VARIATIONAL THEORY AND FLUX SYNTHESIS WITH APPLICATIONS TO FAST REACTOR SPECTRUM CALCULATIONS

by

William George Price, Jr.

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Nuclear Science)
in The University of Michigan
1972

Doctoral Committee:

Assistant Professor James J. Duderstadt, Chairman
Professor Ziya A. Akcasu
Associate Professor Cleve B. Moler
Professor Richard K. Osborn
Associate Professor Fred C. Shure
University Professor Paul F. Zweifel, Virginia Polytechnic Institute and State University

ACKNOWLEDGMENTS

I owe debts of gratitude to many members of my department, both faculty and students, for the advice and encouragement they have given me during the course of my research, but I would particularly like to express my thanks to Dr. James Duderstadt, my advisor, for his continuing confidence in my eventual success.

I gratefully acknowledge the support I have received from the Atomic Energy Commission by the award of a special fellowship in Nuclear Science and Engineering, administered by Oak Ridge Associated Universities, and from the National Science Foundation by the award of a Graduate Fellowship.

A special compliment should be given to the University of Michigan Computing Center for the development and operation of its excellent computer facility, the Michigan Terminal System.

Finally, I must congratulate Madelyn Hudkins for successfully transcribing my handwritten drafts into this readable typescript.

I would like to take this opportunity to dedicate this dissertation to my patient wife, Diane.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF APPENDICES

INTRODUCTION

The calculation of the neutron flux distribution in nuclear reactor cores is of fundamental importance to nuclear reactor design. Since the equation which governs this distribution (the neutron transport equation) is well known, it would seem that the only difficulty in performing such calculations should be in obtaining the physical parameters of the reactor: the neutron-nuclear interaction cross sections.

Unfortunately, the enormous complexity of realistic reactor parameters (with variations in space, angle, energy, and even time) makes it impossible to solve the transport equation itself, and we are left casting around for approximations which are simple enough to be solvable while still imparting useful information. Discretization is the basis for most currently popular approximations. Each independent variable (space, energy, angle, time) is partitioned into intervals, and the behavior of the flux in each interval is assumed to be known (and usually simple); then the approximate equations are solved to find scale factors to be applied in each interval.

The problem with this scheme is due to limitations on the calculational capacity of current computers, which put a limit on the total number of mesh intervals. With such a constraint in effect, the use of small intervals in a region where detail is necessary implies the use of gross intervals to treat the rest of the problem. (Two examples: three group calculations for water reactors where fine spatial detail is essential; one-dimensional calculations for fast reactors where the spectral detail must be resolved.)

Lately there has been a growing interest in synthesis methods of approximation (of which the discretization methods are a special case).

Synthesis is based on the representation of the flux dependence as a combination of simpler (but realistic) "prototype" fluxes. The power of the synthesis methods derives from the fact that any details or trends in the flux distribution which can be predicted a priori can be built into the trials, rather than calculated anew. Further, when this is done the gross coupling coefficients represent shifts of emphasis between realistic modes, and thus express the gross behavior of the flux more clearly than a table of fine-group discrete mesh fluxes.

A particularly suitable application of the synthesis technique should be in the analysis of fast reactors. Here it is known that the energy spectrum shifts fairly slowly and smoothly from region to region, but because of the fine detail in the cross sections any multigroup treatment requires many energy groups. Thus the ability to treat the spatial dependence is limited by the need to calculate redundant spectra everywhere on the spatial mesh. This limitation can be lifted, however, by spectral synthesis - the treatment of the energy dependence as a combination of typical spectra characterizing regions in the core which are of greatest importance.

Unfortunately, spectral synthesis has not really been accepted for this application, principally because the savings in computational effort has not been as dramatic as expected. A secondary factor has been the lack of an analysis of those few situations in which the synthesis apparently degenerate and produces absurd results. This dissertation describes an investigation of the foundations of spectral synthesis techniques.

In the following chapters we shall: review from various sources the theoretical justification of synthesis as a form of variational approximation; analyze the necessary conditions for solution; demonstrate

an equivalence between variational and perturbative methods; develop a computer-oriented form of spectral synthesis; and show how Wielandt's method can be applied to regain the computational advantages originally anticipated for synthesis.

In Chapter I we show that the Calculus of Variations provides a theoretical framework for the derivation of approximation methods. If $F[u]$ is a functional defined on some space of functions $\{u\}$, then in general its first variation $\delta F[u]$ will be zero only for certain particular functions $\bar{u}$. The characterization of any $\bar{u}$ as the stationary point of $F[u]$ is equivalent to its characterization as the solution of some appropriate equation $H\bar{u} = s$, where H is an operator related to the functional derivative of F.

This equivalence can be exploited in two ways: first, the value of $F[u]$ is equal to $F[\bar{u}]$ plus terms of second order in the error $(u-\bar{u})$ and so $F[u]$ can be used as an estimator; second, the stationary points of F in any trial subspace of $\{u\}$ will be an approximation (in the sense of being an early member of a sequence) to $\bar{u}$.

The concept of seeking a stationary point $\tilde{u}$ in a subspace of $\{u\}$ is explored in Chapter II. The point $\tilde{u}$ can be found by solving a reduced equation $\tilde{H}\,\tilde{u} = \tilde{s}$ which is related to but (presumably) simpler than $H\bar{u} = s$; this equation is analyzed in terms of an eigenvector basis for the subspace. A case is described in which there is no solution to the reduced equations; this particular case is unlikely to occur in practice, but it does case some light on the occasional "anomalous" failures of the method.

In Chapter III an alternate application of variational methods is demonstrated. A functional $F[u,z]$ whose value at $u(x)$ is $u(z)$ plus terms

of second order can be applied iteratively to generate a sequence of approximants to $\bar{u}$. This brings to mind Perturbation Theory expansions, and in Chapter IV it is shown that the higher order variational methods are equivalent to the various perturbation expansions of $\bar{u}$, in those cases where the base operator H_0 of the perturbation method is consistent ($H_0 u_0 = s$) with the trial function of the variational scheme.

The use of the Roussopoulos functional $R[v,u] = \langle z,u \rangle + \langle v,s \rangle - \langle v, Hu \rangle$ to derive the multigroup neutron diffusion equations is reviewed in Chapter V, and in Chapter VI these equations are approximated still further by the application of spectral synthesis. The discretized neutron flux is expanded as a sum of known group dependent spectra multiplied by unknown space (and mode) dependent coefficients. In Chapter VII matrix equations for these expansion coefficients are derived, and the MACH-1 one-dimensional neutron diffusion code is adapted to solve them. The spectral synthesis equations feature full coupling of every mode to every other, whereas the multigroup equations were coupled only by downscattering. Because of this, the great computational savings expected from replacing many groups by a few modes does not appear.

It is noted in Chapter VIII that Wielandt's method for extracting eigenvectors has not been applied to the multigroup diffusion equations for this same reason: the extra cost of solving the new system with full energy coupling compensates for the savings derived from accelerated convergence. Wielandt's method can be applied to the spectral synthesis equations, however, without paying this extra penalty (the complication is already there), and so this combination of methods should provide a fairly cheap method of solving fast reactor eigenvalue problems. The synergistic

effect has been demonstrated with the MACH/360 code, incorporating both spectral synthesis and Wielandt's method.

CHAPTER I

REVIEW OF THE CALCULUS OF VARIATIONS

The basic theory of the Calculus of Variation is quite old, with some parts deriving from the work of Euler (1707-1783). The elements of the theory [1, 2] will be reviewed (not with full rigor) to establish the relation between more traditional variational techniques and the practical methods to be developed in later chapters.

We will be dealing with scalars, functions, operators, and functionals. Assuming that all but the last are familiar, we define a functional to be a correspondence which assigns a definite scalar to each function belonging to an appropriate class (its domain). In the same way that we speak of a function f(r) assigning a value to a point r in Euclidean space, we shall say that a function F[u] assigns a value to the "point" u where u(r) is a function in a function space $\{u\}$. Note that a functional can also be an ordinary function of some independent parameter.

For example, consider the set of all curves connecting some fixed point $\underline{r}$ with some other point $\underline{p}$: the length of each curve is a functional of the function describing the curve; the space of acceptable functions is that set describing continuous curves passing through $\underline{r}$ and $\underline{p}$; and if $\underline{p}$ is allowed to vary, the length is also an ordinary function of $\underline{p}$.

We will assume that the argument functions u are members of a real inner product space U, i.e. a linear space with an inner product $\langle v, u \rangle$ and a norm $\|u\|^2 = \langle u, u \rangle$. A functional F[u] is a linear functional if $F[\alpha u + \beta v] = \alpha F[u] + \beta F[v]$; the Riesz-Fischer Theorem states that any bounded linear functional on an inner product space can be written as $\langle w, u \rangle$, where w is an element which is uniquely determined

by F[u]. Using these concepts we can define the variations and derivatives of a functional.

The variations of a functional are analogous to the differentials of an ordinary function. Let $\Delta F[u,h] \equiv F[u + h] - F[u]$ be the increment of F[u] corresponding to a finite variation h(r) in the argument function. If $\Delta F[u,h] = V[u,h] + \|h\| E[u,h]$, where V[u,h] is a linear functional of h(r) and $E[u,h] \to 0$ as $\|h\| \to 0$, then V[u,h] is called the (first) variation of F at u, $\delta F[u,h]$. Higher variations are defined in the same manner: the second variation is the quadratic part of ΔF, etc.

Derivatives of a functional can be defined also. Since the first variation is a linear functional, we define the first derivative of F[u] as that element $\frac{\delta F}{\delta u}[u,r]$ of U such that

$$\delta F[u,h] = V[u,h] = \left\langle \frac{\delta F}{\delta u}[u,r],\ h(r) \right\rangle .$$

Similarly we use

$$\delta^2 F[u,h] = \frac{1}{2} \left\langle \left\langle \frac{\delta^2 F}{\delta u^2}[u,r,p],\ h(r) \right\rangle ,\ h(p) \right\rangle$$

to define the second derivative, etc.

Using the Dirac Delta notation ($\delta(p-r)$ or just δ_p) to represent that generalized function [3] derived from any of the sequences $u_n^p(r) \in U$ with the property that $\lim_{n\to\infty} \langle v(r),\ u_n^p(r) \rangle = v(p)$, we see that we can formally set $h(r) = \delta(p-r)$ and write expressions like

$$\frac{\delta F}{\delta u}[u,n] = \left\langle \frac{\delta F}{\delta u}[u,r] ,\ \delta(n - r) \right\rangle$$

for the derivatives of F.

Use as a Lagrangian

Continuing the analogy with ordinary functions, we investigate the possible existence of extremal points of F[u], points $\bar{u}$ such that the limit of $\Delta F[u,h]$ as $\|h\| \to 0$ has the same sign for all h. A basic theorem states that a necessary condition for $F[\bar{u}]$ to be an extremum is that $\bar{u}$ be a "stationary point" of F, that is, that the first variation vanish at $u = \bar{u}$.

The proof arises directly from the definition of δF as the linear part of the increment of F. If $\delta F[\bar{u},h]$ is not zero, then, for sufficiently small $\|h\|$, $\Delta F[\bar{u},h]$ will have the sign of $\delta F[\bar{u},h]$, but $\Delta F[u,-h]$ will have the opposite sign, and so this $\bar{u}$ cannot be an extremum.

We now introduce the Fundamental Lemma of the Calculus of Variations, which allows us to specify the condition for $\bar{u}$ to be a stationary point without incorporating references to all possible increments h. The lemma states that if $\langle v(r), h(r) \rangle = 0$ for every h(r), then v(r) = 0. (The validity of this can be demonstrated by considering the particular $h(r) = \delta(p-r)$).

Recalling $\delta F[u,h] = \left\langle \frac{\delta F}{\delta u}[u,r], h(r) \right\rangle$, we see immediately that $\delta F[u,h] = 0$ implies that $\frac{\delta F}{\delta u}[u,r] = 0$ for all r. Once again this is a direct parallel to the theory of ordinary functions, in which the first derivative is required to be zero at a function extremum.

The condition that $\frac{\delta F}{\delta u}[u,r] = 0$ is known as Euler's equation, and can be used to find candidates for extrema of F. The solutions are only candidates, of course, because $\delta F = 0$ is only a necessary condition, and at least $\delta^2 F$ must be examined for sufficiency. There are, however, many cases in which the stationary points themselves are of interest, and in these cases the stationary property and Euler's equation provide alternate methods, global vs. local, for determining the $\bar{u}$. The "Variational Principle" that will be used in the rest of this work to derive approximate functions exploits this duality.

The specification of a function as the solution of Euler's equation was originally developed in the field of mechanics. Taking q(t) to be the set of generalized coordinates of a finite dimensional mechanical

system, and $L(q, \dot{q}, t)$ to be the Lagrangian of the system (kinetic energy minus potential energy), Hamilton's Principle states that the time dependence of the actual $\bar{q}(t)$ is such that $H[\bar{q}] = \int_{t_1}^{t_2} L[\bar{q}, \dot{\bar{q}}, t]dt$ is stationary. Euler's equation for this functional is $\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} = 0$, and this is recognized as the Lagrangian equation of motion.

Constraints

It is not uncommon to want to restrict the class of functions to which F is applied, and then to seek the stationary points in this restricted class. (These points do not necessarily form a subset of the stationary points in the full space since the function variations h are restricted also.) Typically this can be expressed as an examination of the variations $F[u,h]$ subject to the constraint that some other functional $G[u + h]$ have a particular value g. Fortunately this form of constraint can be incorporated directly into a Variational Principle by the method of Lagrange Multipliers. Assuming such a stationary point $\tilde{u}$ does exist and that $\tilde{u}$ is not simultaneously a stationary point of G, then there is a scalar λ such that $\tilde{u}$ is a stationary point of the functional $L[v,\lambda] = F[v] + \lambda G[v]$ with respect to arbitrary variations h in the full space. The method for finding $\tilde{u}$ is to find all the stationary points $\tilde{v}(\lambda)$ of L; then $\tilde{u} = \tilde{v}(\lambda)$ for that λ such that $G[\tilde{v}(\lambda)] = g$.

The validity of the Lagrange Multiplier method is demonstrated by examining explicity the variations of F when $h(r)$ is broken into two parts, one arbitrary and the other chosen so that the constraint is satisfied; then the Euler equation incorporating this class of function variation is seen to be the Euler equation of $L[v,\lambda]$. This line of argument can easily be adapted to more complicated constraints. For example,

when the constraint is that $g(u(x)) = g$ for all x, the appropriate Lagrange Multiplier is a function $\lambda(r)$ and $L[v, \lambda] = F[v] + \int_R \lambda(r) g(u(r)) dr$.

Functional Evaluation

So far, Variational Principles have been shown to be interesting alternate schemes for exactly specifying certain functions. What is their connection with Approximation Theory? A very useful one [4,5] since there are two methods for developing approximations to the stationary point (the solution of the Euler equation) of a given functional.

The first involves the actual evaluation of the functional with trial functions close to the stationary point. Since $\delta F[\bar{u},h] = 0$, $F[\bar{u} + h] - F[\bar{u}] = \|h\| \cdot E[\bar{u},h]$, and $E \to 0$ as $\|h\| \to 0$. Thus if $F[\bar{u}]$ is a value of interest, the error in approximating it by $F[u]$ will be of an order higher than the order of the error $\|u - \bar{u}\|$ in u. If $F[\bar{u}]$ is an extremum of F, then $|F[u] - F[\bar{u}]|$ can be used as a "pseudo-norm" to rank the accuracy of a set of approximates u_i to $\bar{u}$; the larger (or smaller) $F[u_i]$ is, the closer u_i is to $\bar{u}$. Finally, if $F[\bar{u}]$ has a known value (e.g. zero), or if $F[u]$ and $G[u]$ are extremum principles bracketing a desired value g, the functional evaluation can provide an absolute measure of the error in any trial u and an absolute limit on the error of using the functionals to estimate g.

As an example of the bracketing approach, Pomraning [6] has created two functionals which can be used to bound the extrapolated endpoint for the Milne problem. Using asymptotic fluxes as trial functions, he generated upper and lower limits which (for strong absorbers) excluded the previously accepted "exact" numerically calculated values; these had been directly evaluated without an estimate of the error in the numerical procedure used.

Variational Approximations

The more important use of variational methods for function approximation is through the derivation of reduced Lagrangians [7,8] i.e. functionals whose stationary points are related to the exact functions but whose Euler equations are easier to solve. Consider the case when $\bar{u}$ is the extremum, in some large space $\{u\}$, of some functional F. If $\{v\}$ is some subspace, and $\bar{v}$ is the extremum of F in this space, what is the relation between $\bar{u}$ and $\bar{v}$? Well, it is possible (but not likely) that $\bar{v} = \bar{u}$. More generally, $\bar{v}$ can be regarded as the "best approximation" to $\bar{u}$ in the space $\{v\}$, in the sense of minimizing $|F[v] - F[\bar{u}]|$, and further, as more degrees of freedom are added to $\{v\}$ so that it approaches $\{u\}$, $\bar{v}$ will become a better approximation to $\bar{u}$ in the sense of this "pseudo-norm" derived from F. Assuming that the Euler equations of F are easier to solve when variations are restricted to the simpler spaces $\{v\}$, the variational principle can be used to generate a sequence of better and better approximations to $\bar{u}$, each approximation requiring a little additional work to evaluate.

This line of reasoning about "best approximations" does not do much good when the goal function is not the extremum of a functional, since then the stationary points in a given restricted space can be arbitrarily higher or lower than $F[\bar{u}]$. The use of reduced Lagrangians can still be justified, however, on the basis that the stationary principle will pick out $v = \bar{u}$ if the subspace $\{v\}$ is made large enough to include $\bar{u}$. It is (unfortunately) up to the analyst to decide how much freedom must be allowed in the restricted space $\{v\}$ in order to obtain an accurate approximation.

A further practical argument can be made to justify the use of variational methods with non-extremal functionals: they give internally consistent approximations. That is, the approximation is determined completely by the

manner in which the subclass of the original full function space is chosen. Since there are no intermediate steps of unrelated approximations involved, the analysis of errors should be (formally) easier to perform. This argument should not be taken lightly, because it, coupled with the fact that they usually work, is the only theoretical justification for most neutron flux synthesis methods.

Generalization to Complex Spaces

Before showing how to construct some useful functionals we must introduce the concept of functionals with argument functions from complex spaces. A common example of such a functional is the inner product on a complex Hilbert space (e.g. $\langle v,u \rangle = \int v^*(r)u(r)dr$ or $\langle v,u \rangle = v^T \cdot u$; this is considered a functional of two arguments. Note that while it is linear in the second argument it is conjugate linear in the first, thus spurring a reconsideration of the definitions of the variations and derivatives of a functional.

We will work with the partial variations and derivatives of a functional with multiple complex arguments, and revise the definition of the first variation in the following manner:

Letting $\Delta_w F[v,u,w,..] = F[v,u,w+h,...] - F[v,u,w,..]$ be the increment of F corresponding to a variation h(r) only in the argument function w(r); if

$$\Delta_w F = V[v,u,w,...h] + \|h\| \cdot E[v,u,w,..]$$

where $E \to 0$ as $\|h\| \to 0$, and if V is either a linear or conjugate linear

functional of h(r), then $\delta_w F = V$ is the (first) variation of F with respect to w(r) at F[v,u,w,...].

We define the partial derivative $\frac{\delta F}{\delta w}[v,u,w,\ldots,r]$ by using the Riesz-Fischer theorem as before. If $\delta_w F$ is linear, then $\left\langle \frac{\delta F}{\delta w}, h \right\rangle = \delta_w F$ defines the derivative; but if $\delta_w F$ is conjugate linear, then we use $\left\langle h, \frac{\delta F}{\delta w} \right\rangle = \delta_w F$ to define the derivative. We are now almost ready to examine some practical functionals.

One last definition is that of the "adjoint operator". If H is an operator on U, then we define H^+ to be its adjoint operator on V if $\langle v, Hu \rangle = \langle H^+ v, u \rangle$ for all possible choices of v and u. The properties of the adjoint operator thus depend on H, U, V and the definition of $\langle v, u \rangle$. Often V is referred to as the adjoint space (to U), and the solution to an equation $H^+ v = z$ will be called the adjoint to the solution of $H u = s$. Adjoints will be used formally throughout the remainder of this work, on the assumption that they can be constructed when needed for any practical problem.

Least Squares

In order to try out the approximation methods suggested earlier we must construct some functionals with useful Euler equations.

Perhaps the most obvious of these is the Least Squares [5] functional; using the inner product as a functional which is linear in its second argument but conjugate linear in its first, consider the functional $L[u] = \langle Hu - s, Hu - s \rangle = \| Hu - s \|^2$. Clearly $L[\bar{u}] = 0$ if $H\bar{u} = s$, while for all other trial functions L[u] is the norm of the residual error, ≥ 0. Thus we can characterize $\bar{u}$ either as the solution of $H\bar{u} = s$ or as the minimizing point of L[u]. The latter condition implies that $\bar{u}$ is a sta-

tionary point at $L[\tilde{u},h] = 0$; unfortunately, it is not necessarily the only stationary point. In fact,

$$L[u] = \langle Hu, Hu \rangle - \langle s, Hu \rangle - \langle Hu, s \rangle + \langle s, s \rangle$$

so that

$$\delta L[u, \delta u] = \langle H\,\delta u, Hu \rangle + \langle Hu, H\,\delta u \rangle - \langle s, H\,\delta u \rangle - \langle H\,\delta u, s \rangle$$

$$= \langle H\,\delta u, Hu - s \rangle + \langle Hu - s, H\,\delta u \rangle$$

$$= \langle \delta u, H^{\dagger}(Hu - s) \rangle + \langle H^{\dagger}(Hu - s), \delta u \rangle$$

$$= \langle H^{\dagger}(Hu - s), \delta u \rangle^{*} + \langle H^{\dagger}(Hu-s), \delta u \rangle$$

$$= 2\,\mathrm{Re}\,\{\langle H^{\dagger}(Hu-s), \delta u \rangle\}$$

The stationary requirement is that $\delta L[\tilde{u}, \delta u] = 0$, and clearly this will be true if $H\tilde{u} = s$, but in general this will not be the only $\tilde{u}$.

There would be no problem if the solution u to Euler's equation were really the function of interest, but problems of this form are not very common. It is unfortunate that the Least Squares Principle leads to this over-complicated, over-generous variational principle, because the absolute error bounding properties of this functional could be quite useful.

Roussopoulos Functional

A much more useful functional is that known as the Roussopoulos functional [9], $R[v,u] = \langle z,u \rangle + \langle v,s \rangle - \langle v,H\,u \rangle$. The usefulness is due to its simple structure and its variational evaluation of the functional $\langle z,H^{-1}s \rangle$. It can be motivated in terms of a goal functional subject to a constraint [10].

First assume that we wish to find a stationary point $\bar{u}$ of the functional $K[u] = \langle z,u \rangle$, so that we can evaluate $K[u]$ with errors of second order in $\|u - \bar{u}\|$. But in addition, we want to restrict the variations of u to allow only those functions which satisfy $H\bar{u} = s$. This is just the specification of a constrained variational problem, and so we look for stationary points $\tilde{u}(v,x)$ of the functional $R[v,u] = \langle z,u \rangle - \langle v, Hu - s \rangle$ (where $g(u) = Hu - s = 0$ is the constraint and $v(x)$ is a Lagrangian multiplier function). There should be a particular $\bar{v}$ such that $H\,\tilde{u}(\bar{v},x) - s(x) = 0$ and $K[\tilde{u}(\bar{v},x)]$ is stationary.

To try to find $\tilde{u}(v,x)$ we set the variation of R with respect to u equal to zero: $\langle z - H^{+}v, \delta u \rangle = \delta_u R = 0$.
By the fundamental lemma, $0 = z - H^{+}v$, which is a condition on v, not u! R is stationary in U only when $v = \bar{v}$, the solution of $H^{+}\bar{v} = z$, and in this case $R[\bar{v},u] = \langle \bar{v}, s \rangle$ for all u. This doesn't do much good in evaluating $\langle z,u \rangle$, but it does suggest looking at the variations of R with respect to v:

$$\delta_v R = \langle \delta v, s - Hy \rangle$$

and so $\delta_v R = 0$ implies $\langle \delta_y, s - Hu \rangle = 0 = s - Hu$. At last we have a condition on u; in fact when u is the solution $\bar{u}$ of $H\bar{u} = s$, then $R[v,\bar{u}] = \langle z, \bar{u} \rangle$ for every v.

The arguments motivating the construction of R are now in a shambles, but we can make use of R anyhow by thinking of it as an ad hoc functional

of the function pair (v,u) and examining its usefulness from scratch. The first variation of R is the sum of $\delta_u R$ and $\delta_v R$, so that $\delta R = \langle \delta v, s - Hu \rangle + \langle z - H^+ v, \delta u \rangle$, and $\delta R[\bar{v},\bar{u}] = 0$ when $H\bar{u} = s$ and $H^+\bar{v} = z$. At this point $R[\bar{v},\bar{u}] = \langle z, \bar{u} \rangle = \langle \bar{v}, s \rangle$ and $\delta^2 R = -\langle \delta v, H \delta u \rangle$, a second order term proportional to the <u>product</u> of the error norms $\| v - \bar{v} \| \cdot \| u - \bar{u} \|$.

Thus the stationary point of R is equivalent to the simultaneous solutions of a direct problem $H\bar{u} = s$ and an adjoint problem $H^+\bar{v} = z$. Furthermore the value of R[v,u] at points close to the stationary point is an estimate of the value $\langle z, \bar{u} \rangle$, as desired originally. The fact that $\langle z,\bar{u} \rangle = \langle \bar{v}, s \rangle$ gives rise to an interpretation of the adjoint function $\bar{v}$ as an "importance" function, since it measures the "importance" of an arbitrary source to the goal $\langle z,\bar{u} \rangle$.

An alternate justification for the construction of R has been given by Selengut [11]. He proposes to evaluate some arbitrary functional F which depends on both the state function $\bar{\phi}$ and an importance function $\bar{\phi}^+$ of some physical system. Requiring that $F[\phi^+, \phi]$ be expandable in a Taylor series with no terms higher than second order leads to an expression

$$F[\phi^\dagger, \phi] = A_0 + \langle A_1^\dagger(r), \phi(r) \rangle + \langle A_1(r), \phi^\dagger(r) \rangle + \langle \phi^\dagger(r), A_2 \phi(r) \rangle$$

Now requiring that F be insensitive to small perturbations in ϕ^+ and ϕ , i.e. that the first variation of F be zero, leads to the requirements that

$$A_1(r) + A_2 \phi(r) = 0 \equiv s - H\bar{u}$$

$$A_2(r) + A_2 \phi^\dagger(r) = 0 \equiv z - H^\dagger \bar{v}$$

so that $F[\bar{\phi}^+, \bar{\phi}] = A_0 + \left\langle A_1^+(r)\bar{\phi}(r)\right\rangle = A_0 + \left\langle z, \bar{u}\right\rangle$.

Thus the goal of writing an insensitive, simple functional for evaluating some property of a physical system can be met provided that the desired property can be evaluated as $A_0 + \langle z,u \rangle$ and the state of the system can be specified by linear equations of the form $H\bar{u} = s$ and $H^+\bar{v} = z$; the functional which does this is just the Roussopoulos functional (with an arbitrary constant term A_0).

Although the formal Roussopoulos functional seems to provide a very flexible method for generating functionals whose Euler equations have desirable solutions, there are certain classes of problems which cannot be handled in this simple manner. One class is that in which the goal functional, for which a second order estimate is desired, is not linear. Pomraning [10] has considered this problem and has shown how to construct a generalized form of the Roussopoulos functional, $P[v,u] = L[u] + \langle v, s-Hu\rangle$, which estimates an arbitrary goal functional $L[\bar{u}]$ to second order when the trial functions are approximations to $H\bar{u} = s$ and $H^+\bar{v} = \frac{dL}{du}[\bar{u}]$.

Rayleigh Principle

An equally important class to treat is that set of problems for which the desired Euler equations are eigenvalue equations: $H\bar{u} = \lambda\bar{u}$ and $H^+\bar{v} = \lambda^*\bar{v}$. A functional related to the Rayleigh Principle can be developed for this class of eigenvalue problems in much the same way that the Roussopoulos functional was developed: we seek stationary points of the functional $K[v,u] = \langle v, Hu\rangle$ subject to the constraint $G[v,u] = \langle v,u\rangle =$ constant g. As a motivation for this, assume that H and H^+ have complete orthonormal sets of eigenvectors u_n and v_n and expand u and v in terms of these: $u = \sum_n \alpha_n u_n$, $v = \sum_m \beta_n v_n$.

Now $K[u,v] = \sum_n \lambda_n \alpha_n \beta_n^*$ while $G[u,v] = \sum_n \alpha_n \beta_n^*$. With the value of G constrained to a constant g, the value of K should contain some information about the eigenvalues λ_n.

Using the Lagrange Multiplier technique we write $E[v,u,\lambda] = K[v,u] - \lambda G[v,u]$ or $E[v,u,\lambda] = \langle v, Hu \rangle - \lambda \langle v,u \rangle = \langle v, (H-\lambda)u \rangle$. The stationary point $(\bar{v},\bar{u})$ of E is determined by requiring $\delta E = \langle \delta v, (H-\lambda)\bar{u} \rangle + \langle (H^+ - \lambda^*)\bar{v}, \delta u \rangle = 0$, so $\bar{v}$ and $\bar{u}$ must satisfy $H\bar{u} = \lambda \bar{u}$ and $H^+\bar{v} = \lambda^* \bar{v}$, which is possible if and only if $\bar{u} = u_n$, $\bar{v} = v_n$, and $\lambda = \lambda_n$. In this situation $E[v_n, u_n, \lambda_n] = 0$ and $K[v_n, u_n] = \lambda_n g$; each of the stationary points characterizes a different eigenvalue.

Since $K[v_n,u_n] = \lambda_n g = \lambda_n G[v_n,u_n]$, we are tempted to examine the functional $Y[v,u] = K[v,u]/G[v,u]$. We already know that $Y[v_n,u_n] = \lambda_n$; perhaps this is a stationary point also (of course it is - Y is the Rayleigh Principle for evaluating eigenvalues [12] - but we shall prove it anyway). Consider $E[v,u,\lambda]$:

$$E[v_n+\delta v, u_n+\delta u, \lambda_n+\delta\lambda] = \langle v, (H-\lambda_n)u_n \rangle$$

$$+ \langle v_n, (H-\lambda_n)\delta u \rangle + \langle \delta v, (H-\lambda_n)\delta u \rangle - \delta\lambda G[v,u]$$

$$= -\delta\lambda G[v,u] + \langle \delta v, (h-\lambda_n)\,\delta u \rangle$$

$$= K[v,u] - (\lambda_n + \delta\lambda)\, G[v,u]$$

Rearranging terms,

$$\frac{K[v,u]}{G[v,u]} = \lambda_n + \frac{\langle \delta v, (H-\lambda_n)\delta u \rangle}{G[v,u]}$$

or

$$Y[v,u] = \lambda_n + \frac{\langle \delta v, (H-\lambda_n)\delta u \rangle}{\langle v_n,u_n \rangle + \langle \delta v, \delta u \rangle}$$

Clearly (v_n, u_n) is a stationary point of Y, since the errors are of second order. The functional $E[v,u,\lambda]$ is probably more useful than Y for deriving approximations, however, because of the non-linear nature of the latter. The Rayleigh Principle can always be applied after the approximating is done, to extract the eigenvalue estimate.

Pomraning has gone a step further [13,14] in the derivation of functionals and produced a sort of hybrid of R and E. The stationary conditions of this new functional are the eigenvalue equations $Hu - \lambda u = 0$ and $H^+v - \lambda^* v = \frac{\delta L}{\delta u}$, and the value of the functional at its stationary point is a second order estimate of an arbitrary homogeneous goal functional L[u]. This is quite reasonable; although the solution $\bar{u}$ of a critical system has an arbitrary normalization, homogeneous (ratio) functionals still have validity. The forms of the functional and the trial functions for v are not simple, but the ability to calculate more than just the eigenvalue in a critical system may make the extra complication worthwhile.

CHAPTER II

VARIATIONAL APPROXIMATIONS

In neutron transport theory (and in most other disciplines) there are very few realistic problems which can be solved exactly. We can write formally $H\bar{u} = s$, or equivalently $\delta F[\bar{u}] = 0$, but solving $\bar{u} = H^{-1}s$ or finding this stationary point is out of the question. We have to be satisfied with an approximation $\tilde{u}$, where $\tilde{u}$ is "like" $\bar{u}$ (in some arbitrary sense), and our success in working with $\tilde{u}$ instead of $\bar{u}$ depends on the appropriateness of our criterion for "likeness".

Restricted Domain

The Variational Principle, which is used to convert locally determined problems ($H\bar{u} = s$) into globally determined problems ($\delta F[\bar{u}] = 0$), can be put to work in a Variational Approximation Technique [4,8] to find $\hat{u} \cong \bar{u}$ in the following manner: let $\bar{u}$ be the stationary point of $F[u]$; to get an easier problem we restrict the trials and variations to some subspace U of the full inner product space U. The stationary point $\tilde{u}$ in this restricted space, i.e. the solution of the reduced Euler equations, is taken to be the approximation to $\bar{u}$. With this scheme, $\tilde{u}$ is "like" $\bar{u}$ in the sense that both are chosen by the same variational procedure; the accuracy will depend on the nature of the subspace $\tilde{U}$; and the hope is that the use of the "global" method will smooth out and reduce the error in $\tilde{u}$ in a way that could not be achieved with "local" approximations to H.

Extremum Functionals

Fortunately there are situations in which the accuracy of the Variational Approximation Technique can be assessed directly, thereby in-

creasing confidence in its general application. When F is a maximum functional this is particularly easy, because then $F[u \in \tilde{U}] \leq F[\tilde{u}] \leq F[\bar{u}]$. This gives unambiguous meaning to the statement that u is the best approximation to $\bar{u}$ in the reduced space $\tilde{U}$ in the sense of the "pseudo-norm" derived from F. It also provides a means for evaluating the suitability of this trial space; as degrees of freedom are added to $\tilde{U}$ the evaluated F[u] always gets closer to $F[\bar{u}]$, and the changes in F[u] are indicators of the importance of those extra degrees of freedom.

Unfortunately our primary interest is in the functionals R[v,u] and $E[v,u,\lambda]$, which because of the trial adjoint functions do not in general provide extremum principles. In those special cases, however, when the operator H is self adjoint <u>and</u> the real source s is the same as the adjoint source z we can derive extremum principles even from R and E.

If $z = s$ and $H^+ = H$, we can rewrite $R[u] = R[u, u] = 2\langle s,u\rangle - \langle u, H u\rangle$; so $R[\bar{u}] = \langle s, \bar{u}\rangle$, and $\delta^2 R[u, \delta u] = -\langle \delta u, H \delta u\rangle$. If H is a positive (or negative) operator, i.e. $\langle w, Hw\rangle$ is always positive (or negative), then R[u] provides an extremum principle for $\langle s, u\rangle$, and the analysis above can be used in approximating $\bar{u}$.

For eigenvalue problems, $H^+ = H$ implies that $v_n = u_n$ and $\lambda_n^* = \lambda_n$, so we can write

$$E[u,u,\lambda] = \langle u, (H-\lambda)u\rangle = E[u, \lambda] \text{ and } Y[u,u] = \frac{\langle u, Hu\rangle}{\langle u, u\rangle} = Y[u];$$

furthermore,

$$\delta^2 Y[u] = + \langle \delta u, (H-\lambda_n)\delta u\rangle / \langle u,u\rangle ,$$

$$\langle u,u\rangle \, \delta^2 Y[u] = \langle \delta u, H\delta u\rangle - \lambda_n \langle \delta u, \delta u\rangle .$$

If H is positive, then there are two numbers λ_0 and λ_N such that

$$\lambda_N \geq \lambda_n \geq \lambda_0 > 0,$$

and so

$$\lambda_N \langle u, u \rangle \geq \langle u, Hu \rangle \geq \lambda_0 \langle u, u \rangle .$$

Clearly

$$\langle u, u \rangle \ (Y[u] - \lambda_N) = \langle u, Hu \rangle - \lambda_N \langle u, u \rangle \leq 0$$

while

$$\langle u, u \rangle \ (Y[u] - \lambda_0) = \langle u, Hu \rangle - \lambda_0 \ \langle u, u \rangle \geq 0$$

and upon rearranging, we see that $\lambda_N \geq Y[u] \geq \lambda_0 > 0$.

(Similarly when H is negative $\lambda_N \leq Y[u] \leq \lambda_0 < 0$) thus Y[u] is an extremum principle for both the largest and smallest eigenvalues, and can be used as described above to evaluate approximations to u_0 and u_N.)

Bracketing

There is a subclass of the extremum functionals whose members provide an even better error estimating property than the monotonic convergence of the ordinary extremum functionals. These are the bracketing functionals - extremum functionals evaluating known limits, or pairs of functionals providing both upper and lower bounds for the same value.

The prime example of the former type is the class of Least Squares functionals [5]. As described earlier, $L[u] = \langle Hu - s, Hu - s \rangle \geq L[\bar{u}] = 0$. Thus L[u] is directly a measure of the error in u; in fact by construction it is the norm of the residual source.

There are few examples other than the Least Squares of functionals which evaluate known quantities, so the few methods for constructing "reciprocal" functionals - methods which provide the corresponding bracketing functional when one extremum is known - are of great (theoretical) interest.

Pomraning [6] describes how to do this when starting out with a positive definite, self-adjoint operator H and the extremum Roussopoulos functional $R[u] = 2\langle s,u\rangle - \langle u, Hu\rangle \leq \langle s,\bar{u}\rangle$. He shows that it is always possible to decompose H into the sum $L + T^{\dagger}T$, where L and $T^{\dagger}T$ are both non-negative, so that $H\bar{u} = s$ can be rewritten as the coupled pair of equations $w=T\bar{u}$ and $L\bar{u} + T^{\dagger}w = s$. Assuming that the inverse of L is calculable (and this is probably a big assumption) we can use a new functional

$$P[u] = R[u] + \left\langle (Hu - s), L^{-1}(Hu - s)\right\rangle$$

for the other side of the bracket, since $P[\bar{u}] = R[\bar{u}]$ by inspection and $P[u] \geq \langle s,\bar{u}\rangle$.

The proof that this is a minimum principle and the proofs of several other similarly constructed brackets [16, 17] can be written as special cases of a generalization and extension of Hamilton's canonical transformation [15].

As the starting point in the transformation we assume that we are given some functional $G_1[u,w]$ and we define $G_2[u] = G_1[u,Ku]$. If $\bar{u}$ is the stationary point of G_2 and $\delta^2 G_2[\bar{u}] \geq 0$, then G_2 is a minimum principle for some value $D = \min_u G_2[u]$, and G_1 is also a minimum principle when w is restricted to equal Ku. (Note that $\bar{u}$ is determined by the Euler equation $\frac{\delta G_1}{\delta u}[\bar{u}, K\bar{u}, r] + K^{+}\frac{\delta G_1}{\delta w}[\bar{u}, K\bar{u}, r] = 0$). In order to avoid the restriction w = Ku, we use the Lagrange Multiplier technique to define a new functional $G_3[u,w,v] = G_1[u,w] + \left\langle v,(Ku-w)\right\rangle$ where v is restricted to those functions for which G_3 has a stationary point. Denoting this stationary point of G_3 (for variations of v and w with v held constant) as $(\tilde{u}(v), \tilde{w}(v), v)$, we see that $\tilde{u}$, $\tilde{w}$, and v must satisfy the Euler equations

$$\frac{\delta G_1}{\delta u}[\tilde{u}, \tilde{w}, r] = -K^{+}v(r) \text{ and } \frac{\delta G_1}{\delta w}[\tilde{u}, \tilde{w}, r] = v(r).$$

If $\delta^2 G_3 \geq 0$ for variation of u and w we can define

$$G_4[u, v] \equiv \min_w G_3[u, w, v]$$

and $$G_5[v] \equiv \min_u G_4[u, v] = \min_u \min_w G_3[u, w, v]$$

and show an interesting relation between these functionals and D. Recalling the Euler equation for G_1, we see that the set

$$\bar{u}, \bar{w} = K\bar{u}, \text{ and } \bar{v} = \frac{\delta G_1}{\delta w}[\bar{u}, K\bar{u}, r]$$

is a stationary point of G_3, and that

$$G_3[\bar{u}, \bar{w}, \bar{v}] = G_4[\bar{u}, \bar{v}] = G_5[\bar{v}] = D.$$

Now since

$$G_5[v] = \min_u G_4[u, v] = \min_u [\min_w G_3[u, w, v]]$$

for any v, and this is

$$G_5[v] \leq \min_{u, w=Ku} G_3[u, w, v] = \min_u G_2[u] = D,$$

the conclusion is that $$D = \max_v G_5[v] = \max_v \min_u G_4[u, v].$$

$G_2[u]$ and $G_5[v]$

are known as Reciprocal or Involutory functionals for D, while $G_4[u,v]$ is known as the canonical form. The bracketing properties depend on the fact that $G_2[\tilde{u}]$ is an extremum and that $G_3[u,w,v]$ has an extremum with respect to u and w. If these conditions do not hold the Canonical and

Involutory transformations are still valid, but the intermediate steps and the final functionals are all just stationary principles.

When brackets do exist, they can be used to evaluate D within known error limits, as was done by Pomraning in his calculation of the Milne problem extrapolated endpoint described earlier [6]. They can also be used to evaluate the error in function approximations, even though $\bar{u}$ and $\bar{v}$ presumably satisfy different Euler equations, by relating them both to a third function $\bar{w}$. If there are operators such that $\bar{u} = U(\bar{w})$ and $\bar{v} = V(\bar{w})$, then the values of $G_2[U(w)]$ and $G_5[V(w)]$, for any w, can be used as indicators of the accuracy of $U(w) \approx u$ and $V(w) \approx v$.

Non Self-Adjoint Systems

One other method of generating bracketing principles, that of Buslick [18], deserves examination. This method is based on the transformation of a non-self-adjoint system $Hu = s$ and $H^+v = z$ into self-adjoint and anti-self-adjoint parts. First we define $H_s = \frac{1}{2}(H + H^+)$, $H_a = \frac{1}{2}(H - H^+)$, $w_s = \frac{1}{2}(u + v)$, $w_a = \frac{1}{2}(u - v)$ and finally $q_s = \frac{1}{2}(s + z)$, $q_a = \frac{1}{2}(s - z)$. Substituting these definitions in the original equations, we derive an equivalent set $H_s w_s + H_a w_a = q_s$, $H_a w_s + H_s w_a = q_a$. Now we assume that H_s^{-1} can be calculated, and eliminate $w_a = H_s^{-1}(q_a - H_a w_s)$ to derive $(H_s - H_a H_s^{-1} H_a)\, w_s = B w_s = q_s - H_a H_s^{-1} q_a$. This coupled pair of equations is also equivalent to the original pair, but each is written in terms of a self-adjoint operator (since $B = H_s - H_a H_s^{-1} H_a = B^+$). $\langle f, Bf \rangle \geq 0$ if and only if $\langle f, H_s f \rangle \geq 0$. The Roussopoulos functional can now be written for the w_s equation:

$$R[w] = 2 \left\langle q_s - H_a H_s^{-1} q_a, w \right\rangle - \langle w, Bw \rangle ,$$

with

$$R[w] \geq R[w_s] = \langle z, \bar{u} \rangle + \frac{1}{4} \langle (s-z), H_s^{-1}(s-z) \rangle$$

provided that B (or equivalently H_s) is positive. Repeating the derivation with z replaced by (-z) does not affect the extremum property but does give a new functional R'[w], with the property that

$$R'[w] \geq -\langle z, \bar{u} \rangle + \frac{1}{4} \langle (s+z), H_s^{-1}(s+z) \rangle .$$

Therefore, since the second term in each of these expressions is a known constant, the two functionals can give upper and lower bounds on the ordinary goal functional $\langle z,u \rangle$.

The particular value of this method is that it shows a way to construct bracketing functionals for non-self-adjoint problems, problems which never even have extremals when treated with the ordinary functionals. Of course a price is paid, and that is the much greater complication in the effective operator B.

In fact, the complication inherent in each of these methods is quite apparent and is the reason they are hardly ever used to generate approximate solutions. There is no reason to expect that these bracketing functionals will produce better approximations than those which the usual Roussopoulos functional does (the experience has been the reverse [18]) - what they do provide is a means of assessing the accuracy of an approximation once it has been generated. From a practical standpoint the preferred procedure is to generate approximations using simple, easy to handle functionals and hold these special functionals in reserve for occasional verification of those results.

Reduced Functionals

Once a suitable functional has been chosen the trial functions are allowed to vary only in a chosen subset $\tilde{\mathcal{U}} \subset \mathcal{U}$. In the following section we will assume $\tilde{U}$ is the set of all functions U(x), where U is a mapping taking elements of the linear space X into U, and that $\tilde{V}$ is the set of all functions V(y), where V is a mapping taking elements of the linear space Y into V. (Examples might be that X is a vector space and U a transformation matrix, or X is the set of scalars and U(x)=exp(ax)). We assume that the mappings U and V can be absorbed into the functional: and write F[V(y), U(x)] = F[y,x]. (We call this the "reduced" functional since only x and y are variable). Now we seek the stationary point F[y,x] in the reduced spaces Y and X; the Variational Approximations $\tilde{v}$ and $\tilde{u}$ are seen to be $V(\tilde{y})$ and $U(\tilde{x})$.

For the reasons given in the previous section we are assuming that "practical" methods of approximation [4,7] will be based on the use of the Roussopoulos functional R[v,u] (or E[v,u] for eigenvalue problems). If the nature of the transformations U(x) and V(y) can be specified more closely, it is possible to perform a small amount of error analysis even for the non-extremum cases.

We will not (yet) assume that U or V are linear operators or that $\{U(x)\}$ and $\{V(y)\}$ are full linear subspaces of U and V. We do assume, however, that for any small perturbation δx in x we can expand $U(x+\delta x) = U(x) + U'(x) \cdot \delta x + O\|\delta x\|^2$, and similarly for V(y). Thus it is clear that if

$$F[y, x] = \langle z, U(x) \rangle + \langle V(y), s \rangle - \langle V(y), HU(x) \rangle$$

then $\delta F[y, x, \delta y, \delta x] = \langle z, U'(x) \cdot \delta x \rangle + \langle V'(y) \cdot \delta y, s \rangle$

$$- \langle V(y), HU'(x) \cdot \delta x \rangle - \langle V'(y) \cdot \delta y, HU(x) \rangle$$

$$= \langle z - H^{+}V(y), U'(x) \cdot \delta x \rangle + \langle V'(y) \cdot \delta y, s - HU(x) \rangle .$$

The stationary requirement that $\delta F[\tilde{y}, \tilde{x}, \delta y, \delta x] = 0$ then gives the Euler equations

$$\langle z - H^{+}V(\tilde{y}), U'(\tilde{x}) \cdot \delta x \rangle = 0 \qquad \forall\ \delta x$$

and

$$\langle V'(\tilde{y}) \cdot \delta y, s - HU(\tilde{x}) \rangle = 0 \qquad \forall\ \delta y.$$

However, because of the linearity of the derivative term we can expand $U'(x) \cdot \delta x = \sum_{i}^{I} U_i'(x)\ \delta x_i$, i.e. we can expand δx in components δx_i corresponding to a basis for the space X (even when this is not valid for $U(x)$ itself). Similarly we expand $V'(y) \cdot \delta y = \sum_{j}^{J} V_j'(y)\ \delta y_j$, and then the Euler equations can be rewritten

$$\sum_i \langle z - H^{+}V(\tilde{y}), U_i'(\tilde{x}) \rangle \delta x_i = 0 \qquad \forall\ x$$

$$\text{or} \quad \langle z - H^{+}V(\tilde{y}), U_i'(\tilde{x}) \rangle = 0 \qquad \forall\ i\ ,$$

$$\text{and} \quad \sum_j \delta y^{*}_j \langle V'_j(\tilde{y}), s - HU(\tilde{x}) \rangle = 0 \qquad \forall\ \delta y$$

$$\text{or} \quad \langle V'_j(\tilde{y}), s - HU(\tilde{x}) \rangle = 0 \qquad \forall\ j.$$

Already some dangerous situations are apparent. For example, if there is any $\hat{x}$ such that $\langle V_j'(\tilde{y}), H\ U(\hat{x}) \rangle = 0$ for every j, then the stationary point $\tilde{x}$ may contain arbitrary amounts of $\hat{x}$. More dangerously, if there is any δy such that

$$\langle V'(\tilde{y}) \cdot \delta y, s \rangle \neq 0$$

even though $\langle V'(\tilde{y}) \cdot \delta y, H\ U(x) \rangle = 0$ for every x then there is no stationary point.

We assume that a stationary point $(\tilde{v},\tilde{u}) = (V(\tilde{y}), U(\tilde{x}))$ does exist so that we can continue the analysis, and we build this analysis on an expansion of the trial functions $U(x)$ and $V(y)$ in eigenvectors of H. Assuming there exist complete biorthonormal sets of eigenvectors for H^+ and H, $Hu_m = \lambda_m u_m$ and $H^+ v_n = \lambda^*_n v_n$, $\langle v_n, u_m \rangle = \delta_{nm}$,

we can expand any

$$u = \sum_k u_k c_k, \quad v = \sum_l v_l d_l,$$

$$s = \sum_m u_m e_m, \text{ and } z = \sum_n v_n f_n.$$

(We pause here to notice that

$$0 = \langle v_l, s - H\bar{u} \rangle = \langle v_l, \sum_r u_r e_r - H \sum_k u_k \bar{e}_k \rangle$$

$$= e_l - \langle v_l, \sum_k \lambda_k u_k \bar{c}_k \rangle = e_l - \lambda_l \bar{c}_l$$

so that the exact stationary point can be written in this eigenvector notation as $\bar{u} = \sum_k u_k e_k / \lambda_k$, $\bar{v} = \sum_l v_l f_l / \lambda_l$). Continuing, we expand $U(x) = \sum_k u_k c_k(x)$ and $U'_i(x) = \sum_k u_k a_{ki}(x)$ in eigenvectors, and also

$$V(y) = \sum_l v_l d_l(y) \text{ and } V'_j(y) = \sum_l v_l b_{lj}(y).$$

Substituting these expansions, the first Euler equation becomes

$$\langle \sum_l v_l b_{lj}(\tilde{y}), \sum_m u_m e_m \rangle = \langle \sum_l v_l b_{lj}(\tilde{y}), H \sum_k u_k c_k(\tilde{x}) \rangle$$

or, using the orthonormality conditions,

$$\sum_l b^*_{lj}(\tilde{y}) e_l = \sum_l b^*_{lj}(\tilde{y}) \lambda_l c_l(\tilde{x}) \qquad \forall\ j.$$

Similarly the second Euler equation reduces to

$$\sum_k a^*_{ki}(\tilde{x}) f_k = \sum_k a^*_{ki}(\tilde{x}) \lambda^*_k d_k(\tilde{y}) \qquad \forall\ i.$$

The Euler equations for the reduced functional thus have been transformed into a set of simultaneous (not necessarily linear) equations for the coefficients $c_\ell(\tilde{x})$ and $d_\ell(\tilde{y})$ of the eigenvalue expansions of the stationary point $V(\tilde{y})$, $U(\tilde{z})$.

Linear Subspaces of Trials

Not much information can be obtained from the stationary conditions in this form, so again we simplify, finally assuming that U and V are linear transformations from the spaces X and Y into subspaces of U and V. (From here on the analysis will not apply to nonlinear operators like $U(x) = \exp(ax)$, for example). Because of the linearity of U and V, we write

$$u(r) = U(r,x) = \sum_i^I U_i(r)\, x_i \qquad \text{and} \qquad U_i'(x) = U_i$$

but recall that we have already written $U(x) = \sum_k u_k c_k(x)$. By using the expansion of each of the $U_i' = U_i$ in eigenvectors $U_i = \sum_k u_k a_{ki}$ we see that

$$U(x) = \sum_i^I \sum_k u_k a_{ki} x_i = \sum_k u_k \sum_i a_{ki} x_i$$

so that

$$C_k(x) = \sum_i a_{ki} x_i$$

Similarly we write

$$v(r) = V(r,y) = \sum_j^J V_j(r)\, y_j \qquad \text{and} \qquad V_j'(y) = V_j = \sum_l v_l\, b_{lj}$$

so that

$$V(y) = \sum_l v_l \sum_j b_{lj} y_j \text{ and } d_l(y) = \sum_i b_{lj} y_j.$$

Substituting c and d into the Euler equations yields

$$\sum_l b^*_{lj} e_l = \sum_l b^*_{lj} \lambda_l \sum_i a_{li} \tilde{x}_i \qquad \forall\ j$$

and

$$\sum_k a^*_{ki} f_k = \sum_k a^*_{ki} \lambda^*_k \sum_j b_{kj} \tilde{y}_j \qquad \forall\ i.$$

We now have two uncoupled sets of linear equations which can be solved for the components of $\tilde{y}$ and $\tilde{x}$, the stationary points in the reduced spaces.

Adopting a matrix and vector notation to simplify the expression of these linear equations, we write the alternate forms

$$U(x) = \underline{U} \cdot \underline{x} = \underline{u} \cdot \underline{\underline{a}} \cdot \underline{x},\ V(y) = \underline{V} \cdot \underline{y} = \underline{v} \cdot \underline{\underline{b}} \cdot \underline{y},$$

$$s = \underline{u} \cdot \underline{e} \text{ and } z = \underline{v} \cdot \underline{f},$$

where the elements of $\underline{u}$ and $\underline{v}$ are all the eigenvectors u_{ℓ} and v_{ℓ}, the elements of $\underline{\underline{a}}$ and $\underline{\underline{b}}$ are the expansion elements a_{ki} and b_{lj}, and the I components of $\underline{x}$ and the J components of $\underline{y}$ determine the functions x and y in the X and Y spaces.

Starting over again with the original Euler equations, we see that

$$\left\langle V'(\tilde{y}) \cdot \delta y,\ s - HU(\tilde{x}) \right\rangle = 0 \qquad \forall\ \delta y$$

becomes

$$\underline{\delta y}^T \cdot \left\langle \underline{v} \cdot \underline{\underline{b}},\ \underline{u} \cdot \underline{e} - H\, \underline{u} \cdot \underline{\underline{a}} \cdot \tilde{\underline{x}} \right\rangle = 0 \quad \forall\ y$$

or

$$\underline{\delta y}^T \cdot \underline{\underline{b}}^T \cdot \left\langle \underline{v},\ \underline{u} \cdot \underline{e} \right\rangle = \underline{\delta y}^T \cdot \underline{\underline{b}}^T \cdot \left\langle \underline{v},\ H\, \underline{u} \cdot \underline{\underline{a}} \cdot \tilde{\underline{x}} \right\rangle$$

or

$$\underline{\underline{b}}^T \cdot \left\langle \underline{v},\ \underline{u} \right\rangle \cdot \underline{e} = \underline{\underline{b}}^T \cdot \left\langle \underline{v},\ H\, \underline{u} \right\rangle \cdot \underline{\underline{a}} \cdot \tilde{\underline{x}} \quad ,$$

and similarly the equation for y becomes

$$\underline{f}^T \cdot \left\langle \underline{v},\ \underline{u} \right\rangle \cdot \underline{\underline{a}} = \tilde{\underline{y}}^T \cdot \underline{\underline{b}}^T \cdot \left\langle H^+ \underline{v},\ \underline{u} \right\rangle \cdot a \ .$$

Now if H(and H^+) had complete orthonormal sets of eigenvectors then $\langle \underline{v},\underline{u} \rangle$ would be the identity matrix and $\langle \underline{v}, H\underline{u} \rangle = \langle H^+\underline{v}, \underline{u} \rangle$ would be the diagonal matrix of eigenvalues, so we denote $\underline{\underline{\tilde{I}}} = \langle \underline{v},\underline{u} \rangle$ and $\underline{\underline{\tilde{D}}} = \langle \underline{v}, H\underline{u} \rangle$, and define $\underline{\underline{\tilde{H}}} = \underline{\underline{b}}^T \cdot \underline{\underline{\tilde{D}}} \cdot \underline{\underline{a}}$ in order to write the stationary conditions as

$$\underline{\underline{b}}^T \cdot \underline{\underline{\tilde{I}}} \cdot \underline{e} = \underline{\underline{b}}^T \cdot \underline{\underline{\tilde{D}}} \cdot \underline{\underline{a}} \cdot \underline{\tilde{x}} = \underline{\underline{\tilde{H}}} \cdot \underline{\tilde{x}} = \underline{\tilde{e}}$$

and

$$\underline{f}^T \cdot \underline{\underline{\tilde{I}}} \cdot \underline{\underline{a}} = \underline{\tilde{y}}^T \cdot \underline{\underline{b}} \cdot \underline{\underline{\tilde{D}}} \cdot \underline{\underline{a}} = \underline{\tilde{y}}^T \cdot \underline{\underline{\tilde{H}}} = \underline{\tilde{f}}^T$$

or $\underline{\underline{\tilde{H}}}^T \cdot \underline{\tilde{y}} = \underline{\tilde{f}}$

The elementary theory of linear equations shows that in general (arbitrary s and z functions) there will be undetermined components of $\underline{\tilde{x}}$ unless $J \geq I$, and undetermined components of $\tilde{y}$ unless $I \geq J$, implying that I must equal J to ensure a unique stationary point $(\tilde{y}, \tilde{x})$ unless special assumptions can be made about the source functions. Actually, there may be multiple solutions even when I = J, because if $\underline{\underline{\tilde{H}}} \cdot \underline{\hat{x}} = 0$ or $\underline{\underline{\tilde{H}}}^T \cdot \underline{\hat{y}} = 0$, then arbitrary multiples of $\underline{\hat{x}}$ and $\underline{\hat{y}}$ can be added to any particular solutions $\underline{\tilde{x}}$ and $\underline{\tilde{y}}$. Uniqueness of the stationary point may not be a requirement of the approximation problem, but if it is, it is good to be aware of potential failure.

A more dangerous situation arises if one of the stationary equations cannot be solved. The condition for this can be stated precisely using the notion of the rank of a matrix - the maximum number of linearly independent rows or columns (note that rank $(\underline{\underline{\tilde{H}}})$ = rank $(\underline{\underline{\tilde{H}}}^T) \leq$ smaller of I,J). There will be a solution to $\underline{\underline{\tilde{H}}} \cdot \underline{\tilde{x}} = \underline{\tilde{e}}$ if and only if the rank of $\underline{\underline{\tilde{H}}}$ is equal to the rank of the augmented matrix $(\underline{\underline{\tilde{H}}} \oplus \underline{\tilde{e}})$. Likewise, there will

be a solution to $\tilde{\underline{\underline{H}}}^T \cdot \tilde{\underline{y}} = \tilde{\underline{f}}$ if and only if the rank of $\tilde{\underline{\underline{H}}}^T$ is equal to the rank of the augmented matrix $(\tilde{\underline{\underline{H}}}^T \oplus \tilde{\underline{f}})$. These formal statements are not really very useful (determining the rank is roughly equivalent to trying to solve the system) but they suggest caution rather than blind confidence, particularly when I is not equal to J and both are greater than the rank of $\tilde{\underline{\underline{H}}}$. A case which is unlikely to ever occur in practice but in which there will be no stationary point of the reduced system occurs if the trial functions U(x) form a linear subspace of U spanned by a set of eigenvectors u_k, the trial functions V(y) form a linear subspace of V spanned by a set of eigenvectors v_l, and the v_l are not the corresponding adjoint eigenvectors to the u_k.

Eigenvalue Problems

This analysis of the stationary point of the reduced functional can also be applied to the eigenvalue functional $E[v,u,\lambda] = \langle v,(H-\lambda I)u\rangle$. We define the reduced functional $E[y,x,\lambda] = E[V(y), U(x), \lambda]$, and pass immediately to the assumption that U and V are linear transformations from X and Y to subspaces of U and V.

Using the same expansions as before, we see that

$$\langle V'(\tilde{y}) \cdot \delta y, (H - \tilde{\lambda} I)\, U(\tilde{x})\rangle = 0 \qquad \forall\, \delta y$$

becomes

$$\sum_i (\sum_l b^*_{lj} \lambda_l a_{li})\tilde{x}_i = \tilde{\lambda} \sum_i (\sum_l b^*_{lj}\, a_{li})\tilde{x}_i \qquad \forall\, j$$

and

$$\langle (H^+ - \tilde{\lambda}^* I)V(\tilde{y}),\, U'(\tilde{x}) \cdot \delta x\rangle = 0 \qquad \forall\, \delta x$$

becomes

$$\sum_j (\sum_k a^*_{ki} \lambda^*_k b_{kj})\tilde{y}_j = \tilde{\lambda}^* \sum_j (\sum_k a^*_{ki} b_{kj})\tilde{y}_j \qquad \forall\, i$$

In matrix notation (defining $\underline{\underline{\tilde{K}}} = \underline{\underline{b}}^T \cdot \underline{\underline{\tilde{I}}} \cdot \underline{\underline{a}}$),

$$\underline{\underline{b}}^T \cdot \underline{\underline{\tilde{D}}} \cdot \underline{\underline{a}} \cdot \underline{\tilde{x}} = \tilde{\lambda} \underline{\underline{b}}^T \cdot \underline{\underline{\tilde{I}}} \cdot \underline{\underline{a}} \cdot \underline{\tilde{x}}$$

and

$$\underline{\tilde{y}}^T \cdot \underline{\underline{b}}^T \cdot \underline{\underline{\tilde{D}}} \cdot \underline{\underline{a}} = \tilde{\lambda} \underline{\tilde{y}}^T \cdot \underline{\underline{b}}^T \cdot \underline{\underline{\tilde{I}}} \cdot \underline{\underline{a}}$$

or

$$\underline{\underline{\tilde{H}}} \cdot \underline{\tilde{x}} = \tilde{\lambda} \underline{\underline{\tilde{K}}} \cdot \underline{\tilde{x}}$$

and

$$\underline{\underline{\tilde{H}}}^T \cdot \underline{\tilde{y}} = \tilde{\lambda}^* \underline{\underline{\tilde{K}}}^T \cdot \underline{\tilde{y}}$$

In comparison, we see that the original pair of ordinary eigenvalue problems $H\bar{u} = \lambda \bar{u}$ and $H^+\bar{v} = \lambda \bar{v}$ has by these approximations been transformed into a pair of generalized eigenvalue problems, since there is no reason to expect $\underline{\underline{b}}^T \cdot \underline{\underline{\tilde{I}}} \cdot \underline{\underline{a}}$ to be the identity matrix.

These equations can behave even more poorly than those for the reduced Roussopoulos functional, even when I = J, because there potentially can be valid solutions for any value of $\tilde{\lambda}$ at all. This is terrible, because the whole point of the reduction is to obtain approximations $U(\tilde{x})$, $V(\tilde{y})$, and $\tilde{\lambda}$ to the exact solutions $\bar{u}$, $\bar{v}$ and λ of the more difficult exact problem. Assuming I = J, we see that one way the problem may arise is if there is some $\hat{x}$ such that $H \cdot \hat{x} = 0$ and $K \cdot \hat{x} = 0$, i.e. if the null spaces of $\underline{\underline{\tilde{H}}}$ and $\underline{\underline{\tilde{K}}}$ overlap. In this case an arbitrary multiple of $\underline{\hat{x}}$ can be added to any valid $\underline{\tilde{x}}$ (rendering it non-unique), and $\underline{\hat{x}}$ itself is a valid solution for any choice of $\tilde{\lambda}$. (Similarly there will be a $\underline{\hat{y}}$ if the null spaces of $\underline{\underline{\tilde{H}}}^T$ and $\underline{\underline{\tilde{K}}}^T$ overlap). As before, an example of this behavior occurs in the case of trial function spaces composed of non-corresponding sets of eigenvectors.

Fortunately, experience indicates that these degenerate cases do not occur very often in practice, so we make the working assumption that the

reduced Euler equations can be solved. When this "anomalous behavior" has occurred, it has usually given clearly very bad approximations [50].

An idea which comes to mind when generating approximate eigenvectors is to use the Rayleigh Principle with $U(\tilde{x})$ and $V(\tilde{y})$ to obtain a direct estimate of λ which might be better than the indirect estimate $\tilde{\lambda}$. Since

$$R[v,u] = \frac{\langle v, Hu \rangle}{\langle v, u \rangle} \quad ,$$

$$R[V(\tilde{y}), U(\tilde{x})] = \frac{\langle V(\tilde{y}), HU(\tilde{x}) \rangle}{\langle V(\tilde{y}), U(\tilde{x}) \rangle} = \frac{\langle \underline{v} \cdot \underline{\underline{b}} \cdot \tilde{y}, H\underline{u} \cdot \underline{\underline{a}} \cdot \tilde{x} \rangle}{\langle \underline{v} \cdot \underline{\underline{b}} \cdot \tilde{y}, \underline{u} \cdot \underline{\underline{a}} \cdot \tilde{x} \rangle}$$

$$= \frac{\tilde{y}^T \cdot \underline{\underline{b}}^T \cdot \langle \underline{v}, H\underline{u} \rangle \cdot \underline{\underline{a}} \cdot \tilde{x}}{\tilde{y}^T \cdot \underline{\underline{b}}^T \cdot \langle \underline{v}, \underline{u} \rangle \cdot \underline{\underline{a}} \quad \tilde{x}} = \frac{\tilde{y}^T \cdot \tilde{\underline{\underline{H}}} \cdot \tilde{x}}{\tilde{y}^T \cdot \tilde{\underline{\underline{K}}} \cdot \tilde{x}}$$

$$= \tilde{\lambda} \tilde{y}^T \cdot \tilde{\underline{\underline{K}}} \cdot \tilde{x} / \tilde{y} \cdot \tilde{\underline{\underline{K}}} \cdot \tilde{x} = \tilde{\lambda}.$$

Thus the Rayleigh Principle will not provide a better estimate of an eigenvalue than that calculated while solving the reduced Euler equations of $E[v,u,\lambda]$.

A similar situation arises in the use of $R[v,u]$, equal to $\langle z,\bar{u} \rangle$ at its stationary point, to evaluate approximations $V(\tilde{y})$ and $U(\tilde{x})$. It turns out that $R[V(\tilde{y}), U(\tilde{x})]$ is equal to $\langle z, U(\tilde{x}) \rangle$; a variational functional cannot be used to improve this sort of evaluation when the trial function has been determined by applying the Variational Approximation technique to the same functional [19,20].

Weighted Residuals

Having completed this analysis of the pitfalls of the Variational Approximation technique, it is interesting to consider the relation of

the variational method to the method of Weighted Residuals [21, 22, 23]. The WR method is motivated and justified by the argument that a good approximation to the solution of $H\tilde{u} = s$ can be determined by requiring that when the residual source $s - H\tilde{u}$ is weighted with J different functions measuring some kind of source "importance", each such weight should be zero:

$$\left\langle w_j,\ s - HU(\tilde{x}) \right\rangle = 0 \qquad \forall\ j$$

This is easily recognized as the variational condition obtained from a reduced Roussopoulos functional when $u = U(\tilde{x})$ and $V(\tilde{y})$ is chosen so that $V_j'(\tilde{y}) = w_j$; the ancient and honorable Weighted Residual method is seen to be analyzable as a particular form of Variational Approximation.

Numerical Example

A genuine example might be appropriate at this point as a reminder that the goal here is to develop practical approximation methods for real numerical problems. We will use the Least Squares functional to calculate approximate solutions for the Milne Problem: the distribution of neutrons in angle and space in a semi-infinite source-free half-space with an asymptotic source at infinity.

The mathematical statement of this problem is an equation with boundary conditions for $\psi(x,\mu)$:

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

(where 1-c is the absorption probability)

and $\psi(0,\mu) = 0$ for $0 \leq \mu \leq 1$

and $\psi(x,\mu) \to \frac{c\nu_0}{2}\frac{1}{(\nu_0+\mu)}\exp(x/\nu_0)$ as $x\to\infty$

(where ν_0 is the positive solution of $1 = c\nu_0 \tanh^{-1}\frac{1}{\nu_0}$)

It can be demonstrated [24] that the solution of this problem may be written

$$\psi(x, \mu) = \frac{c\ \nu_0\ \exp(x/\nu_0)}{2\ (\nu_0 + \mu\ )} + a\ \frac{c\ \nu_0\ \exp(-x/\nu_0)}{2\ (\nu_0 - \mu\ )} + \int_0^1 A(\nu)\ \phi_\nu(\nu)\ e^{-x/\nu}\ d\nu$$

where the $\Phi_\nu(\mu)$ are the singular eigenfunctions

$$\phi_\nu(\mu) \equiv \frac{c\nu}{2} P[\frac{1}{\nu-\mu}] + \delta(\nu-\mu)\ [1-\nu c\ \tanh^{-1}\nu]$$

and where the scalar "a" and the expansion coefficient function $A(\nu)$ can be determined from the boundary condition that

$$0 = \frac{c\nu_0}{2}\ (\nu_0+\mu)^{-1} + a\ \frac{c\nu_0}{2}\ (\nu_0-\mu)^{-1} + \int_0^1 A(\nu)\phi_\nu(\mu)d\nu$$

Explicit formulas for a and $A(\nu)$ can be found by the application of the singular eigenfunctions half-range orthogonality relations, but both the discovery of these relations and the evaluation of the resulting expressions are difficult processes. We take the attitude that an approximate solution for $A(\nu)$ will be satisfactory, and more specifically that a polynomial approximation is desired. (Polynomials are particularly suitable, because the integral of $\phi_\nu(\mu)$ times a polynomial is very easy to perform analytically, and it is the difficulty of handling the ϕ_ν that makes the exact problem so difficult to solve). Note that theoretically $A(\nu)$ could be expanded in a polynomial of arbitrarily high degree and therefore have an arbitrarily small error with respect to the exact function.

The values of a and the function A(v) can be specified as the stationary point of the Least Squares functional

$$L[b,B] = \int_0^1 [\phi_{0-}\ (\mu) + b\phi_{0+}(\mu) + \int_0^1 B(\nu)\phi_\nu(\mu)d\nu]^2 d\mu$$

The approximate solution is found by restricting $B(\nu)$ {= polynomial with parameters C_1, C_2, ...C_n} and integrating out of L first the ν dependence and then the μ dependence, to obtain a reduced functional $L[b,c_1,..c_n]$. Notice that L has been reduced to an ordinary function (a common occurance in practice) so that the stationary point in the reduced space can be found simply by setting the ordinary derivatives equal to zero.

This was done for several cases involving different types of polynomials, with typical results shown in Figure 1. The trial function used for $B(\nu)$ had only four degrees of freedom, so finding the stationary point required the solution of only a five-by-five system of linear equations. The function actually used for the initial series of calculations was a quartic polynomial in $(1-\nu)$, since it was known that $A(1) = 0$.

For cases with high absorption (small values of c) the use of quartics, cubics, etc. as trials does not lead to good approximation, because of the difficulty of representing the "corner" which appears in the true solution when ν is near one. To avoid this problem, the functional was reduced again over a trial function space consisting of natural cubic splines with n joints. [Cubic splines [25] are piecewise continuous cubic polynomials with piecewise continuous first and second derivatives. Discontinuities are allowed in the third derivative at n different points known as the joints; the term "natural" implies that the second and third derivatives are zero outside the region between the smallest and largest joints. These functions have n degrees of freedom, where n can easily be changed by adding or removing joints. They are smooth but flexible, and they are capable or representing strangely shaped curves, because the joints may be arbitrarily placed.] The five-joint spline approximation

(with four degrees of freedom since $A(1) = 0$) plotted in figure 1 is noticeably better than the quartic,giving weight to a recommendation to use splines wherever polynomial approximations are desired.

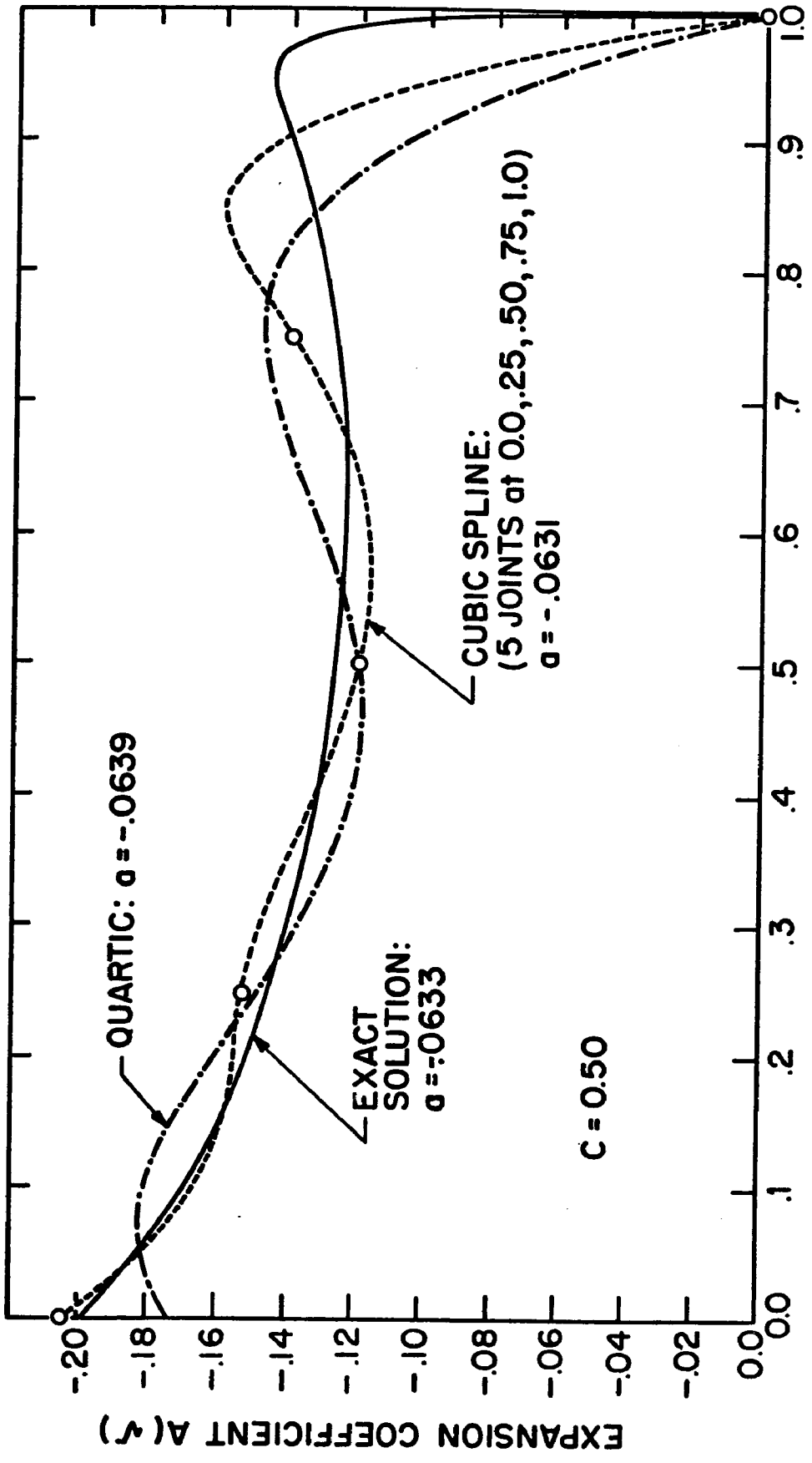


Figure 1. Milne Problem Eigenfunction Expansion Coefficient

CHAPTER III

ITERATIVE METHODS

Since so little can be proven about the accuracy of the Variational Approximation Principle there is good cause for interest in a procedure which is (almost) guaranteed to generate good approximate solutions. Such a procedure results from the iterative use of a variational functional whose value is a second order estimate of its stationary function.

The goal then is a functional which will provide

$$u^{(1)}(p) = F[u_0(r),p] = \bar{u} + ||\delta u|| \cdot E[\bar{u},\delta u]$$

where $E \rightarrow 0$ when $\|\delta u\| = \|u - \bar{u}\| \rightarrow 0$.

(Notice for later reference that this functional is explicitly an ordinary function of the new independent parameter p.)

If such a "Bootstrap" functional could be found, then it would be possible to pick some initial approximation u_0 to $\bar{u}$, determine $u^{(1)} = B[u_0]$, then $u^{(2)} = B[u^{(1)}]$, etc. until $u^{(N)} = B[u^{(N-1)}]$ was deemed a satisfactory point to stop the iteration. Reliance on the appropriate choice of restricted trial spaces would not be needed.

Bootstrap Functionals

Such a functional does exist. Symbolically using the Dirac Delta function to represent the sequence of functions with the property that $\lim_n \langle v_n^p, u \rangle = u(p)$, we can set $z(r) = \delta(p-r)$ in the Roussopoulos functional and define the result as $B[v(r,p),u(r)] = \langle \delta_p, u \rangle + \langle v(p), s \rangle - \langle v(p), Hu \rangle$ with the Euler equations

$$H\,\bar{u} = s \text{ and } H^+\,\bar{v}(p) = \delta_p.$$

For $u_0 = \bar{u} + \delta u$ and $v(p) = \bar{v}(p) + \delta v(p)$,

$$B[v(p),u_0] = \bar{u}(p) - \langle \delta v(p), H\,\delta u \rangle = u^{(1)}$$

so that if $\|\delta u\| \sim \eta$ and $\|\delta w(p)\| \sim \epsilon$ then

$$u^{(1)} = \bar{u} + \text{terms of order } (\epsilon\eta).$$

At the stationary point

$$-\delta^2 B[\bar{v}(p),\bar{u}] = \langle \delta v(p),\ H\ \delta u \rangle$$

so B cannot provide an extremal estimate of $\bar{u}$ unless H is (positive or negative) definite and $v(p)-\bar{v}(p) = u - \bar{u}$. This latter condition will only be met, however, when $H^+ = H$, $z = s$, and $v = u$, and so we must be satisfied with a second-order error of unknown sign in our improved approximation.

Green's Functions

The Bootstrapping functional requires two arguments, a trial for the desired function and a trial for some sort of adjoint which incorporates an extra free parameter. We will see that this adjoint is the Green's function [12] for the problem $H\bar{u} = s$. The Green's function is defined as the contribution to u(p) of a unit point source at r; i.e. g(r,p) is that function for which $u(p) = \langle g(r,p),\ s(r) \rangle$ is true. To find an equation for g(r,p), we substitute $s(r) = H\ \bar{u}(r)$:

$$\bar{u}(p) = \langle g(r,p),\ H\ \bar{u}(r) \rangle = \langle H^+ g(r,p),\ \bar{u}(r) \rangle$$

which implies that $H^+\ g(r,p) = \delta(p-r)$, which is the defining equation for the adjoint v(p).

The problem now is to generate the initial trial functions. When using the Variational Approximation Principle we were using trials with unknown parameters to be fixed later, but now we need good approximations

to start with (because if the errors ϵ and η are not small, "second order" processes may actually increase them). We shall assume that the Green's function trials to be used are formally the solutions of the equation $K^+ v(r,p) = \delta(p-r)$ (although these solutions do not have to be explicitly calculated). Here K is assumed to be an operator which approximates H but which is simple enough to allow solution of problems like $K w = q$.

This might look a little strange, trading all the freedom in picking $v(r,p)$ trials for the choice of K, but it really is an adaptation to the practical problem of generating <u>good</u> approximate Green's functions. By assuming $K^+ v(r,p) = \delta(p-r)$ we ensure that all the adjoint trials are consistent, and by assuming K approximates H we hope the overall error with respect to $H^+ g(r,p) = \delta(p-r)$ will be small. We will see that we can reformulate the variational functional so that the adjoint trials themselves never appear, being replaced by K^{-1}, so that the ability to solve $K w = q$ is a practical necessity. (Finally, we note that an operator which satisfies the requirements for K should not be hard to find, because the trial u_0 presumably is generated precisely by the procedure of finding some operator $\tilde{H}$ which approximates H and then solving the equation $\tilde{H} u_0 = s$.)

Inserting now the assumption that $K^{+\,-1}\,\delta(p-r) = v(r,p)$, we find

$$B[v(p),u_0] = \langle \delta_p, u_0 \rangle + \langle K^{+\,-1}\delta_p, s \rangle - \langle K^{+\,-1}\delta_p, H\, u_0 \rangle$$

$$B[H_0,u_0,p] = u_0(p) + \langle \delta_p, K^{-1}(s - H\, u_0) \rangle = u^{(1)}(p)$$

or $$u^{(1)} = u_0 + K^{-1}\,(s - H\, u_0).$$

Higher Order Methods

The notation $u^{(1)}$ has been used for this improved approximation because of the obvious capability to iterate the improvement process [26]. We define

$$u^{(n)}(p) \equiv B[v(r,p),\ u^{(n-1)}(r)]$$

so that if the error in $u^{(n-1)}$ was of order $(\epsilon^{n-1}\eta)$, where $\epsilon = \|v-\bar{v}\|$ and $\eta = \|u_0-\bar{u}\|$, then the error in $u^{(n)}$ is of order $(\epsilon^n \cdot \eta)$.

Using the K form of the Bootstrap functional, we see $u^{(n)} = u^{(n-1)} + K^{-1}(s - Hu^{(n-1)})$ and so by induction $u^{(n)} = u_0 + \sum_{m=1}^{n} u_m$, where we define $u_m \equiv K^{-1}(s - Hu^{(m-1)})$. Recognizing that $s = H\bar{u}$, and $H = K + (H-K)$, we see that $u_m = [I + K^{-1}(H-K)]\ (\bar{u} - u^{(m-1)})$. Subtracting u_m from u_{m+1}, we derive a recursion relation:

$$u_{m+1} = -K^{-1}(H-K)\ u_m$$

and by induction get the relation $u_m = (-1)^m\ [K^{-1}(H-K)]^m u_0$. This provides a direct expression for u(n) in terms of u_0:

$$u^{(n)} = u_0 + \sum_{m=1}^{n} (-1)^m\ [K^{-1}(H-K)]^m\ u_0$$

which presumably will converge to some function $\hat{u}$ as n approaches infinity if

$$||K^{-1}(H-K)|| = ||K^{-1}H-I|| < 1$$

i.e. if K^{-1} is sufficiently close to H^{-1}. We will return in Chapter IV to consider the fact that this expression for u(n), the approximation to $\tilde{u}$ iteratively improved by a variational functional, strongly resembles the Perturbation Theory expansion of $\bar{u}$.

Compound Iteration

First, however, we have one more item to explore. The error in $u^{(n)}$ was seen to be of order $(\epsilon^n \cdot \eta)$, where $(u_0 - \bar{u})$ was of order η and $(v(r,p) - g(r,p))$ was of order ϵ. The iterative error reduction is thus determined by the goodness of the adjoint trials. This suggests trying to "compound" the improvement [10] by calculating better v trials as well as u trials on each iteration.

A potential method, written symbolically with Delta functions again, would work like this:

$$BB[v,w,p,t] = \langle \delta(p-r), w(r,t) \rangle + \langle v(r,p), \delta(t-r) \rangle - \langle v(r,p), Hw(r,t) \rangle$$

has the Euler equations

$$H\,\bar{w}(r,t) = \delta(t-r) \quad \text{and} \quad H^+\,\bar{v}(r,p) = \delta(p-r)$$

and evaluates at its stationary point to

$$\begin{aligned} w(p,t) &= \langle \delta(p-r), \bar{w}(r,t) \rangle = \langle H^+\,\bar{v}(r,p),\ \bar{w}(r,t) \rangle \\ &= \langle \bar{v}(r,p),\ H\,\bar{w}(r,t) \rangle = \langle \bar{v}(r,p),\ \delta(t-r) \rangle = \bar{v}(t,p) \end{aligned}$$

and so we see that

$$g(t,p) = \bar{v}(t,p) = \bar{w}(p,t) = BB[g,g].$$

Using BB[v,v] we can generate improved Green's function trials up to any desired order:

$$g^{(n)} = BB[g^{(n-1)},\ g^{(n-1)}] = 2g^{(n-1)} - \langle g^{(n-1)},\ Hg^{(n-1)} \rangle$$

and since $\delta g^{(n)}$, the error in $g^{(n)}$, is of order $\|\delta g^{(n-1)}\| \cdot \|\delta g^{(n-1)}\|$ we see that $\delta g^{(n)}$ is of order $\|\delta g_0\|^{2^n}$, which represents a very rapid error reduction. At any point we can stop and recover the function

$u_g^{(n)} \equiv \langle g^{(n)}, s \rangle$ which is an approximation to $\bar{u}$ with error of order $\| \delta g_o \|^{2^n}$.

The problem with this scheme of compound improvement (there must be a problem since the results look so good) is that each $g^{(n)}$ is a function of two independent variables, and the work involved in calculating $g^{(n)}(r,p)$ for every r and every p from $\langle g^{(n-1)}(t,r), H g^{(n-1)}(p,t) \rangle$ is too enormous to contemplate for a realistic problem.

Perturbation Theory

The iterative nature of these schemes to reduce the error in approximate functions (and also the form of the equations they involve) is reminiscent of the treatment of operators with small perturbations. It will be useful to digress somewhat here to develop the perturbation theory properly [12] so that a comparison with variational methods can be made.

The problem to be solved is $H u = s$, as always. We assume that there is some "simple" operator K which approximates H, and propose to find a series expansion of u whose terms can be calculated by solving equations no harder than $K w = q$.

The easiest way to generate such an expansion (and the least rigorous) is to expand in a power series the effect upon the solution of $K u_o = s$ of a small perturbation to K. In particular, we assume that the perturbation is $\epsilon(H-K)$, where ϵ is some small scalar, and we expand the solution of

$[K + \epsilon(H-K)]\, u(\epsilon) = s$ in powers of ϵ:

$$u(\epsilon) = u_o + \epsilon u_1 + \epsilon^2 u_2 + \ldots = u_o + \sum_{m=1}^{\infty} \epsilon^m u_m$$

with the consequence that

$$K u_o + \epsilon[K u_1 + (H-K)u_o] + \epsilon^2[K u_2 + (H-K)u_1] + \ldots = s$$

This equation is valid for all values of ϵ, so we see that $Ku_0 = s$ as assumed, and that

$$K u_1 = -(H-K) u_0,$$

$$K u_2 = -(H-K) u_1, \text{ etc.}$$

so $K u_m = -(H-K) u_{m-1}$

which implies that $u_m = (-1)^m[K^{-1}(H-K)]^m u_0$. The infinite expansion for $u(\epsilon)$ can now be written

$$u(\epsilon) = u_0 + \sum_{m=1}^{\infty} (-\epsilon)^m [K^{-1}(H-K)]^m u_0.$$

Of course, this expansion is invalid if it is not convergent. Convergence will be assured if $\| \epsilon K^{-1}(H-K) \| < 1$; if this is true for $\epsilon = 1$, i.e. $\| K^{-1}H - I \| \equiv \xi < 1$, then we have found the required expansion for $\bar{u}$:

$$\bar{u} = u_0 + \sum_{m=1}^{\infty} (-1)^m [K^{-1}(H-K)]^m u_0.$$

We can define an approximation $u(n)$ to $\bar{u}$ by truncating this expansion at $m=n$:

$$u^{(n)} = u_0 + \sum_{m=1}^{n} (-1)^m [K^{-1}(H-K)]^m u_0$$

and recognize that this is the iterated variational approximation to $\bar{u}$, as suspected. Note that we can now express the error in $u^{(n)}$ in terms of the neglected terms of the infinite series: $\| u^{(n)} - \bar{u} \|$ is of order $\| K^{-1}H - I \|^{n+1} = \xi^{n+1}$.

Successive Approximation

A version of perturbation theory which is better adapted to practical use than the infinite series presented above is the method of Successive

Approximations. The name is derived from the procedure of solving a series of equations of the form $Kw = q$, where each step gives a better approximation to $\bar{u}$ than the last [12,28].

If $H\bar{u} = s$ and $K \approx H$ is easy to work with, consider $H\bar{u} = (K + H-K)\bar{u} = K[I + K^{-1}(H-K)]\bar{u} = K(I - T)\bar{u} = s$, where we introduce $T \equiv -K^{-1}(H-K) = (I - K^{-1}H)$. Now defining $u_0 = K^{-1}s$ also, we have $(I - T)\bar{u} = K^{-1}s = u_0$, or $\bar{u} = u_0 + Tu$. To get an approximate solution, we assume $T\bar{u} \approx Tu_0$, and define $u^{(1)} = u_0 + Tu_0 \approx \bar{u}$. Generalizing this process, we let $u^{(n)} \equiv u_0 + Tu^{(n-1)}$ to get the sequence of approximations that was sought. Note that the operator T is not really required; in practice $u^{(n)}$ is formed by solving the equivalent equation $Ku^{(n)} = s + (K-H)\,u^{(n-1)}$ as was proposed.

But does this sequence really converge to $\bar{u}$? Defining $u_n = u^{(n)} - u^{(n-1)}$, we see that $Ku_n \equiv Ku^{(n)} - Ku^{(n-1)} = (K-H)u^{(n-1)} - (K-H)u^{(n-2)} = (K-H)u_{n-1}$ or $u_n = -K^{-1}(H-K)\,u_{n-1} = (-1)^n[K^{-1}(H-K)]^n u_0$. Thus the Successive Approximation method is equivalent to the ϵ-expansion method, and converges, just as it does, when $\|K^{-1}H-I\| < 1$.

Formal Derivation

The perturbation theory formulas developed above by arguments about the successive terms of an approximation can be derived in a more satisfying (and rigorous) manner from consideration of the formal inverse of an operator which is "close" to the identity. This analysis also has the benefit of suggesting another "perturbation" expansion with much faster convergence properties.

The problem we wish to solve is $H\bar{u}=s$, and the solution is formally $\bar{u} = H^{-1}s$; the problem is to compute H^{-1}. If we rewrite the equation as $H\bar{u} = [K + (H-K)]\,\bar{u} = s = K[I + K^{-1}(H-K)]\bar{u}$, we can then formally invert

the K operator to get

$$[I+K^{-1}(H-K)]\,\bar{u} = K^{-1}s = (I-T)\,\bar{u}$$

(where we have defined $T = -K^{-1}(H-K) = F - K^{-1}H$). Now the solution can be written $\bar{u} = (I - T)^{-1}K^{-1}s$. However, $(I-T)^{-1}$ can be expanded, provided $\|T\| < 1$, as $\sum_{m=0}^{\infty} T^m$, (the Neumann expansion) so that

$$\bar{u} = K^{-1}s + \sum_{m=1}^{\infty} T^m K^{-1}s = u_o + \sum_{m=1}^{\infty} (-1)^m\,[K^{-1}(H-K)]^m\,u_o$$

As indicated, we expected this to be the usual perturbation expansion and it is, but analysis in terms of $(I-T)^{-1}$ has raised the interesting possibility of developing a new expression for $\bar{u}$ based upon the <u>product</u> expansion of $(I-T)^{-1}$ rather than the Neumann expansion.

Although the latter is well known, the equivalent expansions

$$(I-T)^{-1} = (I+T)\,(I+T^2)\,(I+T^4)\,\ldots$$
$$(I-T)^{-1} = (I+T+T^2)\,(I+T^3+T^6)\,(I+T^9+T^{18})\,\ldots$$

etc. appear to be not as well appreciated. In fact, there is a generalized product expansion

$$(I-T)^{-1} = \prod_{m=1}^{\infty} \left(\sum_{K=1}^{P} T^{(K-1)P^m}\right) \text{ for any } P \geq 2$$

with the property that the partial product of the first n factors is equal to the partial sum of the first term through the term (p^n-1) of the summation expansion. Clearly even the order-2 product expansion reaches greater accuracy than the (usual successive approximation) series expansion very quickly. Therefore we define the approximations $u_o \equiv K^{-1}s$ and $u^{(p,n)} \equiv \sum_{K=1}^{P} T^{(K-1)P^n} u_o$; these have errors of order $\|T\|^{P^n}$. This

method, called the "P-Hyperpower" method, has recently been proposed for use in reactor analysis by Devooght [27]. It is capable of dramatic accuracy, but for a number of good reasons it is only rarely used.

The problems with applying the hyperpower method are practical, not theoretical. First, powers of T must be calculated and applied, which can involve great amounts of work. Second and more important, the actual inverse of K must be found, since the fast convergence of the hyperpower method depends on the explicit application of T^{p^m}, where $T = -K^{-1}(H-K)$. The numerical calculation of matrix inverses is susceptible to errors and is expensive, so a significant advantage of the normal perturbation method is that K only appears in equations of the form $K\,w = q$ which can be solved for w without inverting K. For these reasons, the hyperpower method is seldom used.

Operators with Parameters

All of these methods for forming high-order approximations implicitly assume that the base operator K is chosen and fixed and that the desired degree of accuracy is obtained by iterating until $\left\| I-K^{-1}H \right\|^{n}$ is sufficiently small. There is, however, no reason why K cannot be chosen as an operator containing free parameters [28]. These parameters (call them "x")will also affect the approximate solution $u^{(n)}_{(x)}$ and can be chosen so as to make $K_{(x)}\,w = q$ easy to solve or, more likely, to reduce the potential error by minimizing $\left\| K^{-1}_{(x)}\,(H-K_{(x)}) \right\|$.

A widely used iterative method which incorporates a free parameter is the Successive Over-Relaxation method [32] for the solution of sets of linear equations. The relaxation methods are perhaps not thought of as perturbation expansions, but they are in fact examples of the success-

ive approximation formulation. To see their perturbative nature explicitly, and to show how the parameter is introduced, we start with the matrix equation $H\,u = (B-R-R^T)u = s$. (Where typically H is a finite difference approximation of the Laplacian).

The simplest relaxation method is Simultaneous Relaxation: take K=B so that $T=B^{-1}(R+ R^T)$. B is assumed to be relatively easy to invert, hence we can solve for $u^{(n)} = B^{-1}s + B^{-1}(R+R^T)u^{(n-1)}$. We note, however, that if R is a lower triangular matrix and if the vector $u^{(n)}$ is evaluated from the top down, then $R\,u^{(n-1)}$ uses only those elements of $u^{(n-1)}$ which have already been "improved", and we could compute $u_{SR}^{(n)} = B^{-1}s + B^{-1}R\,u^{(n)} + B^{-1}R^T u^{(n-1)}$ with presumably greater iterative improvement. This is called Successive Relaxation, and can be seen to be an ordinary iteration with $K = B-R$ and thus $T = (B-R)^{-1}R^T = (I-B^{-1}R)^{-1}B^{-1}R^T$.

The parameter is introduced now in an attempt to accelerate the convergence of u(n):

$$u_{sor}^{(n)} = u^{(n-1)} + \omega(B^{-1}(s+Ru^{(n)}+R^T u^{(n-1)}) - u^{(n-1)}]$$

or

$$u_{sor}^{(n)} = \omega B^{-1}s + \omega B^{-1}Ru^{(n)} + [(1-\omega)I+\omega B^{-1}R^T]u^{(n-1)}$$

This scheme (Successive Over-Relaxation) is an iteration with $K=(\frac{1}{\omega}B-R)$, and since the intention is to iterate to convergence we want to choose ω so as to minimize the norm of $T = \left\| (B-\omega R)^{-1}\,((1-\omega)\,B+\omega R^T) \right\|$. For some problems (as when H is the finite-difference diffusion operator) this norm can be evaluated analytically and formulas for the optimum ω obtained.

Fixing the Parameters

In general, what sort of freedom can be allowed? Linear combinations suggest themselves immediately, so we try the form $K(x) = \sum_{\ell} x_\ell K_\ell$, where presumably each K_ℓ is itself an approximation to H. Although this would seem to be the simplest sort of expansion, it actually leads to great difficulties in trying to solve the prototype equation $K(x)\, w = \sum_{\ell} x_\ell K_\ell\, w = q$ since the inverse of a sum of terms is needed. This suggests the alternate approach of forming $K^{-1}(x) = \sum_{\ell} x_\ell K_\ell^{-1}$ directly. Since K itself is never used, only the solution of $Kw = q$, obtaining $K(x)^{-1}$ directly is definitely more useful than obtaining K(x). The greater difficulty of choosing trials K_ℓ^{-1} which approximate H^{-1} instead of K_ℓ which approximate H is offset by the flexibility which the free parameters incorporate in $K(x)^{-1}$. The degenerate kernel technique, using a kernel (inverse) which is a sum of simpler kernels (trial inverses) is an example of this method.

Having inverted the parameters, it becomes necessary to fix their values. Some particular choice $\tilde{x}$ must be made, but the criterion can depend on the application of $\bar{u}$. If the actual goal is to approximate $\bar{u}$ as closely as possible, then presumably the iterative procedure will be applied until the error is very small; choosing $\tilde{x}$ so as to minimize $\left\| I - K^{-1}(\tilde{x})H \right\|$ will minimize the potential work in getting to a satisfactory $u^{(n)}$, since the norm of the correction terms $\left\|u_m\right\| \leq \left\| K^{-1}(H-K) \right\| \cdot \left\|u_{m-1}\right\|$ will be decreased as rapidly as possible. On the other hand, it would not be unreasonable to ask for a small error after only a fixed small number of iterations.

Goldstein [29] attempts to achieve this by fixing $\tilde{x}$ so that when

some functional $L[u]$ is evaluated with the $u^{(n)}_{(x)}$, $L[u_{o(\tilde{x})}] = L[u^{(1)}_{(\tilde{x})}]$. The justification is analagous to that used with the ordinary variational method: $L[u_{o(\tilde{x})}]$ has only second order errors. To see this, recall that $u^{(n)}$ has errors of order $\|K^{-1}(H-K)\|^{n+1}$, so that $u^{(1)}$ has errors of order $\|K^{-1}(H-K)\|^{2}$. Then notice that since $L[u_{o(\tilde{x})}] = L[u^{(1)}_{(\tilde{x})}]$, the error is of order $\|K^{-1}(\tilde{x})\ (H-K(\tilde{x}))\|^{2}$. Of course, this does not guarantee that the error $L(u_{o(\tilde{x})}] - L[\bar{u}]$ will itself be small, only that it will be relatively smaller than the error $\left(u_{o(x)} - \bar{u}\right)$.

Assuming $L[u]$ is a linear functional written $L[u] = \langle z,u \rangle$, this method requires that

$$0 = \langle z, u^{(1)}_{(\tilde{x})} \rangle - \langle z, u_{o(\tilde{x})} \rangle = \langle z, u^{(1)}_{(\tilde{x})} - u_{o(\tilde{x})} \rangle$$

or

$$0 = \langle z,\ u_{1(\tilde{x})} \rangle$$

This is known as the "Zero Residual" method since $\langle z,\ u_{1(\tilde{x})} \rangle$ approximates $\langle z,\ \bar{u} - u_{o(\tilde{x})} \rangle$, the residual error in the functional evaluation. Since

$$u_{1(x)} = -K^{-1}_{(x)}\ (H - K_{(x)})\ u_{o(x)}$$

we have

$$0 = -\langle z,\ K^{-1}_{(\tilde{x})}\ (H - K_{(\tilde{x})})\ u_{o(\tilde{x})} \rangle$$

or

$$0 = \langle K^{+-1}_{(\tilde{x})}\ z,\ (H-K_{(\tilde{x})})\ u_{o(\tilde{x})} \rangle$$

so that if we think of $K^{+}(\tilde{x})\ v_{o(\tilde{x})} = z$ or defining an approximate adjoint analagous to $u_{o(\tilde{x})}$, this is a requirement that $(H-K)$ be small in the sense of contributions to $L[\bar{u}]$. With $\tilde{x}$ chosen this way, we see that $L[u_{o(\tilde{x})}] = L[u^{(1)}_{(\tilde{x})}]$, which has second order error, and we find an extra

bonus in that $L[u^{(2)}_{(\tilde{x})}]$, which has a third order error, takes a particularly simple form:

$$L[u^{(2)}_{(\tilde{x})}] = \quad z,K^{-1}HK^{-1}HK^{-1}s$$

This explanation of the Zero Residual method shows how to fix only one parameter $\tilde{x}$; if x represents a set of parameters x_i, more conditions are needed. There are two obvious ways of extending this procedure, one more elegant and the other more useful. The elegant method requires that $L[u_{o(\tilde{x})}] = L[u^{(n)}_{(\tilde{x})}]$ where n = 1 through I, the number of parameters. This set of conditions should be large enough to determine all the x_i, and implies that the evaluation of $L[u_{o(x)}]$ will have errors of order $\|K(H-K)\|^{I+1}$. The calculation of all the $u^{(n)}_{(x)}$ in terms of the free variables will lead, however, to impractically large systems of (probably non-linear) equations.

More useful is a procedure which specifies that $L_n[u_{o(\tilde{x})}] = L_n[u^{(1)}_{(\tilde{x})}]$ for n = 1 to I different functionals. This leaves the error at second order, but (hopefully) produces an approximation $u_{o(\tilde{x})}$ which can be used with confidence to evaluate a wide variety of functionals. An interesting point to notice here is that if the functionals are all linear of the form $\langle z_n, u \rangle$, then the definition $K(x)\, v_n(\tilde{x}) = z_n$ transforms the ZR equations into

$$\langle v_{n(\tilde{x})}, (H-K_{(\tilde{x})})u_{o(\tilde{x})} \rangle = 0 \qquad \forall\ n$$

or

$$\langle v_{n(\tilde{x})}, Hu_{o(\tilde{x})} - s \rangle = 0 \qquad \forall\ n$$

which bears a remarkable resemblance to the Variational Approximation Principle.

CHAPTER IV

PERTURBATIVE VS. VARIATIONAL METHODS

In the previous chapters we have noted on several occasions the similarity between the variational analysis and the perturbation analysis. We will formally analyze these similarities here, and show that the two methods are formally equivalent in many cases.

Formal Equivalence

We assume that the goal is an approximation to the function $\bar{u}$ which is the solution of the equation $H\bar{u} = s$. Further, we assume that the variational approximation will be formed by iterative application of the Bootstrap functional to some $u(r) = U(r,x) \approx u$(where x is a free parameter), while the perturbative approximation will be formed by successive approximations with the operator $K_p(x) \approx H$. Using the definition $T_p(x) = -K_p^{-1}(x) \cdot [H-K_p(x)]$, we recall that the perturbative expansion is

$$u_p^{(n)}(x) = u_p(x) + T_p(x)u_{p(x)}^{(n-1)} = u_p(x) + \sum_{m=1}^{n} u_m^p(x),$$

where

$$K_p(x)u_p(x) = s \text{ and } K_p(x)u_m^p(x) = -(H-K_p(x))u_{m-1}^p(x).$$

On the other hand, for the variational expansion we start with $u_v(x)$, which we assume can be generated by some operator such that $K_v(x)u_v(x) = s$, and we write the iterated approximation as

$$u_v^{(n)}(x) = u_v^{(n-1)}(x) + K_v^{-1}(x)[s-Hu_v^{(n-1)}(x)] = u_v(x) + \sum_{m=1}^{n} u_m^v(x),$$

where

$$K_v(x)u_v(x) = s \text{ and } K_v(x)u_m^v(x) = -(H-K_v(x))u_{m-1}^v(x).$$

Now we see clearly that $u^{(n)}_{v(x)} = u^{(n)}_{p(x)}$ for those situations in which the two approximate operators are equal, or effectively, when $K_p(x)u_v(x)=s$. Furthermore, consider the Roussopoulos functional which evaluates some goal functional L[u]:

$$R[v,u] = L[u] + \langle v,\ s-Hu \rangle \ .$$

The source for the adjoint problem for this functional is $z = \frac{\delta L}{\delta u}$, and if we assume that $K^+_v(x)\ \tilde{v} = z = \frac{\delta L}{\delta u}$ (for consistency with $u_o(x)$) we see that

$$\begin{aligned} R[\tilde{v},u^{(n-1)}(x)] &= L[u^{(n-1)}(x)] + \left\langle K_v^{+\,-1}(x)\ \frac{\delta L}{\delta u},\ s-Hu^{(n-1)}(x) \right\rangle \\ &= L[u^{(n-1)}(x)] + \left\langle \frac{\delta L}{\delta u}\ ,\ K_v^{-1}(x)\ (s-Hu^{(n-1)}(x)) \right\rangle \\ &= L[u^{(n-1)}(x)] + \left\langle \frac{\delta L}{\delta u}\ ,\ u_n(x) \right\rangle \end{aligned}$$

while the value of L[u] upon inserting the approximation directly is

$$L[u^{(n)}(x)] = L[u^{(n-1)}(x) + u_n(x)] = L[u^{(n-1)}(x)] + \left\langle \frac{\delta L}{\delta u}\ ,\ u_n(x) \right\rangle + \text{higher orders}$$

which is the same as $F[\tilde{v},u^{(n-1)}_{(x)}]$ in the instance that the goal functional L[u] is linear [26].

Thus not only are the high order variational and perturbative approximations to $\bar{u}$ formally equivalent, but also the direct perturbative evaluation of a linear goal functional is equivalent to the variational evaluation with an approximation of one lower order, i.e. $R[\tilde{v},u^{(n-1)}] = L[u^{(n)}]$. In particular, this applies to evaluations with second order error: L[u] evaluated directly with a first order perturbation approximation will be equivalent to the variational evaluation of L[u] using the (zeroth-order) un-iterated trial function, provided again that the perbation base operator and the trial functions satisfy the relation $K_v(x) \cdot U_p(x) = s$.

Practical Differences

Since second order evaluation is the most common type, we see that the accuracy of the results really does not depend on the nature of the method used - it depends on the quality of the trials, and the manner of fixing the parameter. Whether a particular method is called perturbative or variational will be determined by the order of the approximation and the manner in which the parameters are handled.

Low order methods involving no parameters or only a few tend to be considered perturbation methods, in the sense that they are used to calculate small effects when given very good initial trials. When the initial trials are not too good, high-order iteration is thought of as a perturbation process; whereas when the initial trials have a great many parameters the variational procedure fixes them all at once. The aim of the discussion in this section is to encourage consideration of the alternate formulation of any problem to make certain that good approximations are not overlooked because they are unusual.

Example

As an example of this technique of trying the alternate formalism, we introduce the Intermediate Resonance method of Goldstein [29]. This is a method for evaluating $L[\psi] = \int \sigma_a(u)\,\psi(u)du$, the absorption of neutrons by a resonance, when the flux $\psi(u)$ cannot be treated in the traditional Narrow Resonance (NR) or Narrow Resonance Infinite Mass (NRIM) approximations.

The energy dependence of the flux of neutrons slowing down in a medium consisting of a light moderator and a heavy resonance absorber is determined by the equation

$$(\Sigma_m + \Sigma_s(u) + \Sigma_a(u))\ \psi(u) - G(\Sigma_s\psi) = \Sigma_m$$

where Σ_m is the moderator cross section, and the scattering operator is

$$G(\Sigma_s\psi) = \int_{u-\ln 1/\alpha}^{u} \frac{-(u-u')}{(1-\alpha)}\ \Sigma_s(u')\ \psi(u')du'$$

In the NRIM limit (called the Wide Resonance limit by Goldstein) $G(\Sigma_s\psi)$ approaches $\Sigma_s(u)\psi(u)$ so that $\psi_{WR}(u) = \frac{\Sigma_m}{\Sigma_m+\Sigma_a}$, while in the Narrow Resonance limit $G(\Sigma_s\psi)$ approaches Σ_p, the scattering cross section of the absorber, so that $\psi_{NR}(u) = \frac{\Sigma_m+\Sigma_p}{\Sigma_m+\Sigma_a+\Sigma_s}$; Goldstein assumes that between these limits (the Intermediate Resonance region) the expression $\psi_{IR}^{\lambda}(u) = \frac{\Sigma_m+\lambda\Sigma_p}{\Sigma_m+\Sigma_a+\lambda\Sigma_s}$, which depends on λ as a free parameter, is a reasonable approximation to $\psi(u)$.

To get an improved approximation, however, he assumes there is an operator $K\phi$ which approximates $(\Sigma_m+\Sigma_s+\Sigma_a)\phi - G(\Sigma_s\phi)$ and which has the property that $\psi_{IR}^{\lambda}(u) = K^{-1}\Sigma_m$, i.e. that $\psi_{IR}^{\lambda}(u)$ is the first order perturbation approximation to $\psi(u)$ when iterating with K. The value of λ is then fixed by using the Zero Residual principle within the perturbative formalism: iterating (formally) to get a second order approximation $\tilde{\psi}_{IR}^{\lambda}$ and then requiring $L[\psi_{IR}^{\lambda}(u)] = L[\tilde{\psi}_{IR}^{\lambda}(u)]$. With this value, $L[\psi_{IR}^{\lambda}(u)]$ gives a second order estimate of the resonance absorption.

This has been a very successful method, due to the cleverness in introducing a parameter which interpolates between limiting cases and then using an iterative technique to boost the accuracy at the same time the parameter is fixed. Since its introduction the IR technique has been extended to allow treatment of material heterogeneity [30], multi-nuclide systems [31], etc. By its usefulness it justifies the transformation from

the trial function formulation (where the non-linear parameter would complicate a variational approximation) to the (seemingly) more flexible iterative approach.

Hybrid Applications

An interesting concept for getting the best out of both variational and perturbative methods comes from the following observation: the Successive Approximation form of the iterative expansion is phrased formally in terms of the inverse of K, but it was pointed out that in fact all that is required is the ability to solve systems $Kw = q$. These, however, are approximations to $H\tilde{w} = q$, and so we might propose to apply the Variational Approximation Technique to the functional $L[u] + \langle v, q-Hu \rangle$ and define K implicitly by defining w to be the stationary point $\tilde{w}$ in an appropriate reduced space. Assuming the reduced space is reasonably well chosen, each $\tilde{w}$ should be a good approximation to its $H^{-1}q$; so that the "effective" K should be a good approximation to H, and the Successive Approximations sequence should converge rapidly.

In effect, this is what has been done in some [33,34] of the applications of the Synthetic method of Kopp [35] (really a version of the Successive Approximations method written in a complicated form).

Kopp proposes to solve the problem $\bar{u} = A\bar{u} + s$, when a simpler operator $B \approx A$ is available, by the following cyclic process:
first take $w_1 = s$ and define $x_1 = \bar{u} - w_1$;
this implies $x_1 = A(x_1 + w_1)$, or $x_1 = (I - A)^{-1}A\, w_1$; now try $x_1 = y_1 + x_2$, where $y_1 = B(y_1 + w_1) \approx x_1$; which implies $x_2 = A\, s_2 + w_2$, where $w_2 = (A-B)\cdot(y_1 + w_1)$. The approximation to $\bar{u}$ accumulates as $u^{(n)} = \sum_{m=1}^{n}(w_m + y_m)$ where $y_m = B(y_m + w_m)$ and $w_{m+1} = (A-B)\cdot(y_m + w_m)$. We can greatly simplify

this, however, by defining $u_m = y_{m+1} + w_{m+1}$, which implies $u_m = (I-B)^{-1} \cdot (A-B)\, u_{m-1}$ with $u_0 = (I-B)^{-1}s$, so that u(n) is just the ordinary perturbation expansion for $\bar{u} = (I-A)^{-1}s$ when $K = (I-B)$.

In the applications noted above, the problem $(I-A)\bar{u} = s$ to be solved was the transport equation for the angular neutron flux distribution, while the most suitable choice for the simplified equation $(I-B)\, y = Bw$ was shown to be the neutron scalar flux diffusion equation. The diffusion equation, however, is a variational approximation to the transport equation, and although that was not the basis for its choice in these applications, it illustrates very well the usefulness of hybrid techniques.

CHAPTER V

NEUTRON FLUX DISTRIBUTION

The techniques for approximation presented up to this point have been formal methods with only a few examples. Now we shall start to apply them to the central problem of reactor physics, the calculation of the neutron flux distribution. As was pointed out in the Introduction, the equation governing this distribution is known but unsolvable in practice, and the goal will be to use the Variational Approximation Technique to eliminate unneeded complexity and to incorporate known detail. Variational methods, which are based on the accurate evaluation of some goal functional rather than the calculation of the flux itself, are very suitable for this problem because in fact the real quantities of interest are various weighted averages - functionals - of the unobservable flux.

Neutron Transport

We begin by introducing the definitions needed to state the exact problem. The unknown function for which we want to solve is the expected flux of neutrons $\phi(\underline{r}, E, \underline{\Omega}, t)\, d^3r dE d^2\Omega$ in a small d^3r about the point $\underline{r}$, traveling within a solid angle $d^2\Omega$ about the angle $\underline{\Omega}$, with energy in the range dE about the energy E, at time t. This flux (defined on a linear space with seven independent dimensions) satisfies the neutron transport equation, which describes both the motion of neutrons through space and their interaction with the material comprising the reactor core.

The transport equation can be written [24]:

$$\frac{1}{v}\frac{\partial\phi}{\partial t} + \underline{\Omega} \cdot \nabla\phi + \Sigma\phi = Q + \int dE' \int d^2\Omega'\ \Sigma_s\, \phi(\underline{r}, E', \underline{\Omega}', \underline{t})$$

with the following definitions:

v = velocity of a neutron with energy E

$\nabla\phi$ = gradient of the flux

$Q(\underline{r},E,\underline{\Omega},t)d^3rdEd^2\Omega$ = arbitrary source of neutrons

$\Sigma(\underline{r},E,\underline{\Omega},t)\phi d^3rdEd^2\Omega$ = rate of removal of neutrons from the elements $d^3rdEd^2\Omega$ at $(\underline{r},E,\underline{\Omega},t)$

$\Sigma_s(\underline{r},E'\rightarrow E,\underline{\Omega}'\rightarrow\underline{\Omega},t)\phi(\underline{r},E',\underline{\Omega}',t)d^3rdE'dEd^2\Omega'd^2\Omega$ = rate of scattering of neutrons from energy E' and angle $\underline{\Omega}'$ to energy E and angle $\underline{\Omega}$.

The neutron flux distribution after time t_0 within a region R with surface S is uniquely determined by the initial distribution in R at time t_0, the incident flux on the surface S, the internal source Q, and the transport equation written within R.

The difficulty in solving this equation derives primarily from the complex form of realistic cross sections. Since the nature and distribution of the interacting material is arbitrary, the cross section can change drastically over small regions in space, causing variations in the angular and spatial flux dependence which make the gradient term fluctuate. Furthermore, the energy dependence of the cross sections can vary wildly, impressing corresponding variations on the energy distribution of the flux. Various degrees of approximation will be made, either to ignore detail or extract it in advance, so that the remaining calculational problem provides only enough information to answer specific questions, rather than the complete solution of the exact equation.

Angular Reduction

In preparation for this, we want to express the angular dependence of the transport equation in a different manner [36]. For most materials the scattering cross section is a function of the angle between $\underline{\Omega}'$ and $\underline{\Omega}$ rather than each of these angles independently:

$$\Sigma_s = \frac{1}{2\pi} \cdot \Sigma_s(\underline{r}, E' \to E, \underline{\Omega}' \cdot \underline{\Omega}, t).$$

Before rewriting the scattering term with this dependence we also split Σ_s into $\Sigma_s(E')$, the total scattering cross section, and $F(E' \to E, \underline{\Omega}' \cdot \underline{\Omega})$, a conditional probability distribution, so that

$$\int dE' \int d^2\Omega' \Sigma_s \phi = \int dE' \int d^2\Omega' \phi(\underline{r}, E', \Omega', t) \frac{1}{2\pi} \Sigma_s(\underline{r}, E', t) F(\underline{r}, E' \to E, \underline{\Omega}' \cdot \underline{\Omega})$$

Now we expand the angular probability distribution in a complete set of Legendre polynomials:

$$F(\underline{r}, E' \to E, \underline{\Omega}' \cdot \underline{\Omega}) = \sum_{n=0}^{\infty} \frac{2n+1}{2} f_n(\underline{r}, E' \to E) P_n(\underline{\Omega}' \cdot \underline{\Omega}),$$

where

$$f_n(\underline{r}, E' \to E) \equiv \int_{-1}^{+1} d\mu \, P_n(\mu) F(\underline{r}, E' \to E, \mu),$$

so that

$$\int dE' \int d^2\Omega' \Sigma_s \phi = \sum_{n=0}^{\infty} \frac{2n+1}{2} \int dE' \Sigma_s(r, E', t) f_n(\underline{r}, E' \to E) \int d^2\Omega' \phi \cdot P_n(\underline{\Omega}' \cdot \underline{\Omega}).$$

This is still an exact representation, but in an alternate form.

We must, however, start eliminating parts of the exact equation, and this Legendre expansion was performed because most of the angular dependence will be the first to go. In many cases, particularly cases involving power reactors, the spatial and spectral dependence of the flux is most important, and the angular detail is sacrificed to allow better calculations in these other variables. The traditional way of

doing this has been to expand the angular flux into terms with known angular dependence and then use the Weighted Residual method to derive equations for the coefficients of these terms.

Thus, for example, the flux can be expanded in a complete set of spherical harmonics,

$$\phi(\underline{r},E,\underline{\Omega},t) = \frac{1}{4\pi}\phi_0(\underline{r},E,t) + \frac{3}{4\pi}\underline{\Omega}\cdot\underline{J}(\underline{r},E,t) + \ldots,$$

and the equations for the coefficients can be determined by setting equal to zero the integrals of the transport equation multiplied by the functions orthogonal to the spherical harmonics used in the expansion. This produces a representation of the exact flux as an infinite series of harmonics which is truncated to the degree of angular detail desired (giving the so-called P-N equations).

A better way of eliminating angular complexity is by using the Variational Approximation Technique. (This method is preferred because the procedure used to derive the approximate equations do not depend on the properties (e.g. orthogonality) of the trial functions, so that detail can be built in without increasing the angular "degrees of freedom"). We want to find a variational functional whose Euler equation is equivalent to the transport equation and its boundary condition. Given this, we restrict the trial function space by allowing only certain kinds of angular dependence and find the reduced Euler equation.

Rather than write down such a functional for the full transport equation, we shall illustrate the procedure as applied to the one-dimensional form of the transport equation in the one-speed approximation:

$$H\phi = Q = \mu\frac{\partial\phi}{\partial r} + \Sigma\phi - \sum_{n=0}^{\infty}\frac{2n+1}{2}\Sigma_s f_n P_n(\mu)\int_{-1}^{1} d\mu'\, P_n(\mu')\phi(\underline{r},\mu').$$

Here the flux is a function only of the position r and the cosine μ of the angle relative to the r-axis. Σ , Σ_s and f_n are functions of r only. The term $P(\Omega' \cdot \Omega)$ has been transformed (using the addition theorem for spherical harmonics and integrating over $d\phi$) into the product $P(\mu') \cdot P(\mu)$.

A suitable functional for this equation is

$$F[\phi^+,\phi] + \int dr \int d\mu [Q^+\phi + \phi^+(Q-H\phi)],$$

the Roussopoulos functional, where we see that an adjoint function satisfying $H^+ \phi^+ = Q^+$ must be introduced. We have avoided mentioning the boundary conditions to be associated with this flux equation because that topic brings up a lot of confusion without much compensating insight. Briefly, the boundary conditions can be treated either by requiring all trial functions to satisfy them or else by including them with extra Lagrange multipliers in the functional.

An approximate solution is obtained by assuming expansions for ϕ and ϕ^+. Pomraning [37] effectively uses

$$\phi(r,\mu) = \sum_{m=0}^{N} \frac{2m+1}{2} \phi_m(r) P_m(\mu)$$

and similarly

$$\phi^+(r,\mu) = \sum_{\ell=0}^{N} \frac{2\ell+1}{2} \phi_\ell^+(r) P_\ell(\mu)$$

and performs the angle integrals in F to obtain a reduced functional

$$F[\phi_0^+, \phi_1^+, \ldots \phi_N^+, \phi_0, \phi_1, \ldots \phi_N]$$

Setting

$$0 = \frac{\delta F}{\delta \phi_\ell^+} = \int d\mu \, P_\ell(\mu) [H \sum_{m=0}^{N} \phi_m(r) P_m(\mu) - Q] \qquad \forall \ \ell$$

is seen to provide as Euler equations the ordinary P_N equations. With N = 1, and defining $\Phi_0(r) = \frac{1}{2}\phi(r)$ and $\phi_1(r) = \frac{3}{2}J(r)$, we have

$$\frac{dJ(r)}{dr} + (\Sigma-\Sigma_s)\phi(r) = Q_0(r) \equiv \int d\mu P_0(\mu)\, Q(r,\mu)$$

and

$$(\Sigma-f_1\Sigma_s)J(r) + \frac{1}{3}\frac{d\phi(r)}{dr} = Q_1(r) \equiv \int d\mu P_1(\mu) Q(r,\mu),$$

showing that the common P-1 approximation (and also Diffusion Theory) can be thought of as variational approximations to the transport equation.

This really isn't too exciting. What is interesting is to note that any other sort of angular dependence can be used just as well to specify the restricted trial space, so that approximations tailored to particular known (or suspected) angular distributions can be easily generated. For example, Kaplan, Davis & Natelson [38] describe the use of angle-expansion functions which are "peaked" in certain directions in an attempt to build in, rather than calculate again and again, the knowledge that the flux will be strongly anisotropic in some regions.

Diffusion Theory Functional

We did point out, however, that in practical reactor problems the angular distribution is of secondary importance, and so we shall not pursue these clever angular syntheses any further. Instead, we shall adopt space and energy dependent diffusion approximations as the "exact" problem for the remainder of this work.

The Diffusion Approximation is developed by restricting the flux to a linear dependence on angle (the P-1 approximation) and assuming in addition that the external source terms are isotropic (independent of angle).

Writing the restricted form of the flux in terms of the scalar flux $\phi(\underline{r},E)$ and the net current vector $\underline{J}(\underline{r},E)$, we use the trial functions $\phi(\underline{r},E,\underline{\Omega}) = \phi(\underline{r},E) + \underline{\Omega}\cdot\underline{J}(\underline{r},E)$, in the eigenvalue functional (written for a Boltzmann equation with fission cross sections and a criticality eigenvector) to get $\nabla\cdot\underline{J} + (A - \lambda^{-1}M)\phi = 0$ and $\nabla\phi + D^{-1}\underline{J} = 0$ as the reduced Euler equations. The notation will be defined shortly, but notice that the second equation couples the scalar flux and the current together directly in a form of Fick's law. This coupled-equation form of the Diffusion approximation is known as the Canonical form; while the equation $\nabla\cdot D\nabla\phi + (A - \lambda^{-1}M)\phi = 0$, formed by eliminating the current, is known as the Diffusion equation. (The Diffusion equation features the ∇^2 diffusion term; the Canonical equations can be derived from it by the canonical transformation described earlier. Presumably there is an Involutory form also, but it is not nearly as useful as these).

Returning to the P-1 equations, we define the A, M, and D operators. Total removal and isotropic scattering are combined into

$$A\phi \equiv \Sigma(\underline{r},E)\phi(\underline{r},E) - \int dE'\Sigma_s(\underline{r},E'\rightarrow E)\phi(\underline{r},E')$$

The fission source, which is isotropic, is defined as

$$\frac{1}{\lambda}M\phi \equiv \frac{1}{\lambda}X(E)\int\nu(E')\Sigma_f(\underline{r},E')\phi(\underline{r},E'),$$

where X(E) is the fission spectrum and λ is the eigenvalue for which this homogeneous system has a solution. The current equation has terms for the total removal and anisotropic scattering (where Σ_{s1} is the linearly anisotropic component of the angle-dependent scattering).

$$D^{-1}\underline{J} \equiv 3\Sigma(\underline{r},E)\underline{J}(\underline{r},E)-3\int dE'\Sigma_{s1}(\underline{r},E'\rightarrow E)\underline{J}(\underline{r},E').$$

The choice of the notation $D^{-1}\underline{J}$ is obviously due to the fact that we will eventually want to solve for $\underline{J} = -(D^{-1})^{-1}\nabla\phi = -D\nabla\phi$. This convenient notation, however, is not meant to imply that finding the real $D = (D^{-1})^{-1}$ an easy task.

The question of adjoint operators will arise shortly, when we try to write a Diffusion Theory functional, so we may as well define them here. The inner product we will be using is defined as an integration, over all energies and over the reactor volume R, of the product of the argument functions:

$$\langle\phi^+,\phi\rangle \equiv \int_0^\infty dE \int_R d^3r\phi^+(\underline{r},E) \cdot \phi(\underline{r},E).$$

With respect to this inner product,

$$A^+\phi^+ \equiv \Sigma(\underline{r},E)\phi^+(\underline{r},E)-\int dE'\Sigma_s(\underline{r},E\rightarrow E')\phi^+(\underline{r},E')$$

$$M^+\phi^+ \equiv \nu(E)\Sigma_f(r,E)\int dE'X(E')\phi^+(\underline{r},E'), \text{ and}$$

$$D^{+\,-1}\underline{J}^+ \equiv 3\Sigma(\underline{r},E)\underline{J}^+(\underline{r},E) - 3\int dE'\Sigma_{s1}(\underline{r},E\rightarrow E')\underline{J}^+(\underline{r},E).$$

Now we write a functional (for continuous trial functions which are zero on the outer boundary) whose Euler equations are the Diffusion equation and its adjoint:

$$F_v[v,u] = \int_0^\infty de \int_R d^3r[\nabla v \cdot D\nabla u + v(A-\lambda^{-1}M)u].$$

$$\frac{\delta F_v}{\delta v} = -\nabla \cdot D\nabla u + (A-\lambda^{-1}M)u = 0 \text{ if } u = \phi$$

$$\frac{\delta F_v}{\delta u} = -\nabla \cdot D\nabla v + (A^+-\lambda^{-1}M^+)v = 0 \text{ if } v = \phi^+$$

This functional can be used with any restricted set of trial functions which are continuous and vanish on the boundary; the former conditions are due to the gradient operator; the latter arise from the use of the transformation

$$-\int_R d^3r\nabla v \cdot D\nabla u = +\int_R d^3rv\nabla \cdot D\nabla u = +\int_R d^3ru\nabla \cdot D^+\nabla v.$$

In order to ease the continuity requirement to allow the more general class of trial functions which are continuous within a set of subregions R_K separated by the surface S, we add on Lagrange multiplier terms involving the discontinuities across S:

$$T[v,u,a,b] = \sum_K \int_0^\infty dE \int_{R_k} d^3r[\nabla v \cdot D\nabla u + v(A-\lambda^{-1}M)u]$$
$$+ \int_0^\infty dE \int_S d^2r[a(\underline{r},E)(u_+-u_-)+b(\underline{r},E)(v_+-v_-)]$$

Now the Euler equations require

$$-\nabla \cdot D\nabla u + (A-\lambda^{-1}M)u = 0 \text{ and } -\nabla \cdot D^+\nabla v+(A^+-\lambda^{-1}M^+)v = 0 \text{ in } R_K$$

and also

$$u_+ = u_- \text{ and } v_+ = v_- \text{ across S,}$$

and also

$$a(\underline{r},E) = D_+^+\nabla v_+ \cdot \hat{n} = D_-^+\nabla v_- \cdot \hat{n} \quad \text{on } S,$$

and also

$$b(\underline{r},E) = D_+\nabla u_+ \cdot \hat{n} = D_-\nabla u_- \cdot \hat{n} \quad \text{on } S.$$

This is a very general and useful functional: the functions which make it stationary must satisfy the Diffusion equation within each region R_K and the continuity of flux and (normal) current across the interfaces, but it allows the derivation of approximate solutions which are allowed to be discontinuous. This point is very important, because a requirement of global continuity is bound to conflict with a desire to use special trials tailored to the expected behavior in very different, but adjacent, regions.

Multigroup Expansion

To conclude this chapter we will introduce a trial function space in which the spatial and spectral variables have been separated, and show how this form of restriction can lead to the standard Multigroup Diffusion Equations (among others).

We assume that the expansions

$$u(\underline{r},E) = \sum_{m=1}^{M} \psi_m(\underline{r})X_m(E) \text{ and } v(\underline{r},E) = \sum_{m=1}^{M} \psi_m^+(\underline{r})X_m^+(E)$$

are capable of approximating the solutions of the direct and adjoint diffusion equations well. In addition, we expand

$$a(\underline{r},E) = \sum_{m=1}^{M} \alpha_m^+(\underline{r})\gamma_m^+(E) \text{ and } b(\underline{r},E) = \sum_{m=1}^{M} \alpha_m(\underline{r})\gamma_m(E)$$

(perhaps relating these to the u and v expansions, in view of the known requirements on a and b). To get to the multigroup equations we assume

that the energy dependence of all the trial functions is known, and that they form whole but non-overlapping sets. That is, for every energy E there is one and only one m such that $X_m(E)$ is non-zero (and similarly for X_m^+, γ_m, and γ_m^+).

Inserting these expansions into the Diffusion functional T[v,u,a,b], we see we can integrate out all the energy dependence, and we are left with a reduced functional involving the space-dependent expansion coefficients. If we use subscript notation to indicate the energy-collapse of each operator, i.e.

$$H_{mn} \equiv \int dE X_m^+(E) H X_n(E),$$

the Euler equations of the reduced functional can be written

$$\sum_n -\nabla \cdot D_{mn}\nabla\psi_n + (A_{mn} - \lambda^{-1}M_{mn})\psi_n = 0 \qquad \forall\ m$$

in each R_k, and on the interfaces we have the corresponding conditions

$$\psi_{m+} = \psi_{m-} \text{ on } S \qquad \forall\ m,$$

and

$$\sum_n D_{mn+}\nabla\psi_{n+} = \sum_n D_{mn-}\nabla\psi_{n-} \text{ on } S \qquad \forall\ m.$$

In the case in which the scattering cross-section $\Sigma_{sl}(\underline{r}, E' \rightarrow E)$ in D does not couple energies from the range of one X_m into the range of any other, the collapsed operator D_{mn} will be zero unless m=n, and so the reduced equation will become

$$-\nabla \cdot D_m\nabla\psi_m + \sum_n (A_{mn} - \lambda^{-1}M_{mn})\psi_n = 0 \text{ in } R_K$$

and

$$\psi_{m+} = \psi_{m-} \text{ and } D_{m+}\nabla\psi_{m+} = D_{m-}\nabla\psi_{m-} \text{ on } S.$$

These are the standard Multigroup Diffusion Equations, derived by the Variational Approximation Technique.

The ψ_m are unknown functions (of $\underline{r}$ only) which must satisfy this set of coupled partial differential equations. Unfortunately these equations can be solved exactly only for the most trivial cases, so normally one more level of approximation is applied: the Laplacian term is replaced by a difference operator coupling the values of $\psi_m(\underline{r})$ only at a finite set of points $\underline{r}_n$. The differential equations thus become a set of coupled difference equations which are to be solved for the values $\tilde{\psi}_{mn}$, approximating the ψ_m at each $\underline{r}_n$; we have arrived at the Finite Difference Multigroup Diffusion Equations, which represent the most commonly used approximation method for calculating neutron flux distributions.

CHAPTER VI

FLUX SYNTHESIS

The Finite Difference Multi-Group equations are popular because of their simplicity and their accuracy. The discretized flux representation can be brought as close to the "exact" diffusion theory analytic flux as desired by choosing a sufficiently large number of mesh points and energy groups. Extra points and extra groups are (formally) easy to incorporate because of the simple structure of the equations; this structure also allows a thorough numerical analysis of the resultant matrices [32].

Synthesis Incentives

There are,however, problems in which significant details of the flux distribution can be predicted in advance, and in these situations the full calculation of the discretized flux generates (at great expense) large amounts of redundant information. As an example, consider the analysis of the flux in a fast reactor. An accurate calculation of the flux using the finite difference multi-group equations would require the use of a great many groups (20 to 30) to treat the energy dependence at each point, when in fact it is known that the spectrum shifts fairly smoothly in space from one typical mode to another. The actual "information content" consists of these modal spectra and their relative strengths at each point, so effort is wasted in the finding of the multigroup solution.

Synthesis techniques are designed to treat just such situations, by allowing the construction of approximations which incorporate any known features and require solution for only the remaining unknown parts. (The term "synthesis" is used because this is a process of building up a complicated whole solution out of simpler, but not elementary, component parts).

A synthesis typically involves a series expansion in which each term is the product of a known function of a few of the free variables multiplied by an unknown function not depending on those variables. Both the spherical harmonics expansion and the multigroup expansion fit this description; generally, however, the term "synthesis" is reserved for short series in which the known functions are detailed and tailored to the particular problem, rather than large expansions in relatively simple functions.

The accuracy of a particular synthesis approximation will depend not only on the choice of the expansion functions, but also on the method used to find the expansion coefficients. Occasionally orthogonality properties can be used, but usually some less direct procedure is necessary. Clearly the Variational Approximation Technique, which was used to derive the Diffusion theory, can be applied to fix the unknowns in a synthesis expansion in the same way it was applied to derive the multigroup approximation.

All of the original abstract arguments favoring the variational techniques still apply, but two stand out particularly. The first derives from the fact that the variational approximation principle tries to select a good approximation to the exact stationary function from the reduced space, and that the value of the functional near its stationary point evaluates a goal functional to second order. This is most useful, because the reason for trying to calculate the flux distribution is virtually always to allow the evaluation of some functional of the flux (a reaction rate, etc.) which can be used for the goal functional.

The more important justification of the variational technique is that it keeps all of the approximations visible and consistent. The final

reduced functional obtained by successively restricting the trial space more and more is equivalent to the one which would be obtained by substituting the most restricted trial set into the original functional; thus we can consider all of the approximations to have been made simultaneously at the beginning and all the rest of the method follows by logical derivation. With this viewpoint it is easier to justify the generalization of expansion in elementary functions to expansion in trials chosen because of known closeness to the true solution (rather than ability to form complete sets).

The drawback to using variational synthesis as opposed to weighted residual synthesis (setting integrals of the source residual times some weighting functions equal to zero) is that the variational method seems to require the calculation of adjoint functions, since the multigroup diffusion equations are not self-adjoint. Real consideration of the manner in which the problem is solved, however, shows that only adjoint trial functions are needed. Certainly these may be generated as close approximates to the adjoint, but they can also be chosen to be elementary functions, or equal to the flux trials, or even equal to the weighted residual weights. The fact that this method requires the use of "adjoint" trials does not by itself impose any complications; the difficulties are due to the fact that the accuracy of the solution depends on the choice of the trials. The use of the weighted residual method to avoid having to generate good adjoint approximations implies the acceptance of a correspondingly less accurate approximate solution.

Spatial Synthesis

Flux synthesis has been used for a number of years to generate approx-

imate neutron fluxes, with greatest acceptance of its application to the problem of flux distributions with full three-dimensional dependence. Finite difference diffusion theory calculations require 50 to 500 mesh points along each axis in a typical reactor core, and although a 10,000 point two-dimensional calculation at some cross section through the reactor is feasible, a 1,000,000 point three-dimensional finite difference scheme is really rather impractical.

Because of the manner in which reactors are constructed, however, we know that there will be strong flux variations across any two-dimensional section taken through the fuel assemblies, whereas the variation along flow of coolant will tend to be a smooth shift from one 2-D mode to another, each characterized by the 2-D coolant and control distributions, etc. Synthesis can be applied by expanding the spatial distribution functions $\psi_m(\underline{r})$ into series with known cross-sectional modes multiplied by unknown axial expansion coefficients: $\psi_m(\underline{r}) = \sum_l C_l(z)\ \Theta_l(x,y)$.

This is the basis for the Multichannel flux synthesis scheme [40]. Here the 2-D functions $\Theta_l(x,y)$ are not required to exist over the whole core cross section - instead the section is divided into a set of non-overlapping "channels" and each $\Theta_l(x,y)$ is non-zero in only one. By this method (an analogue of the multigroup expansion) the 2-D detail characteristic of typical regions is incorporated in advance into the solution, and the final numerical problem involves the calculation only of the gross 2-D coupling of channel to channel and the finite-difference expansion of the coupling coefficients in only one dimension.

The actual implementations of spatial synthesis vary somewhat because of differences in the choice of the functional and the nature of the trial functions. If the flux expansion is not required to be contin-

uous then special means must be found to treat the gradient terms in the diffusion equation. One such method [39] was reviewed earlier: the incorporation into the functional of extra surface integrals which add the proper continuity conditions to the Euler equations. An alternate scheme has been to revert to the P-1 functional and use it directly, specifying trial expansions for both the scalar flux and the current [40,41].

A recent development in flux synthesis goes a step further: the axial combining coefficient functions are themselves expanded into series with unknown scalar coefficients multiplying known functions each of which is non-zero only in a particular segment of the axis. The overall result is to partition the reactor into 10-20 intervals in each dimension (producing a stack of blocks or "nodes") and then couple the proposed 3-D trial flux in each one to its neighbors. With this model [42], gross 3-D effects are calculated from a reasonably small number of scalar coupling coefficients, while the detailed flux behavior inside each node is accounted for by the 3-D cell calculations used to generate the node trial fluxes.

Spectral Synthesis

In the sense that the multigroup approximation is a synthesis method, synthesis of the energy dependence of the flux has been used for a long time. In the derivation of the multigroup equations it is assumed that the fine structure of the energy dependence is fairly constant over large energy ranges, and that it is only necessary to compute scale factors to be applied to the pre-calculated spectra within each group. This proved very successful in applications to thermal spectrum reactors, where indeed the dominant energy effect is the coupling of the neutron "birth"

region through the resonance region to the thermal region, but it is not so useful in fast reactor analysis.

The weakness of the standard multigroup theory is not an inability to produce sufficiently accurate results; it is the inability to do this cheaply in certain situations, namely when the details of the energy dependence cannot be assumed to be reasonably constant over large regions of space or when the details cannot be predicted over broad groups in energy. In the former situation, calculations must be performed carrying data for many regions (each with its own characteristic spectrum), an expensive process. In the latter situation, calculations must be performed with a large number of groups (in order to resolve the unknown detail), an even more expensive process.

Both of these problems arise during attempts to calculate the spatial and energy dependence of thermal-energy fluxes. There are only a few principal parameters in this energy range (temperature, leakage, etc) but the resultant flux behavior is complicated, requiring fine energy and spatial calculations. Calame [42,43] attacked this problem with a synthesis expansion in known spectra representing the potential extremes, all covering the full energy range. These overlapping spectra are combined with space-dependent weight coefficients that provide the shift from one region's typical spectrum to another's. The detail (which in view of the small number of physical parameters is clearly mostly redundant) is eliminated and only the few combining coefficients need be calculated.

Fast Spectrum Synthesis

Another application in which the use of spectral synthesis with overlapping modes should be very useful is in the analysis of fast flux distri-

butions. (It is interesting to note that this is at the other end of the spectrum. In typical water moderated reactors the gross coupling of the fast, resonance, and thermal regions is very important, and the details of the spectrum need not be very exact; but when only one energy range is being studied the fine structure is required information).

To do a proper multigroup analysis of a fast reactor requires calculations in 20 to 30 groups because of the effects of the resonances in the fast region. The spatial variations of the flux, however, are fairly smooth because the long mean free paths of fast neutrons make fine structural detail "invisible". This effect compensates somewhat, but not entirely, for the greater number of energy variables (because the number of spatial variables can be reduced) but does not affect the fact that the spectrum is everywhere in transition (so that many material regions should be used). Using 20 groups makes a two-dimensional diffusion code expensive to run and makes three-dimensional analyses almost impossible, but spectral synthesis seems to hold out the promise of solving the space and energy fast flux problem without doing all the (diffusion theory) work [22].

The usual Spectral Synthesis equations are derived from the Diffusion theory functional in the same way that the Multigroup equations are: we expand $u(\underline{r}, E) = \sum_m \psi_m(\underline{r}) \chi_m(E)$, but with the difference that the $\chi_m(E)$ are known functions of energy which span the whole energy range. Now the unknown space-dependent functions represent combining coefficients-relative proportions of the trial modes - rather than scale factors to be applied to each energy group.

Of course, nothing good is free, and spectral synthesis has some definite drawbacks. Potential discontinuities in the flux represent one

such problem, but this is shared with the multigroup method and is treated the same way: either extra interface continuity terms are added to the Diffusion functional or else the P-1 functional is used with separate flux and current trials. Unique to the synthesis technique, however, are the problems of full mode coupling (due to the use of overlapping modes). This problem manifests itself in the structure of the matrices D_{mn} and A_{mn} in the reduced equation: while D was diagonal and A lower triangular (downscattering only) in multigroup theory, D and A are both full matrices coupling every mode to every other in the synthesis method.

The most important question about synthesis, whether it really is an accurate alternative to multigroup theory, seems to have been answered in the affirmative. Several different studies [45, 46, 47] have shown that Spectral Synthesis approximations using three or four trials can achieve (except in occasional cases of anomalous failure) accuracies of a few tenths of a percent in criticality and reaction rate, provided that good flux trials and adjoint trials (weight functions) are used. To ensure that the latter qualification can be met, Cockayne [48] has shown how "Successive Space-Energy Synthesis" can be used to generate trial functions of the required quality. (He proposes an iterative scheme: use energy synthesis to get spatial flux distributions; then use this flux in a spatial synthesis to generate good spectrum trials).

Deterrents

Despite the formal incentives for using Spectral Synthesis and the demonstrations of its reasonable accuracy, this method has received only very limited application to practical problems of reactor analysis. The primary deterrent has been the occasional occurence of "anomalous failures"

of synthesis methods - cases in which the approximate solutions turn out to be extraordinarily poor.

Examples of this have been reported for various types of spatial synthesis [49], for spatial synthesis with group collapsing [50], and for spectral synthesis [47]. (A case of anomalous failure was also discovered during the course of the current investigations.) The effect of these reports has been to scare off potential users of synthesis since it is hard to justify an element of risk when trying to solve genuine, practical problems.

This lack of confidence has been aggravated by the lack of any formal method of analysis which could identify the causes of the anomalies or predict their occurance. When using the finite difference multigroup approximation one can rely on a large body of mathematical analyses giving assurance that iterations will converge, fluxes will be positive, eigenvalues will be real, etc. Unfortunately the synthesis approximation is not susceptible to mathematical analysis. The goal, after all, is to develop a method whereby detailed guesses at the solution of a particular problem can be incorporated into the approximation; but the formal equations for the remaining unknown functions thereby are written in terms of a large amount of unknown information (the detailed guesses) about which any analysis must make the most conservative (worst) assumptions.

All this uncertainty would quickly be put out of mind if the Spectral Synthesis method proved to be significantly cheaper than the competing multigroup methods. A cheap calculation can be repeated or replaced if it turns out badly; a cheap calculational method allows the rapid accumulation of experience in how to avoid dangerous situations. Thus it has been disappointing to find that the synthesis methods do not seem to

achieve the savings implied by the fact that they require the calculation of only about one-tenth as many unknowns as are required by the multigroup methods. The reason for this is that when the number of energy variables is reduced, the nature of the equations coupling them is made more complicated, and thus more effort is required to solve them. Comparisons of alternate solutions of a given problem have shown that the synthesis methods used required from one-sixth [45] to one-half [51] of the calculational time required by multigroup programs. This is not a sufficient savings to justify the abandonment of the experience with and confidence in the finite difference multigroup diffusion methods.

The net evaluation of Spectral Synthesis is that although it is useful, it will not suddenly supplant the more traditional methods. The reason for this is not really due to the anomalies, as might be expected: synthesis results are generally accurate, and as Stacey [42] points out, the anomalies occur very infrequently and are easy to recognize. The real resistance to synthesis is due to the lack of strong economic incentives (the investment in multigroup codes balances the savings of synthesis methods) and also a degree of "mental inertia" — unwillingness to abandon all the accumulated experience in preparing and analyzing multigroup diffusion models.

CHAPTER VII

SPECTRAL SYNTHESIS OF DISCRETIZED FLUXES

In this chapter we shall begin describing the development and computer implementation of a variant of the spectral synthesis developed in the previous chapters. Up to this point, we have treated synthesis methods as alternates to the discretization methods. That is, we used the analytic diffusion theory functional to derive competing approximations based on different types of trial function expansions.

Now we take the point of view that we have reduced the functional whose Euler equations determine the analytic flux to the functional whose Euler equations determine the finite-difference multigroup (F-D M-G) fluxes, and we want to reduce it still further via spectral synthesis. Instead of comparing this new approximation to the "exact" diffusion theory flux $\phi(\underline{r},E)$ at every point $\underline{r}$ and every energy E, we accept as "truth" the F-D M-G flux f_{ng} available only at the points $\underline{r}_n$ and in the groups g.

The set of discretized fluxes f_{ng} (there are NxG of them) satisfy the set of linear equations whose terms are the discretized versions of the operators in the analytic diffusion equation. Writing f_{ng} as a vector, we see that the system of equations (NxG of them) can be written as one matrix equation $\underline{\underline{D}}\underline{f} + \underline{\underline{A}}\,\underline{f} = \underline{\underline{T}}\,\underline{f} + \lambda\,\underline{\underline{F}}\,\underline{f}$ (where λ is the eigenvalue required in a criticality problem) with the following definitions of the component matrices. The matrix derived from the Laplacian term $\nabla\cdot D\nabla\phi$ is $\underline{\underline{D}}$, typically representing 3, 5, or 7 point difference expressions written for every n and every g. The total removal term $\Sigma(\underline{r},E)\,\phi(\underline{r},E)$ becomes the matrix $\underline{\underline{A}}$, which is diagonal (no coupling effects). Scattering is included in the matrix $\underline{\underline{T}}$, which couples only from group to group; and the fission source (effectively an upscattering) is included through the matrix $\underline{\underline{F}}$.

The functional for this approximation (the reduced functional corresponding to the discrete expansion of the diffusion theory scalar flux) can now be written as $E[v,u,\lambda\ ] = v^T(D+A-T)\ u - \lambda\, v^T F u$. We have dropped the underscore notation for vectors and matrices, and replaced f and its adjoint with u and v (since f is the stationary point $\bar{u}$). Since the trial "functions" are now vectors, we see that $E[v,u,\lambda\ ]$ is in reality an ordinary function of the 2 x NxG+1 scalar variables v_{ng}, u_{ng} and λ. The Euler equations are found by setting equal to zero the ordinary derivatives of E.

$$0 = \frac{\partial E}{\partial v_n^g} \Rightarrow (D+A)\bar{u} = (T+\lambda F)\bar{u}$$

and

$$0 = \frac{\partial E}{\partial u_n^g} \Rightarrow \bar{v}^T(D+A) = \bar{v}^T(T+\lambda F)$$

$$\text{or } (D^T+A^T)\bar{v} = (T^T+\lambda^* F^T)$$

(Here we see that the matrices in the equation for the adjoint are the transposes of their counterparts in the direct flux equation.)

Synthesis

The synthesis approximation is obtained by restricting the freedom of the trials v and u and by finding the stationary point in the new reduced space. We expand

$$\underline{u} = \underline{\underline{X}} \cdot \underline{c} \quad \text{or} \quad u_n^g = \sum_{m=1}^{M} x_{nm}^g\, c_{nm}$$

and

$$\underline{v} = \underline{\underline{Y}} \cdot \underline{e} \quad \text{or} \quad v_n^g = \sum_{m=1}^{M} y_{nm}^g\, e_{nm},$$

where the c_{nm} and e_{nm} are the new unknown expansion coefficients and the X and Y matrices provide a specified energy (group)dependence. We consider c and e to be vectors (even though they are doubly subscripted) because, as before, we shall be able to combine all the equations for all the c_{nm} and e_{nm} into two matrix equations.

The matrices X and Y obviously must be dimensioned NxG by NxM, but most of their elements will be zero (so that $\underline{\underline{X}}\cdot\underline{c}$ yields the summation formula given). The non-zero elements X^n_{gm} represent the M different trial spectra (the g dependence) which will be linearly combined $\sum_m X^g_{nm}\, c_{nm}$ to form the trial flux at the point n. The sets of spectra used at different points may, but need not, depend on n. Since the object is to treat the transition in space (with the c_{nm}) of the flux, clearly the strategy to be used in choosing the X^g_{nm} is to try to pick spectra which might be typical or dominant in the regions near n. Sometimes one set of a few such spectra can be used for every point in the model, but more generally different sets will be used in different regions, with less important spectra ignored so as to keep the number of unknowns (NxM) as low as possible.

Note that by starting off with the finite difference multigroup equations we have sidestepped the formal problem of ensuring continuity of the flux trials used for v and u. It is most important to consider this, because in a practical application it will undoubtedly be necessary to use different trial spectra in different regions to achieve reasonable accuracy, and thus discontinuities in the expansion coefficients should be expected. The synthesis solution, however, will be an approximation to the discretized flux, so that "continuity" really is not a meaningful term. We presume that the finite difference equations governing the

discretized flux are derived in an appropriate manner so as to incorporate the special interface terms needed to approximate continuity, so that no further special treatment is needed in subsequent approximations.

When the spectral synthesis expansions are substituted for the F-D M-G fluxes, we find the reduced functional

$$E[e,c,\lambda] = e^T Y^T (D+A-T-\lambda F) X c.$$

Again, we find the reduced Euler equations by setting equal to zero the ordinary derivatives of E as a function of e_{nm} and c_{nm}. In detail:

$$E = \sum_{k,j} \sum_{\ell} y^j_{k\ell} e_{k\ell} \sum_{n,g} (D+A-T-F)^{n,g}_{k,j} \sum_{m} X^g_{nm} C_{nm}$$

$$\frac{\partial E}{\partial e_{k\ell}} = \sum_{j} y^j_{k\ell} \sum_{n,g} (D+A-T-\lambda F)^{n,g}_{k,j} \sum_{m} X^g_{nm} C_{nm} = 0 \qquad \forall\ k,\ell$$

$$\frac{\partial E}{\partial c_{nm}} = \sum_{k,j} \sum_{\ell} y^j_{k\ell} e_{k\ell} \sum_{g} (D+A-T-\lambda F)^{n,g}_{k,j} X^g_{nm} = 0 \qquad \forall\ n,m$$

But these can be written very simply in matrix notation - the stationary points satisfy the equations

$$Y^T (D+A-T-\lambda F) X c = 0$$

$$e Y^T (D+A-T-\lambda F) X = 0$$

or, separating out the eigenvalue term

$$Y^T (D+A-T) X c = \lambda Y^T F X c$$

$$X^T (D^T+A^T-T^T) Y e = \lambda^* X^T F^T Y e$$

(Here we have a good reason for using the term "reduced": the matrix system for the vectors v and u was of order NxG by NxG, whereas this system is of order NxM by NxM, with the number of modes presumably much smaller than the number of groups).

Defining $\tilde{D} = Y^T D X$, $\tilde{A} = Y^T A X$, etc. we can write $(\tilde{D} + \tilde{A} - \tilde{T})c = \lambda \tilde{F} c$ and $(\tilde{D} + \tilde{A} - \tilde{T})^T e = \lambda^* \tilde{F}^T e$ as the equations that will actually be solved (by computer) to approximate the flux distribution needed to analyze a nuclear reactor.

How do we justify this level of approximation in comparison to the real analytic space, energy and angle dependent flux (i.e. the solution of the neutron transport equation). Two steps are required: first we must be sure that the analysis that will be performed does not require knowledge of the angular dependence (other than the net current) or of the fine details of the energy dependence or of the transients in the scalar flux near interfaces. In other words, we must be satisfied with the amount of detail that can be extracted from a multi-modal finite difference diffusion approximation.

Second, assuming that the <u>form</u> of the approximation is valid, we must justify the method for determining the numerical value of the free parameters. For the synthesis method presented here we do this by claiming that the final approximation is the stationary point, in the class of restricted functions accepted as valid, of the original variational functional for the complete neutron transport equation. That this is true can be seen by following the steps of successive approximations for the flux - the P-N method, Diffusion theory, finite difference multigroup diffusion theory, and finally spectral synthesis of the discretized flux. Each approximation is the solution of the Euler equations of the same further and further restricted functional. We do not claim that this procedure is guaranteed to generate the best approximation (since the true stationary point is not an extremum). Rather, the formal strength

is that it is systematic and self-consistent so that variations can be explored and sensitivities tested without having to generate and justify completely new approximations after every change. (It is comforting to know, however, that the method does seem to work fairly well in practical applications.)

Generalized Eigenvalue Problem

Returning now to the practical problem at hand - finding an approximate solution for $(D + A - T)\, f \equiv Hf = \lambda Ff$ - we perhaps should comment on the use of the functional form

$$E[v,u,\lambda] = \langle v, Hu \rangle - \lambda \langle v, Fu \rangle$$

which clearly has been appearing ever since the neutron transport problem was introduced in Chapter V.

In the chapters on variational methods we developed and analyzed a related functional

$$E'[v,u,\lambda] = \langle v, Hu \rangle - \lambda \langle v, u \rangle$$

which had Euler equations

$$H\bar{u} = \lambda\bar{u} \quad \text{and} \quad H^{+}\bar{v} = \lambda^{*}\bar{v}.$$

It should be obvious that the new functional E is an extension to provide Euler equations of the form

$$H\bar{u} = \lambda F\bar{u} \quad \text{and} \quad H^{+}\bar{v} = \lambda^{*}F\bar{v},$$

which are known as generalized eigenvalue equations.

Referring to the analysis of the Rayleigh principle we see that it can be generalized also, and that

$$Y[v,u] = \langle v, Hu \rangle / \langle v, Fu \rangle$$

provides a second order estimate of the eigenvalue λ of the generalized problem. (The proof of this is made immediately obvious upon redefining $G[v,u] = \langle v, Fu \rangle$ instead of the $\langle v, u \rangle$ used before).

Because of the occasional anomalous failures of synthesis we also want to re-examine the analysis of the Euler equations of the generalized E functional, for whatever warnings it might give. As before, we assume that v and u are restricted to the subspaces generated by the linear transformations U(x) and V(y), and that these transformations can be defined by the expansions

$$U(x) = \sum_i U_i x_i = \sum_i \sum_k u_k a_{ki} x_i = \underline{u} \cdot \underline{\underline{a}} \cdot \underline{x}$$

and similarly

$$V(y) = \sum_j V_j\, j = \sum_j \sum_\ell v_\ell \tilde{b}_{\ell j} y_j = \underline{v} \cdot \underline{\underline{b}} \cdot \underline{y}$$

where the u_k are now eigenvectors of the generalized problem $Hu_k = \lambda_k F_k u_k$ and the v_l are the eigenvectors of its adjoint problem $H^T v_l = \lambda_l^* F^T v_l$ (forming complete bi-orthonormal sets).

Since the eigenvalue is not applied directly to the argument function, we no longer get the "$\tilde{I}$" matrix in the reduced Euler equations:

$$\underline{\underline{b}}^T \cdot \langle \underline{v}, H\underline{u} \rangle \cdot \underline{\underline{a}} \cdot \underline{x} = \lambda \underline{\underline{b}}^T \cdot \langle \underline{v}, F\underline{u} \rangle \cdot \underline{\underline{a}} \cdot \underline{x}$$

$$\underline{y}^T \cdot \underline{\underline{b}}^T \cdot \langle \underline{v}, H\underline{u} \rangle \cdot \underline{\underline{a}} = \lambda \underline{y}^T \cdot \underline{\underline{b}}^T \cdot \langle \underline{v}, F\underline{u} \rangle \cdot \underline{\underline{a}}$$

but we recall the I matrix disappeared immediately, so that where before we had $\underline{\underline{b}}^T \cdot \langle \underline{v}, \underline{u} \rangle \cdot \underline{\underline{a}}$, now we have $\underline{\underline{b}}^T \cdot \langle \underline{v}, F\underline{u} \rangle \cdot \underline{\underline{a}}$. Thus the reduced equations developed by this eigenvector-synthesis of the generalized eigenvalue problem are formally no more complicated than those for the regular eigenvalue problem; the added difficulty that we would expect to find associated with such a generalization has not appeared in the structure of the problem.

Assuming that the eigenvector - synthesis is possible, we see that a sufficient condition for failure (by allowing arbitrary λ) is that the null spaces of

$$\underline{b}^T \cdot \langle \underline{v}, H\underline{u} \rangle \cdot \underline{a} \quad \text{and} \quad \underline{b}^T \cdot \langle \underline{v}, F\underline{u} \rangle \cdot \underline{a}$$

overlap. Unfortunately, this condition involves so many unknown quantities in such complicated relations that it is probably useless as a predictive tool - it just provides a warning that sometimes things can go wrong.

"Exact" F-D M-G Problem

The method of spectral synthesis of the discretized flux cannot be implemented and tested without explicitly specifying the nature of the "exact" finite-difference multigroup problem. Here we will finally start defining the actual equations which will be incorporated into a synthesis computer code, MACH/360. This code consists of a revised version of the MACH-1 one-dimensional finite-difference multigroup diffusion theory and perturbation analysis code [52], combined with a spectral synthesis module which can be used in place of the multigroup module to calculate the flux.

MACH/360 finds an approximate solution to the multigroup equations

$$-\nabla \cdot D^g(r)\nabla f^g(r) + \Sigma_t^g(r) f^g(r) = d^g(r)$$

$$= \sum_{j=1}^{g} \Sigma_s^{j\to g}(r) f^j(r) + \lambda z^g(r) \sum_{j=1}^{G} \nu\Sigma_f^j(r) f^j(r)$$

with the following important assumptions: there is only one space dimension (in slab, cylinder, or sphere geometry); the material properties (cross sections)are constant within each of a number of regions K; the flux f_g and the current $D \nabla f_g$ are to be continuous across the interface

between any two regions. We have defined $\chi^g(r)$ as the fission spectrum and $\nu\Sigma_f^j$ as the fission production term; $\Sigma_s^{j\to g}$ is the scattering term and is explicitly written for downscattering only; and $\nabla \cdot D(r) \nabla f(r)$ becomes

$$D_K^g \nabla^2 f^g = D_K^g \left(\frac{\partial^2 f^g}{\partial r^2} + \frac{p}{r} \frac{\partial f^g}{\partial r} \right)$$

within each region.

The difference equations solved by the code are related to these equations in the following manner. First, the spatial variable is discretized by choosing N mesh points r_n with constant spacing h_K in each region. Within these regions of constant material parameters the fluxes $f_n^g = f^g(r_n)$ are coupled by the following equations

$$-D_K^g \left\{ \left[\frac{f_{n-1}^g - 2f_n^g + f_{n+1}^g}{h_K^2} \right] + \frac{p}{r_n} \left[\frac{f_{n+1}^g - f_{n-1}^g}{2h_k} \right] \right\} + \Sigma_{tk}^g f_n^g = d_n^g$$

and

$$d_n^g = \sum_{j=1}^{g-1} \Sigma_{SK}^{j\to g} f_n^j + \chi_K^g \sum_{j=1}^{G} \nu \Sigma_{fK}^j f_n^j.$$

(where p = 0, 1, or 2 for slab, cylinder, or sphere geometry). (These central difference approximations to the analytic equations were, to be truthful, formed by substituting the difference approximations for the analytic operators, but in principle they can be derived as Euler equations.) Defining $P_n = p\, h_K/2r_n$ and $a_K^g = h_K^2/D_K^g$, we see that these equations be written more compactly using the form

$$-(1-P_n) f_{n-1}^g + (2 + a_K^g \Sigma_{tK}^g) f_n^g - (1+P_n) f_{n+1}^g = a_K^g d_n^2.$$

These equations are not suitable for those points r_n which lie on interfaces: for these points there are requirements of continuity of flux

(easily satisfied because $f_n^g = f_n^g$) and continuity of current (not possible to satisfy exactly) which can be used to relate the fluxes for $r > r_n$ (the + side) to the fluxes for $r < r_n$ (the - side). The approach used in the FAIM code [53] is used to derive these equations. We know that if r_ℓ and r_m are the boundary points for region K, then

$$I_K \equiv \int_{r_\ell}^{r_m} -D_K \nabla^2 f(r) r^p dr = -r_m^p D_K \nabla f(r_m) + r_\ell^p D_K \nabla f(r_\ell).$$

This is an exact relation, which we can approximate by using trapezoidal integration to show that

$$\begin{aligned} I_K &= \int_{r_\ell}^{r_m} (d(r) - \Sigma_t(r) f(r)) r^p dr \\ &\approx r_m^p \left[- \frac{h_K}{a_m} (1-P_m)(f_m - f_{m-1}) + \frac{h_K}{2} (d_m - \Sigma_{tm} f_m) \right] \\ &\quad - r_\ell^p \left[+ \frac{h_K}{a_\ell} [1+P_\ell)(f_\ell - f_{\ell+1}) - \frac{h_K}{2} (d_\ell - \Sigma_{t\ell} f_\ell) \right] . \end{aligned}$$

Now we equate the coefficients of r_ℓ and r_m to find approximate relations (at an interface) for $J_{\ell+} = -D_K \nabla f_{\ell+}$ and $J_{m-} = -D_K \nabla f_{m-}$. Defining $\gamma_n \equiv (D_{K+} h_{K-})/(D_{K-} h_{K+})$ we equate $J_{n+} = J_{n-}$ to derive

$$\begin{aligned} & - \frac{[1-P_{K-}]}{\gamma_n} f_{n-1} - (1+P_{K+}) f_{n+1} \\ & + f_n \frac{[1+P_{K-}]}{\gamma_n} + (1-P_{K+}) + \frac{1}{2} \left[\frac{a_{K-} \Sigma_{tK-}}{\gamma_n} + a_{K+} \Sigma_{tK+} \right] \\ & = \frac{1}{2} \left[\frac{a_{K-} d_{K-}}{\gamma_n} + a_{K+} d_{K+} \right] \end{aligned}$$

The equations for each interface, together with the equations for each internal point, form a system relating the flux at each point (except

the boundaries) to the fluxes at its two nearest neighbors. At the inner boundary (n = 1, k=1) and the outer boundary (n = N , k = K) we need to find some way to terminate this coupling. This can be done with finite difference approximations for the standard boundary conditions $\beta f + \eta \nabla f = \xi$ (where β , η , and ξ are arbitrary parameters). The appropriate difference equations are

$$[\eta_1(1+\tilde{P} + \frac{a_1\Sigma t1}{2}) - \beta_1 h_1]f_1 - \eta_1(1+\tilde{P})f_2 = \eta_1 \frac{a_1 d_1}{2} - \xi_1 h_1$$

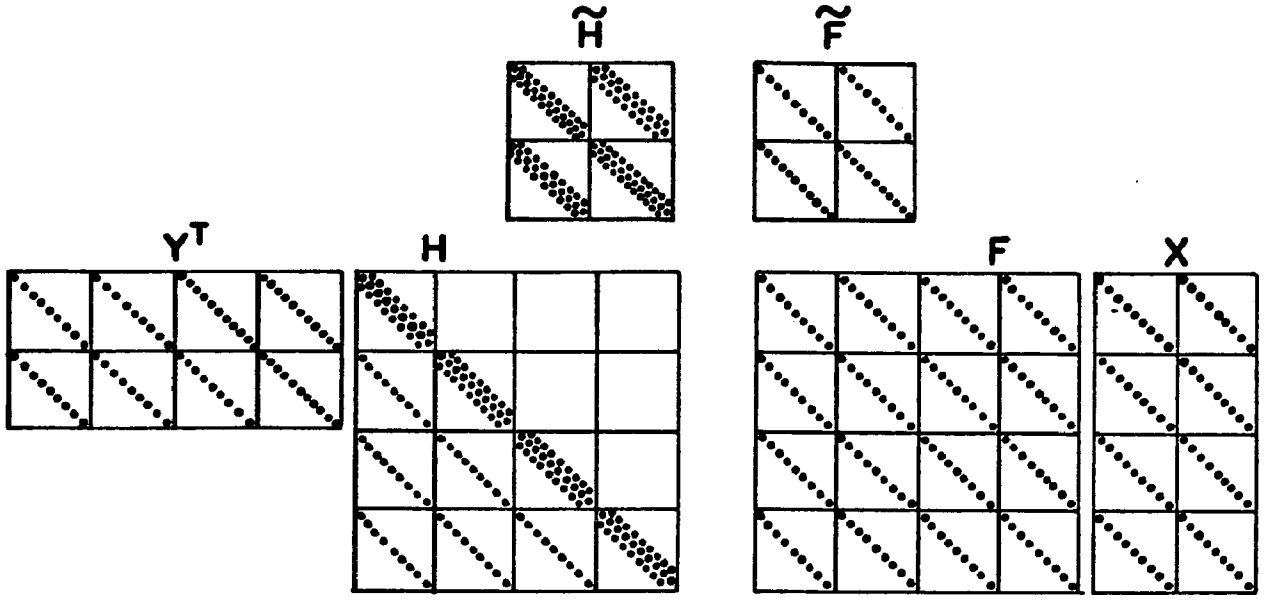
(where $\tilde{P} = P$, if $r_1 \neq 0$; $\tilde{P} = p$ if $r_1 = 0$)

and

$$\eta_N(1-P_N)f_{N-1} + [\eta_N(1-P_N + \frac{a_K\Sigma tK}{2}) + \beta_N h_K] = \eta_N \frac{a_K d_N}{2} + \xi_N h_K$$

This completes the derivation of the F-D M-G equation set. We have NxG unknown fluxes f_n^g, and clearly there are NxG equations to be solved to determine them. The matrix representation $Hf = \lambda Ff$ of these equations uses sparse matrices with a very simple structure (see Figure 2). These matrices can be thought of as being partitioned into a G by G array of blocks, each of which is an N by N array of coefficients coupling the f_n^g for all n but for only one given g. With this arrangement the F matrix, for example, has non-zero entries only on the diagonal of each N by N block. This is due to the fact that the fission process couples each flux f_n^g to the fluxes in every other group f_n^j at the same point.

Similarly, the scattering matrix T has non-zero entries only on the diagonals of the blocks in the lower triangular portion of T: the scattering process only couples fluxes of high energy (in the upper groups)



THE "OBVIOUS" ORDERING (ABOVE)

$$\tilde{H} = Y^T H X$$

$$\tilde{F} = Y^T F X$$

THE "PREFERRED" ORDERING (BELOW)

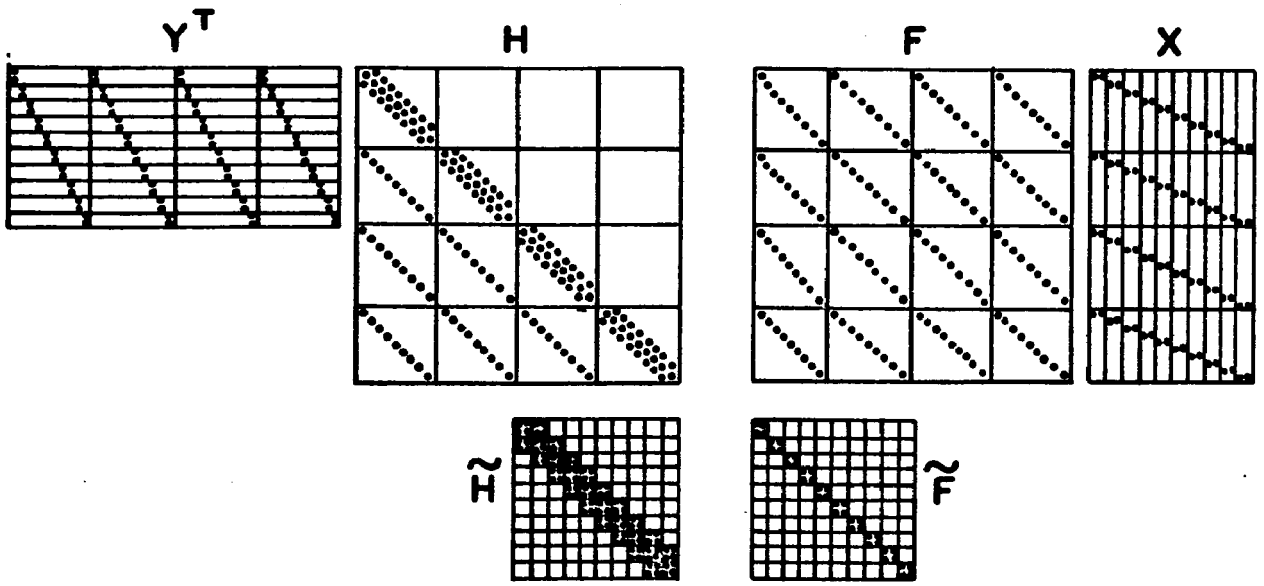


Figure 2. Structure of Multigroup Diffusion Theory Matrices and Spectral Synthesis Matrices (when N=10 points, G=4 groups, M=2 modes)

to fluxes of lower energy. (The triangular nature of T will be seen to have great importance). The removal matrix A has been lumped in with the diffusion matrix D, since both are non-zero only in the block diagonal: in fact, the A matrix is purely diagonal, there is no coupling at all.

Each N by N block on the diagonal of the D matrix in the tridiagonal matrix derived from the difference equations written above to couple the fluxes at adjacent points. If the spatial dependence allowed two or three dimensions these N by N matrices would have five or seven non-zero diagonals (describing the greater degree of spatial coupling) and be considerably harder to solve.

These matrices are used in an iterative solution for the smallest eigenvalue λ and its associated eigenvector [32] (there exists a proof that for the F-D M-G eigenvalue there is an everywhere positive eigenvector associated with the smallest λ). The procedure is the familiar "source" or inverse power iteration: to solve $Hf = \lambda Ff$ we assume some initial $f^{(o)}$ and form $s^{(o)} = Ff^{(o)} / \| Ff^{(o)} \|$; then the iterates are defined by

$$Hf^{(n+1)} = s^{(n)} = \frac{Ff^{(n)}}{\|Ff^{(n)}\|} = \frac{FH^{-1}s^{(n-1)}}{K^{(n)}} = \frac{(FH^{-1})^n s^{(0)}}{\prod_{m=1}^{n} K^{(m)}}$$

as the number of iterations increases, $K^{(n)} \equiv \| Ff^{(n)} \|$ will approach the inverse of the smallest eigenvalue λ ($K^{(n)} \rightarrow K_{eff}$, the multiplication) and $f^{(n)}$ will approach the associated eigenvector - the fundamental mode of the flux.

Note that it is not necessary to form the inverse matrix H^{-1}: it is only necessary to solve equations of the form $Hf = s$. This is very important in view of the potentially very high order of the matrix H. The equation $Hf = s$ is solved easily by working on only one group at a time

(a procedure which can be considered to be either due to or requiring the separation of the fission matrix from the others). Since there is only downscatter coupling in the matrix H (which is block lower triangular) we can solve the N by N system for the fluxes in the first group, then compute the downscatter as pseudo-sources into the lower groups and solve the N by N system for the second group, etc. After the flux has been calculated for every group, a new source vector is computed by multiplying the fission matrix times this flux and normalizing the result. The iteration continues until the differences between successive sources become very small.

(The solution of the N by N system representing the diffusion coupling may be very easy, as in the current case when it is tridiagonal, or very difficult, as in the three dimensional case when there are typically seven non-zero diagonals. For two and three dimensional problems each of the G spatial coupling systems is usually solved approximately with its own iteration scheme, the most common being the successive over-relaxation method discussed earlier as an example of perturbation expansion methods.)

Formation of Trial Functions

To approximate the solution of these equations, we will restrict the form of the flux vectors to conform to the spectral synthesis:

$$u_n^g = \sum_m x_{mn}^g \, c_{mn} \text{ and } v_n^g = \sum_m y_{mn}^g \, e_{mn}$$

The c_{mn} and e_{mn} are now the only free variables, while the x_{nm}^g and y_{nm}^g are the trial spectra to be used to reduce the matrix equations.

We will assume that these trials form an orthonormal set in order to avoid some potential numerical problems: if any direct flux trials were

nearly linearly dependent then the f_n^g would be expanded with large cancelling terms; if any adjoint trials were nearly linearly dependent then the elements of the reduced equations weighted with them would not be well differentiated. Since the synthesis expansion is linear, the requirement of orthonormality presents no real restriction, since any set of (physically motivated) good flux trials can be transformed by the Gram-Schmidt procedure into an equivalent orthonormal set. After the synthesis equations corresponding to this set are found, the flux approximation is recovered by performing the inverse transformation.

Returning now to the problem of what to do with the trial spectra (assuming they have been made orthonormal), we want to develop the matrix forms X and Y in terms of x_{mn}^g and y_{mn}^g so that

$$u = u_n^g = \sum_m x_{mn}^g \, c_{mn} = X_c$$

and

$$v = v_n^g = \sum_m y_{mn}^g \, e_{mn} = Y_e$$

Since we are writing the fluxes u_n^g as elements of the vector u and the expansion mode coefficients c_{mn} as elements of the vector c, the matrix X must be dimensioned to have N x G rows and N x M columns. Most of the NxGxNxM elements are zero, however, since only the NxGxM values of the X_{mn}^g need be incorporated, and so we see that X_{mn}^{kg} equals $x_{mn}^g \, \delta_{kn}$, where δ_{kn} is the Kronecker delta, thus yielding the proper value u = Xc. Similarly we see that Y_{mn}^{kg} must equal $y_{mn}^g \, \delta_{kn}$. (We note here that the matrices X and Y are never really formed during the numerical solution of a synthesis problem - they are too big to store and contain no more information than the trial spectra. They are introduced formally to aid in the derivation of the formulas for the reduced matrices, which will be formed directly.)

Reduced Functional

The synthesis approximation involves solution of the reduced Euler equations of the functional $E[v,u,\lambda] = v^T(D + A - T - \lambda F)\, u = v^T(H - \lambda F)u$. Writing this in component notation,

$$E = \sum_{t,j}^{N,G} v_{tj} \sum_{s,g}^{N,G} (H-\lambda F)_{sg}^{tj}\, u_{sg},$$

and after substituting the synthesis expansion

$$E[e,c,\lambda] = e^T\, Y^T[H-\lambda F]X_c$$

or

$$E = \sum_{t,j}^{N,G} \sum_{\ell,p}^{M,N} Y_{\ell p}^{tj}\, e_{\ell p} \sum_{s,g}^{N,G} (H-\lambda F)_{sg}^{tj} \sum_{m,n}^{M,N} X_{mn}^{sg}\, C_{mn}.$$

The Euler equations are obtained by setting each of the derivatives of E equal to zero:

$$\frac{\partial E}{\partial e_{\ell p}} = \sum_{t,j}^{N,G} Y_{\ell p}^{tj} \sum_{s,g}^{N,G} (H-\lambda F)_{sg}^{tj} \sum_{m,n}^{M,N} X_{mn}^{sg}\, C_{mn};$$

thus

$$0 = \sum_{m,n}^{M,N} \sum_{t,j}^{N,G} Y_{\ell p}^{tj} \sum_{s,g}^{N,G} (H-\lambda F)_{sg}^{tj}\, X_{mn}^{sg}\, C_{mn}$$

for every combination of l and p. These NxM equations can be written in the matrix notation as

$$0 = Y^T(H-\lambda F)X_c \quad \text{or} \quad Y^T H X_c = \lambda Y^T F X_c$$

and similarly the adjoint mode coefficients must satisfy the equation

$$X^T H^T Y e = \lambda^* X^T F Y e.$$

The synthesis program will have to calculate the elements of H and F, obtain the orthonormal trial spectra x_{mn}^g and y_{mn}^g, and then form the re-

duced matrices

$$\tilde{H} = Y^T H X \text{ and } \tilde{F} = Y^T F X.$$

Since all of the matrices involved are very sparse, we may anticipate that the reduced matrices will also have many zero elements, the locations of which are of interest for two different reasons: first, if we can predict where the non-zero elements will lie we need not ever form the full matrices $\tilde{H}$ and $\tilde{F}$ - we only have to reserve storage space for the non-zero values. Second, the fullness and structure of $\tilde{H}$ will determine the nature of the numerical methods which can be used to solve this system. We will return to the question of matrix structure later, after we have seen how $\tilde{H}$ and $\tilde{F}$ are used (formally) to find an approximate eigenvalue and eigenvector.

Inverse Power Iteration

The goal of all this work is to approximate the fundamental mode eigenvector $\bar{f}_n^g$ of the finite difference multigroup diffusion equation and also to approximate its corresponding eigenvalue λ_D, which is the smallest eigenvalue of that equation (and is the inverse of k_D, the multiplication factor). We shall choose as these approximations the smallest eigenvalue λ_o of the reduced equation $\tilde{H}c = \lambda \tilde{F} c$, and the synthesis flux $X\bar{c}$ formed from its associated eigenvector $\bar{c}$. There is no guarantee that these will be good approximations, only that the error $|\lambda_D - \lambda_o|$ will be of second order with respect to the error in the flux. The justification for this choice is by "consistency" - it is derived by the Variational Approximation principle.

We shall extract the desired eigenvector from $\tilde{H}c = \lambda \tilde{F}c$ by using the inverse power iteration method ("inverse" since we want the largest

$1/\lambda$). Rearranging the equation, we have $1/\lambda\ \tilde{H}c = \tilde{F}c$. Defining $k = 1/\lambda$ and $1/\lambda_o = k_o$, the approximation to the multiplication factor, we further derive $kc = \tilde{H}^{-1}\tilde{F}\ c$. (The use of the inverse of $\tilde{H}$ is formal only, and we shall see that we need only to be able to solve equations in $\tilde{H}$.)

To conform with the usual multigroup iteration scheme we shall also introduce the "source" vector $z = \tilde{F}c$, although this no longer can be interpreted so simply as the source of fission neutrons. Introducing this, we finally have the equation upon which the iteration will be based: $k\ z = \tilde{F}\ \tilde{H}^{-1}\ z$.

Choosing an initial arbitrary source $z^{(o)}$, we define the iteration procedure

$$K^{(n)} = ||z^{(n)}||;$$

$$\tilde{H}c^{(n+1)} = z^{(n)}/k^{(n)};$$

and

$$z^{(n+1)} = \tilde{F}\ c^{(n+1)}$$

Clearly

$$c^{(n)} = \left[\prod_{m=0}^{n-1} K^{(m)}\right]^{-1} (\tilde{H}^{-1}\tilde{F})^{n-1}\ \tilde{H}^{-1}\ z^{(0)}$$

so that if there is a smallest λ_o (a largest k_o) then $c^{(n)}$ will approach $\bar{c}$ (normalized so that $||\tilde{H}\bar{c}|| = 1$) while $k^{(n)}$ approaches k_o (regardless of the norm used). The iteration proceeds until the fractional change in $k^{(n)}$ (and optionally the fractional change of every c_{mn}) is less than some small convergence parameter.

Matrix Structure

Recall that we have not yet specified the structure of the trial mode matrices X and Y; this is because the arrangement of the x^g_{mn} and y^g_{mn} can

be chosen so as to induce a useful structure for $\tilde{H}$. Two alternate possibilities for X and Y are shown in Figure 2. The rows must be ordered to correspond to the arrangement of H - with G sets of N points - but the columns can be ordered as M sets of N points or as N sets of M modes. In the former case (the "obvious" ordering) the x^g_{mn} (and y^g_{mn}) appear on diagonals of the NxN blocks, and the reduced matrices have a structure of MxM blocks each with NxN points. As is shown, $\tilde{F}$ is very much like F, but $\tilde{H}$ has been "filled out" with three non-zero diagonals in every block. This is a manifestation of the prime problem of spectral synthesis: the coupling of every mode to every other mode in not only the scattering terms but in the diffusion terms as well. These coupling terms, which resemble a sort of upscattering, prevent the solution of these equations mode-by-mode (the way the multigroup equations were partitioned and solved group-by-group). The equations must be solved simultaneously, either by a direct process (which has to treat the whole NxM by NxM matrix $\tilde{H}$) or else by some iterative technique (e.g. treating the "up-coupling" as a perturbing matrix).

The problem with performing a direct solution of $\tilde{H}c = (\lambda \tilde{F}c)$ when $\tilde{H}$ has this MxM block structure is that the elements of $\tilde{H}$ are so "spread out". The fact that this structure is so regular and so sparse suggests that the rows (equations) and columns (unknown) could be permutted so as to make $\tilde{H}$ easier to work with. This is true, but rather than try to work out such a permutation we shall produce the new $\tilde{H}$ directly from the alternate ordering of X and Y.

With the columns of X and Y arranged as N sets of M columns each (the "preferred" ordering), the matrices $\tilde{H}$ and $\tilde{F}$ consists of an NxN system of blocks, each with MxM elements (see Figure 2, again). Now, the mode

coupling at a given point is expressed by these small MxM blocks, while the spatial coupling manifests itself through the block structure. Since the fission process involves no spatial coupling the $\tilde{F}$ matrix is block diagonal; the $\tilde{H}$ matrix has a block tridiagonal structure of the diffusion terms. (For two or three dimensional problems, $\tilde{H}$ would have five or seven block diagonals.) These matrices are equivalent to the previous versions, but are more amenable to solution.

Block Decomposition

The iterative procedure for finding the eigenvector of $\tilde{H}c = \lambda \tilde{F}c$ will require many successive solutions of equations of the form $\tilde{H}c = z$, where $\tilde{H}$ has non-zero elements only on three block diagonals. Because the size of the blocks is small (MxM) and because tridiagonal matrix equations can be solved directly very rapidly, we shall treat the individual blocks of $\tilde{H}$ as "elements" and apply tridiagonal matrix methods to $\tilde{H}c = z$. In particular, we shall form the block LU decomposition of $\tilde{H}$ before entering the inverse power iteration cycle and save the components L and U for use in each iteration.

The regular LU decomposition [54] consists of the calculation (essentially by the process of Gaussian elimination) of the elements of a lower triangular matrix L and an upper triangular matrix U whose product is equal to $\tilde{H}$. Once this is done (and it can be done faster than $\tilde{H}$ can be inverted) the solution of $\tilde{H}c = z$ can be found by a two step process, $Ld = z$ and then $Uc = d$, as quickly as c could be found by multiplying $\tilde{H}^{-1}z$. Since $\tilde{H}$ is block tridiagonal, however, we prefer the "block" LU decomposition, where the MxM blocks of $\tilde{H}$, L, and U are manipulated as though they were individual elements. With this procedure, L and U are both block bidiagonal, so that the calculation of each c is extremely rapid.

The details of the LU decomposition and the solutions of Ld = z and Uc = d are derived by writing the procedures formally as though the elements were scalars, and then replacing the ordinary arithmatic with equivalent matrix operations. Thus c = a + b is replaced by $\underline{\underline{c}} = \underline{\underline{a}} + \underline{\underline{b}}$ (each being an MxM matrix), and c = a/b is replaced by the procedure of solving $\underline{\underline{c}} \cdot \underline{\underline{b}} = \underline{\underline{a}}$ (formally $\underline{\underline{c}} = \underline{\underline{a}} \cdot \underline{\underline{b}}^{-1}$, but we avoid computing the matrix inverse). This method for obtaining the exact solution of block tridiagonal systems has been successfully incorporated as the heart of the inverse power iteration for the synthesis eigenvector.

CHAPTER VIII

APPLICATION OF WIELANDT'S METHOD

The failure to achieve substantial computational savings through the use of spectral synthesis is due to the increased complexity of the equations which must be solved. Although the expansion of the flux with overlapping trial spectra may allow a significant reduction in the number of trials (and therefore unknown coefficients), the use of overlapping, rather than disjoint, energy functions in the Diffusion theory functional produces Euler equations in which every mode coefficient is coupled to every other.

Cost of Synthesis

The ordinary multigroup equations, derived with the disjoint trial spectra, are never as complicated as the synthesis equations because the diffusion terms, $\nabla \cdot D_g \nabla \phi_g$, are assumed to affect only one group coefficient. In the synthesis equations, however, the energy dependence of the diffusion coefficient forces coupling of each mode to each other mode by diffusion like terms, $\nabla \cdot D_{lm} \nabla \phi_m$, as well as by the expected scattering terms. Thus for a physical problem featuring only downscattering the finite-difference multigroup equations will have terms relating each flux element only to its spatial neighbors in the same group and to its energy successors at lower energies at the same point; on the other hand, the synthesis equations will have terms relating each mode coefficient to the coefficients at the same and adjacent points in every other mode.

Although we have shown how the spectral synthesis matrix can be rearranged into a block-banded form, which can be formally solved by methods

used for one-group diffusion problems, this rearrangement does not eliminate the basic complexity - it just makes it bearable. The greater difficulty of solving the synthesis equations is the price that is paid for the reduction in the number of variables.

The applicability of spectral synthesis to a given problem must be evaluated in terms of: the number of modes which can potentially provide the accuracy of a reference multigroup calculation; the increased effort required to solve the modal equations rather than the multigroup equations; and the probability that inexperience in modal modeling (or anomalous behavior of the synthesis) will cause the solution to be invalid. The question of calculational effort has been considered by Cockayne [55] for one-dimensional problems. He deduces that the ratio of the number of multiplicative operations required to solve an M-mode synthesis compared to the number required to solve a G-group ordinary system is about $(\frac{4}{5})M^3/G^{1.5}$, for modest M and G. This formula indicates that the calculational effort can be cut to about one-fifth by substituting a three mode synthesis for a twenty-two group standard calculation. While this seems attractive enough, it apparently is not sufficient to outweigh the uncertainties associated with the use of synthesis.

Synergism

Since the acceptability of an approximation method is related directly to its cost reduction capability but inversely to the confidence in its results, we are motivated to look for special situations in which the savings resulting from the use of Spectral Synthesis are larger than the factor of two to five achieved in competition with plain multigroup methods. The goal is to find some synergistic combination of methods whose strengths

reinforce each other and whose weaknesses do not, so that the net effect is an improvement in effectivensss which exceeds that expected from the independent application of either.

It is obvious that such a synergistic effect will occur whenever Spectral Synthesis is used in conjunction with another technique which also forces "up-coupling" terms onto the (block) lower triangular multi-group equations. The penalty of directly or indirectly treating the terms above the diagonal only is applied once, but the new technique should combine the advantages of both its components.

For example, consider the reactor time-eigenvalue problem. Here the flux is assumed to have a time dependence $\phi(\underline{r},E)e^{\alpha t}$, and it is desired to find the possible values of α. Since the time-dependent diffusion equation differs from the time-independent equation only by the addition of a term $\frac{1}{v}\frac{\partial \phi}{\partial t}$, we see that ϕ must satisfy a regular diffusion equation incorporating a pseudo-absorption cross section $\frac{\alpha}{v}$. This equation must be satisfied exactly (no eigenvalue can be applied to the fission source), so we see that the only allowable values of are the eigenvalues of the modified equation $(D + A - T - F)\phi = \frac{-\alpha}{v}\phi$. These eigenvalues are usually found by trial and error: test values are used to actually incorporate the $\frac{\alpha}{v}$ term into the absorption matrix, and then $(D + (A+\frac{\alpha}{v}) - T)\phi = \lambda F \phi$ is solved to evaluate the λ required by the guessed α; the guesses are varied until $\lambda = 1$.

The eigenvalues are not extracted directly because the fission matrix F would incorporate upcoupling terms into the (D + A - T - F) matrix, thus greatly increasing the cost of solution: the successive solution of a whole series of regular multiplication factor problems is cheaper. This iterative process can be completely eliminated, however, if the

problem is solved with Spectral Synthesis, because in this case the fission matrix is actually simpler than the others. The use of Spectral Synthesis in the time eigenvalue problem offers potential savings equal to the single synthesis advantage multiplied by the number of eliminated iterations.

Wielandt's Method

A hybrid scheme with far greater potential than the simple direct-α method involves combining Spectral Synthesis with Wielandt's Method [56]. The latter is a technique combining inverse power iteration with a shift of eigenvalue that is viewed most often as a method of extracting eigenvectors but can also be applied to accelerate the extraction of eigenvalues. We will develop a version of Wielandt's Method that is suitable for the generalized eigenvalue problem; it will be clear why the combination with Spectral Synthesis is so useful.

Given a generalized eigenvalue problem $Hu = \lambda Fu$, we will subtract an arbitrary multiple of Fu from both sides, to obtain $(H - \lambda_o F)u = (\lambda - \lambda_o)Fu$. Clearly both equations have the same set of eigenvectors, and the eigenvalues of the second equation are those of the first, shifted by the arbitrary value λ_o.

Now we apply the inverse power iteration procedure to the shifted eigenvalue problem: given an initial guess $u^{(o)}$,

we form $w^{(n)} = F\, u^{(n)}$,

$k^{(n)} = \| w^{(n)} \|$,

and $(H - \lambda_o F)\, u^{(n+1)} = w^{(n)}/k^{(n)}$.

If this iteration converges to a particular $\bar{k}$ and $\bar{u}$, we see that

$(H-\lambda_o F)\bar{u} = (1/k)F\bar{u}$ so that $\bar{u}$ is an eigenvector of $Hu = \lambda Fu$ associated with the eigenvalue $\lambda = \lambda_o + 1/\bar{k}$.

Furthermore, analysis shows that

$$u^{(n)} = \left[\prod_{m=0}^{n-1} k^{(m)} \right]^{-1} \left[(H-\lambda_o F)^{-1} F \right]^n u^{(0)},$$

so that if $u^{(o)}$ can be expanded in a set of eigenvectors $u^{(o)} = \sum_\ell a_\ell u_\ell$, then

$$u^{(n)} = \left[\prod_{m=0}^{n-1} k^{(m)} \right]^{-1} \sum_\ell a_\ell \left[(H-\lambda_o F)^{-1} F \right]^n u_\ell,$$

$$= \left[\prod_{m=0}^{n-1} k^{(m)} \right]^{-1} \sum_\ell a_\ell \left(\frac{1}{\lambda_\ell - \lambda_o} \right)^n u_\ell,$$

or

$$u^{(n)} = \sum_\ell a_\ell u_\ell \prod_{m=0}^{n-1} \left(\frac{1/k^{(m)}}{\lambda_\ell - \lambda_o} \right).$$

If there is a λ_k such that $|\lambda_k - \lambda_o| \leq |\lambda_1 - \lambda_o|$, and if $k^{(n)}$ and $u^{(n)}$ converge, we see from this $\bar{k}$ must equal $(\lambda_k - \lambda_o)^{-1}$ and that $\bar{u}$ is the eigenvector associated with λ_k.

The fact that this method can be used to extract the eigenvector corresponding to any known eigenvalue is interesting, but not particularly useful in reactor analysis. Of greater value is the observation that the rate of convergence of $u^{(n)}$ to $\bar{u}$ can be made very large.

Each undesirable component of $u^{(n)}$ will be reduced by the factor

$$\frac{1/k^{(m)}}{\lambda_\ell - \lambda_o} \sim \frac{\lambda_k - \lambda_o}{\lambda_\ell - \lambda_o}$$

after each cycle of the iteration: The closer λ_o is to λ_k, the faster the convergence. In particular, in the reactor problem we are seeking the

smallest eigenvalue λ (which is presumably unity). The ordinary inverse power iteration corresponds to Wielandt's method with a zero shift, and so the rate of convergence is controlled by λ/λ_1, where λ_1 is the next larger eigenvalue. Choosing some $\lambda_0 < \lambda$ and performing Wielandt's shift we create an iteration scheme with convergence controlled by $\left(\frac{\lambda - \lambda_0}{\lambda_1 - \lambda_0}\right)$ which can be made much less than λ/λ_1.

Wielandt's Method would appear to be of great value in reactor analysis: so many problems require the extraction of the multiplication factor. Why is it not used? Because the shifted multigroup matrix $(H - \lambda_0 F)$ no longer is block-lower-triangular. The acceleration of the convergence would be cancelled out by the greater effort required to solve the shifted equations. However, this argument does not apply to the Spectral Synthesis equation! Shifting over part of the fission matrix causes no complications at all, and so Wielandt's Method for accelerating the eigenvector convergence should be applicable with no additional penalty whatsoever. Only a trivial modification (a matrix subtraction) is required to apply Wielandt's Method to a Spectral Synthesis problem; this modification has been incorporated as an option in the MACH/360 program.

The potential savings available through the use of this combination of (individually unattractive) methods should make this hybrid scheme very attractive for practical reactor analysis problems.

CHAPTER IX

AN EXAMPLE

The usefulness of Spectral Synthesis, especially when used in conjunction with Wielandt's method, has been demonstrated with examples calculated by the MACH/360 core analysis code package. This set of programs provides (among other things) facilities for the solution of the one-dimensional multigroup diffusion theory criticality problem and also for the solution of the Spectral Synthesis approximation to this multigroup flux. (The "MACH" computer codes -- MACH-1, MACH/360, and MACHLIB -- are described more fully in the Appendices.)

Model Problem

The example which we will describe here is based upon the critical experiment ZPR-III Assembly 48. This assembly [57] was one of a series of experiments conducted by the Argonne National Laboratory in support of the liquid metal fast breeder program. The composition of Assembly 48 was chosen to provide a central flux spectrum similar to that expected in a large power reactor, while the geometry was chosen to be as simple as possible so as to make the task of experimental evaluation easier. Assembly 48 has served as a test case for the evaluation of numerical methods since 1966, when it served this function for international intercomparison of fast reactor analysis codes [58]. We have used Assembly 48 as a test case because of the thorough experimental data available with which to check the MACH results.

ZPR-III Assembly 48 was "cylindrical" and "homogeneous" to within the limitation of its 2-inch wide drawers. The equivalent critical cylinder was calculated to be 76.35 cm in height and 41.71 cm in radius,

with radial and axial blankets each 30 cm thick. Since the MACH code allows treatment of only one spatial dimension the following two region finite cylinder was used as an appropriate model: a 42 cm radius core (of Assembly 48 core composition); surrounded by a 30 cm thick blanket (of Assembly 48 radial blanket composition); with extrapolated height of 118.95 cm. The radial axis was divided into 33 intervals (starting at the center) in the following manner: 5 of 2 cm, 6 of 4 cm, 8 of 1 cm (reaching the core-blanket interface at mesh point #20); 6 of 1 cm, 4 of 4 cm, and 4 of 2 cm.

Test Calculations

The multigroup criticality calculation was performed using the 22 group cross section library known as "ANL Set 224". This is a pre-ENDF-B library which was developed for the analysis of fast critical experiments similar to Assembly 48. The cross sections contained in it are out of date, but sufficiently accurate (or typical) to be used for an intercomparison of numerical methods.

The multiplication factor for this model problem was calculated to be 0.999988 by the multigroup diffusion equations (due to a careful choice of the extrapolated height). The flux (normalized to a maximum of 2.0) as a function of lethargy at points 1, 20, and 28 (the center of the core, the core-blanket interface, and a central point in the blanket) is plotted in Figure 3; in Figure 4 we show the radial distribution of the normalized flux in groups 1, 6, and 11.

The calculation of this flux by the inverse power iteration method used in MACH required 10 iterations to meet the criterion that the fractional change in the fission source be less than 10^{-5} at every point.

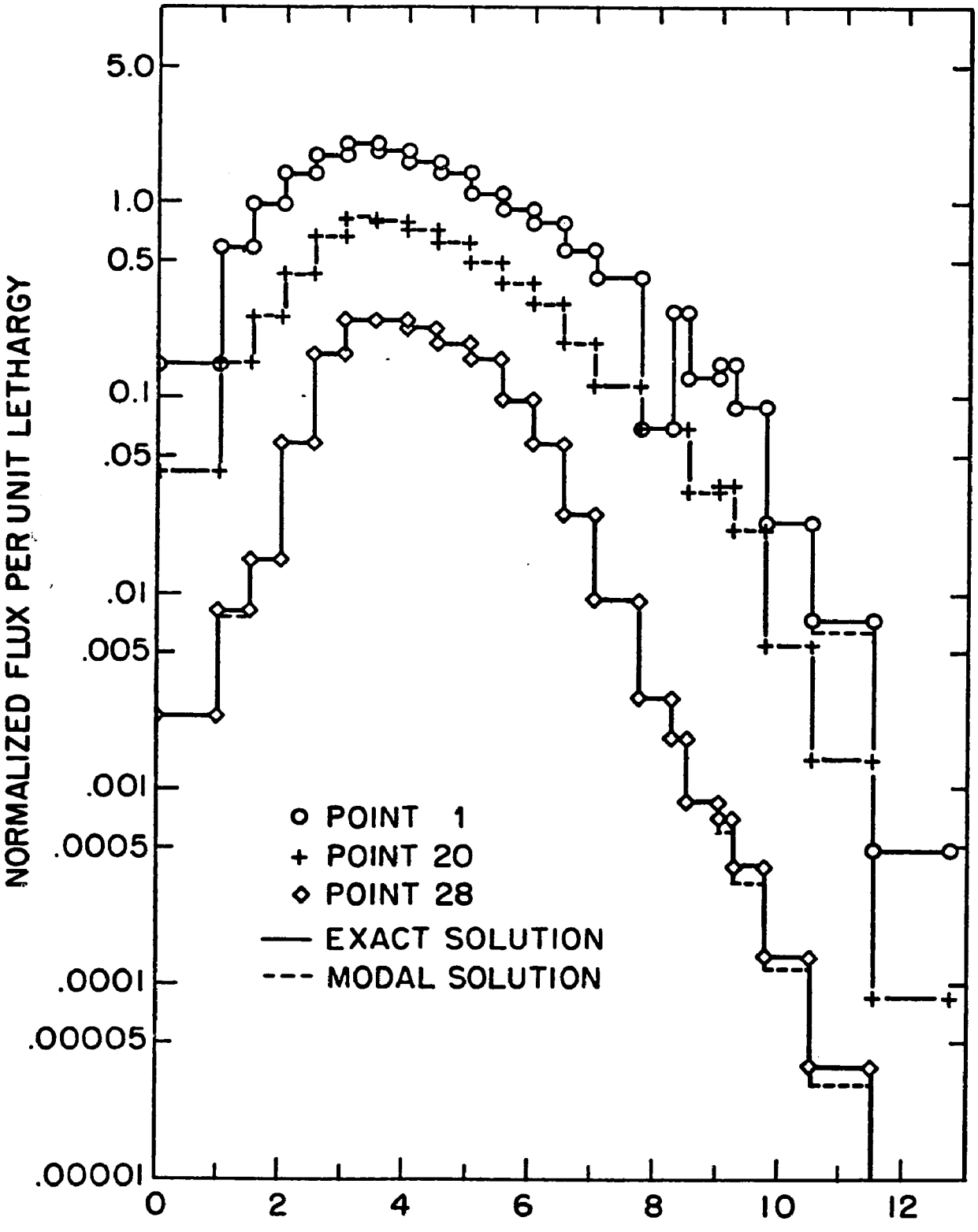


Figure 3. Calculated Spectral Flux Distribution in ZPR III Assembly 48 Test Case

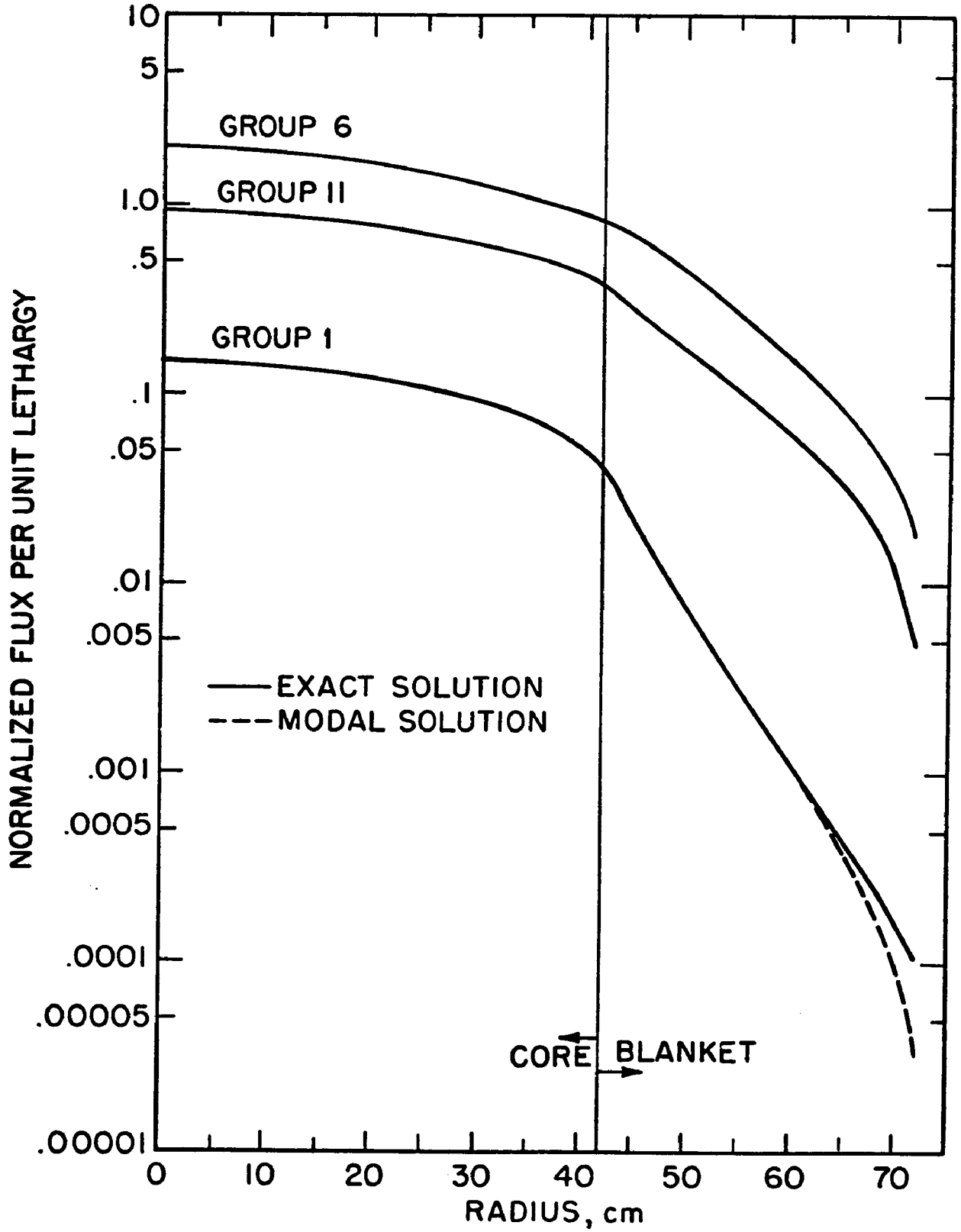


Figure 4. Calculated Spatial Flux Distribution in ZPR III Assembly 48 Test Case

Each iteration required approximately 0.65 seconds of CPU (central processing unit) time on the IBM System 360/Model 67 computer operated by the University of Michigan Computing Center.

The same model problem was solved using the Spectral Synthesis option of MACH/360. Three trial flux spectra and three adjoint trials were provided as input data. The flux trials used were the exact spectra from points 1, 20, and 28 as calculated by the multigroup option; the justification for this is that the goal here is to compare feasibility and computational effort, and that accuracy per se has been demonstrated elsewhere [45, 48]. The adjoint trials were constructed according to the "reaction rate weight" scheme of Neuhold [46]: each adjoint trial is the product of the corresponding flux trial and the (group-dependent) core absorption cross section $\Sigma_{capture} + \Sigma_{fission}$.

With these trials, the Spectral Synthesis eigenvalue equations were solved by ordinary inverse power iteration, yielding a multiplication factor of 1.00814 and the fluxes plotted in Figures 3 and 4. The error of 0.8% in k_{eff} is not out of line with the errors reported from other synthesis tests, and the fluxes appear to approximate the multigroup fluxes very well (except in a few regions of where the flux is very small compared to the peak value). The expansion coefficients of the three modes are plotted in Figure 5: the shifting of the dominant contribution from one mode to another is easily seen.

This solution required 9 iterations to meet the criterion that the fractional change in the modal source vector $(\underline{\underline{\tilde{F}}} \cdot \underline{c})$ be less than 10^{-5} at every point in every mode. Each iteration required approximately 0.17 seconds of CPU time, about 1/4 of the time required for a multigroup iteration.

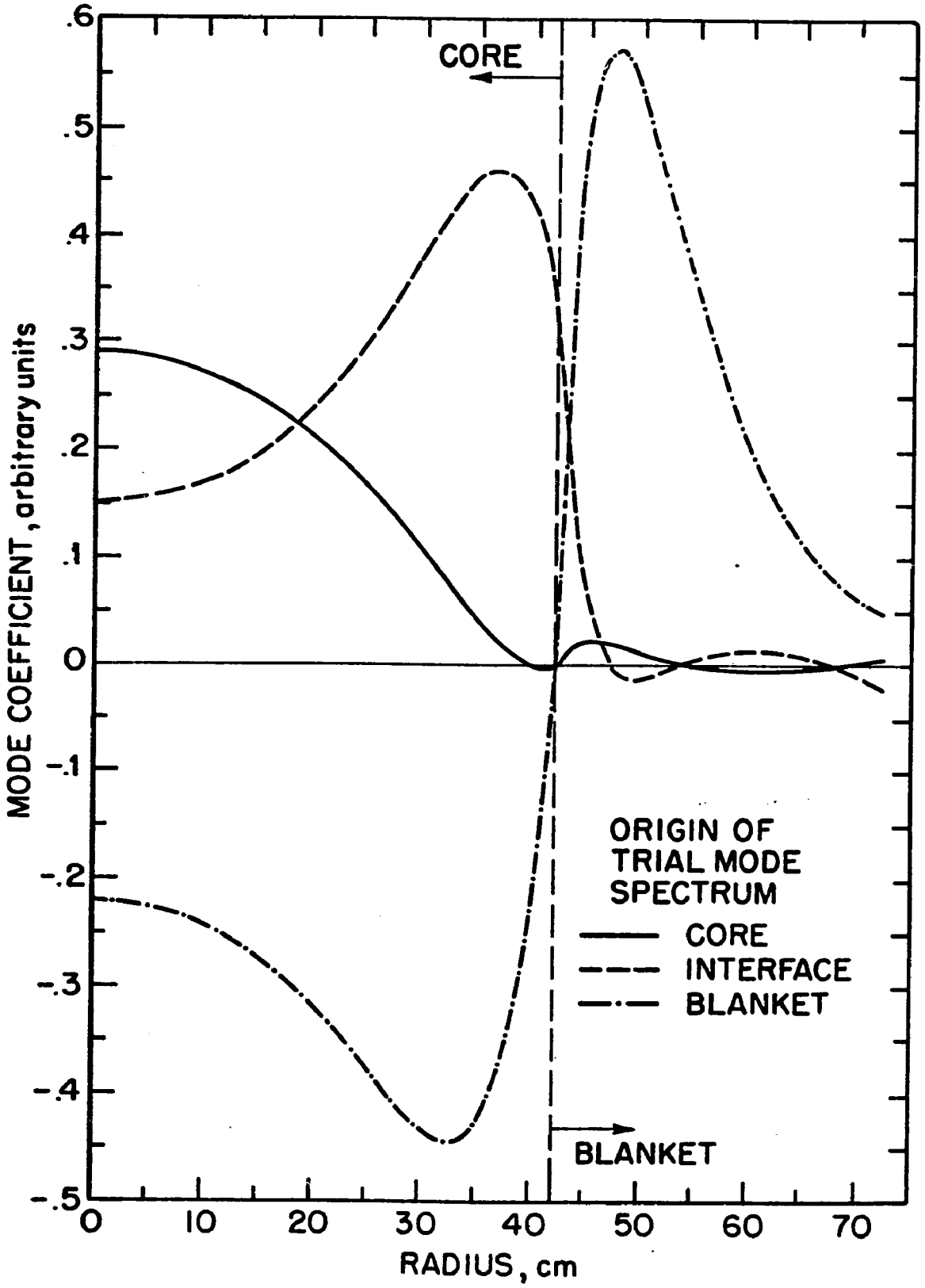


Figure 5. Calculated Modal Expansion Coefficients in ZPR III Assembly 48 Test Case

As a test of Wielandt's Method, this modal problem was executed again with only one change: just before the LU decomposition of H was performed (in preparation for the inverse power iteration) an eigenvalue shift of $\lambda_o = 1/1.009$ was performed. The result was convergence to the same criterion after only 3 iterations, with virtually identical fluxes and multiplication.

The particular application of Wielandt's method reduced the iteration cycle computational effort by a factor of 3, but the synthesis iterations were already a factor of 4 cheaper than the multigroup iterations. Thus we have an example of a savings of 11/12 of the time required to generate the flux eigenvector. If we assume that the Spectral Synthesis equations offer speed improvement factors of 3 to 8, and that Wielandt's method can decrease the number of iterations by a factor from 2 to 10, we see that the potential savings resulting from the use of this combination can be very large.

In closing, we must report one disconcerting fact: the Assembly 48 model case also gave rise to an anomalous failure. One synthesis was attempted with the same flux trials taken from the exact solution, but with adjoint trials also taken from the exact adjoint solution at points 1, 20, and 28; the inverse power iteration of the reduced eigenvalue equations did not converge.

In an attempt to determine the cause of this behavior we arranged to extract all the eigenvalues of the $\tilde{H}c = \lambda \tilde{F} c$ system (to see what might be interfering with convergence to a smallest eigenvalue which should have been about 1/1.003). This extraction was made possible by the use of the newly-developed "QZ" algorithm [59], which solves the generalized matrix eigenvalue problem $Ax = \lambda Bx$ even when B is singular: the components of

the $\tilde{H}$ and $\tilde{F}$ matrices were dumped from core and then fed back as input to QZ.

The unexpected result of this procedure was the discovery that, although the smallest <u>real</u> eigenvalue of this $\tilde{H}c = \lambda \tilde{F} c$ is 0.991 = 1/1.00879 $\approx$ 1/1.003, there is one pair of complex conjugate eigenvalues (.733 $\pm$.571 i) with smaller norm:.929 $<$.991. Since the inverse power iteration eliminates all eigenvectors except those whose eigenvalues have smallest <u>norm</u>, the cause of the non-convergence is obvious: the successive iterates are arbitrary <u>real</u> vectors which exist in the space spanned by the <u>complex</u> eigenvectors whose eigenvalues are the pair having the smallest norm. The mystery of this anomalous case is the mechanism which introduced those complex eigenvalues.

CHAPTER X

CONCLUSIONS

This dissertation is a report of a series of studies performed to assess the potential usefulness for fast reactor analysis of the Spectral Synthesis approximation method. The initial chapters provide a thorough review of the Variational Theory which is used to justify the synthesis methods; the final chapters analyze Spectral Synthesis as it would really be applied -- as an approximation to the finite-difference multigroup diffusion equations.

The investigations of the various types of variational perturbative approximation methods leads to the conclusion that most practical approximation techniques will take one of two forms: either a simple variational synthesis which achieves accuracy by introducing many free parameters, or a perturbation expansion with very few parameters (if any) carried to high order. The intermediate schemes seem to be too complex to ever see much use.

A further conclusion from the theoretical analysis is that virtually all flux synthesis approximations are inherently unreliable, since they are derived from variational principles for non-extremal functionals. Reliance on them must be justified by successful experience.

The analysis of the computer implementation of Spectral Synthesis has shown some potentially important applications, even though it appears that the synthesis equations are sufficiently more complex than the multigroup to prevent their general adoption. There clearly are certain hybrid methods combining Spectral Synthesis with some other marginally useful technique in which a synergistic effect provides great potential for savings.

Wielandt's Method can be made useful in reactor analysis this way; certainly there are other possibilities. The search for and investigation of this type of synergistic hybrid should provide an interesting topic for much future research.

LIST OF REFERENCES

1. Gelfand, I. M., and Fomin, S. V., Calculus of Variation, Prentice-Hall, Englewood Cliffs (1963).

2. Rektorys, K., ed., Survey of Applicable Mathematics, MIT Press, Cambridge (1969).

3. Lighthill, M.J., Introduction to Fourier Analysis and Generalized Functions, Cambridge University Press, Cambridge (1964).

4. Kaplan, S., "Variational Methods in Nuclear Engineering," Advances in Nuclear Science and Technology, Vol. 5, ed. P. Greebler, Academic Press, New York (1969).

5. Becker, M., The Principles and Applications of Variational Methods, MIT Press, Cambridge (1964).

6. Pomraning, G. C., "Complementary Variational Principles and Their Application to Neutron Transport Problems", J. Math. Phys., 8, 2096 (1967).

7. Selengut, D. S., "Variational Analysis of Multi-Dimensional Systems," USAEC Report HW-59126, p. 89 (1959).

8. Selengut, D.S., "The Construction of Approximate Theories by Variational Methods," Trans. ANS, 5, 413 (1962).

9. Roussopoulos, P., "Methodes Variationneles Theorie des Collisions," C. R. Acad. Sci. Paris, 236, 1858 (1953).

10. Pomraning, G. C., "A Derivation of Variational Principles for Inhomogeneous Equations," Nucl. Sci. Eng., 29, 220 (1967).

11. Selengut, D. S., "On the Derivation of a Variational Principle for Linear Systems," Nucl. Sci. Eng., 17, 310 (1963).

12. Friedman, B., Principles and Techniques of Applied Mathematics, John Wiley and Sons, New York (1956).

13. Pomraning, G. C., "The Calculation of Ratios in Critical Systems," J. Nuclear Energy, 21, 285 (1967).

14. Pomraning, G. C., "The Calculation of Neutron Flux Ratios in Critical Systems by the Indirect Variational Method," Nucl. Sci. Eng., 34, 308 (1968).

15. Pomraning, G. C., "Reciprocal and Canonical Forms of Variational Problems Involving Linear Operators," J. of Math. and Physics, 47, 155 (1968).

16. Kaplan, S., and Davis, J. A., "Canonical and Involutory Transformations of the Variational Problems of Transport Theory," Nucl. Sci. Eng., 28, 166 (1967).

17. Yasinsky, J. B., and Kaplan, S., "On the Use of Dual Variational Principles for the Estimation of Error in Approximate Solutions of Diffusion Problems," Nucl. Sci. Eng., 31, 80 (1968).

18. Buslik, A. J., "Extremum Variational Principles for the Monoenergetic Neutron Transport Equation with Arbitrary Adjoint Source," Nucl. Sci. Eng., 35, 303 (1969).

19. Pomraning, G. C., "A Limitation of the Roussopoulos Variational Principle?," Nucl. Sci. Eng., 28, 150 (1967).

20. Dwivedi, S. R., "Virtual Limitation of Variational Principle," Nucl. Sci. Eng., 31, 174 (1968).

21. Pomraning, G. C., "A Numerical Study of the Method of Weighted Residuals," Nucl. Sci. Eng., 24, 291 (1966).

22. Stacey, W. M., Jr., Modal Approximations: Theory and an Application to Reactor Physics, MIT Press, Cambridge (1967).

23. Kaplan, S., "On the Best Method for Choosing the Weighting Functions in the Method of Weighted Residuals," Trans. ANS, 6, 3 (1963).

24. Case, K. M., and Zweifel, P. F., Linear Transport Theory, Addison-Wesley, Reading (1967).

25. deBoor, C., and Lynch, R. E., "On Splines and their Minimum Properties," J. of Math. and Mechanics, 15, 953 (1966).

26. Pomraning, G. C., "A Derivation of Variational Principles for Inhomogeneous Equations," Nucl. Sci. Eng., 29, 220 (1967).

27. Devooght, J., "Higher Order Variational Principles and Iterative Processes," Nucl. Sci. Eng., 41, 399 (1970).

28. Goldstein, R., "Iterative Solutions to Reactor Equations," USAEC Report CONF-670501-15, also BNL-12911 (1967).

29. Goldstein, R., and Cohen, E. R., "Theory of Resonance Absorption of Neutrons," Nucl. Sci. Eng., 13, 132 (1962).

30. Goldstein, R., and Sehgal, B. R., "Intermediate Resonance Absorption in Heterogeneous Media," Nucl. Sci. Eng., 25, 174 (1966).

31. Goldstein, R., "Intermediate Resonance Absorption for Multi-Nuclide Systems," Nucl. Sci. Eng., 30, 304 (1967).

32. Wachspress, E. L., Iterative Solution of Elliptic Systems, Prentice Hall, Englewood Cliffs (1966).

33. Crawford, B. W., and Friedman, J. P., "A Synthetic Transport Theory Method," USAEC Report CONF-710302, 625 (1971).

34. Gelbard, E. M., and Hageman, L. A., "The Synthetic Method as Applied to the S-N Equations," Nucl. Sci. Eng., 37, 288 (1969).

35. Kopp, H. J., "Synthetic Method Solution of the Transport Equation," Nucl. Sci. Eng., 17, 65 (1963).

36. Price, W. G., Jr., "A Derivation of Some P-1 Approximations to the Neutron Transport Equation," unpublished manuscript (1967).

37. Pomraning, G. C., and Clark, M., "The Variational Method Applied to the Monoenergetic Boltzmann Equation, Part I," Nucl. Sci. Eng., 16, 147 (1963).

38. Kaplan, S., Davis, J. A., and Natelson, M., "Angle-Space Synthesis -- an Approach to Transport Approximations," Nucl. Sci. Eng., 28, 364 (1967).

39. Buslik, A. J., "Interface Conditions for Few-Group Neutron Diffusion Equations with Flux-Adjoint Weighted Constants," Nucl. Sci. Eng., 32, 233 (1968).

40. Wachspress, E. L., and Becker, M., "Variational Synthesis with Discontinuous Trial Functions," USAEC Report ANL-7050, p. 191 (1965).

41. Wachspress, E. L., "Numerical Studies of Multichannel Variational Synthesis," Nucl. Sci. Eng., 26, 373 (1966).

42. Stacey, W. M., "Variational Flux Synthesis Methods for Multigroup Neutron Diffusion Theory," USAEC Report FRA-TM-20 (1971).

43. Calame, G. P., Federighi, F. D., and Ombrellaro, P. A., "A Two-Mode Variational Procedure for Calculating Thermal Diffusion Theory Parameters," Nucl. Sci. Eng., 10, 31 (1961).

44. Calame, G. P., and Federighi, F. D., "A Variational Procedure for Determining Spatially Dependent Thermal Spectra," Nucl. Sci. Eng., 10, 190 (1961).

45. Lorenzini, P. G., and Robinson, A. H., "Solutions of the Diffusion Equation by the Spectral Synthesis Method," Nucl. Sci. Eng., 44, 27 (1971).

46. Neuhold, R. J., "Multiple Weighting Functions in Fast Reactor Space-Energy Synthesis," Nucl. Sci. Eng., 43, 74 (1971).

47. Vaughan, E. V., Rose, P. F., and Hausknecht, D. F., "Spectrum Synthesis of Fast Reactor Analysis," USAEC Report AI-AEC-12820 (1969).

48. Cockayne, J. E., and Ott, K. O., "Successive Space-Energy Synthesis for Neutron Fluxes in Fast Reactors," Nucl. Sci. Eng., 43, 159 (1971).

49. Froelich, R., "Anomalies in Variational Flux Synthesis Methods," Trans. ANS, 12, 150 (1969).

50. Adams, C. H., and Stacey, W. M., "An Anomaly Arising in the Collapsed Group Flux Synthesis Approximation," Trans. ANS, 12, 151 (1971)

51. Murley, T. E., and Williamson, J. W., "Space-Energy Synthesis Techniques for Fast Reactor Calculations," Trans. ANS, 11, 174 (1968).

52. Meneley, D. A., Kvitek, L. C., and O'Shea, D. M., "MACH-1, A One-Dimensional Diffusion-Theory Package," USAEC Report ANL-7223 (1966).

53. Baller, D. C., "The FAIM Code", USAEC Report NAA-SR-7137 (1962).

54. Forsythe, G. E., and Moler, C. B., Computer Solution of Linear Algebraic Systems, Prentice-Hall, Englewood Cliffs (1967).

55. Cockayne, J. E., "Improved Fast Reactor Space-Energy Synthesis," Ph.D. Thesis, Purdue University (1970).

56. Wilkinson, J. H., The Algebraic Eigenvalue Problem, Oxford University Press, London (1965).

57. Broomfield, A. M., et al., "ZPR-III Assembly 48: Studies of a Dilute Plutonium-Fueled Assembly," USAEC Report ANL-7320, p. 205 (1966).

58. Davey, W. G., "Intercomparison of Calculations for a Dilute Plutonium-Fueled Fast Critical Assembly," USAEC Report ANL-7320, p. 57 (1966).

59. Moler, C.B., and Stewart, G.W., "An Algorithm for the Generalized Matrix Eigenvalue Problem $Ax = Bx$," Stanford University Report STAN-CS-232-71 (1971).

APPENDIX I

THE MACH-1 CODE

The computer code MACH-1 is really a coupled set of programs incorporating principally the features of the AIM-6 one-dimensional multigroup diffusion theory code and the DEL perturbation analysis code. The reference report MACH-1 is ANL-7223, by D.A. Meneley, L.C. Kvitek, and D.M. O'Shea; for AIM-6 the reference is NAA-SR-MEMO-9204, by H.P.Flatt and D.C. Baller; for DEL the reference is ANL-7052 by L.C. Kvitek.

MACH-1 solves almost the same set of finite difference multigroup equations described in Chapter VII: a slightly different (less accurate?) equation is used for points lying on an interface. The solution is carried out by assuming the formula $\phi_{n+1} = (\phi_n + \beta_{n-1})/(1 + \delta_n)$ is valid. A forward sweep is made through the points n = 1 to N to compute β_{n-1} and δ_n and then a backward sweep is made to compute each ϕ_n. MACH-1 will also compute the adjoint flux.

Source decks for this code are available as program #262 from the Argonne Reactor Code Center. One version is written in FORTRAN for the CDC-3600 - this presumably is the original version and should work properly. A second version, written in FORTRAN for the CDC-6500, is available (7000 + cards) but is <u>not reliable</u>. This version is the basis of the MACH/360 code, which is (supposed to be) MACH-1 plus a Spectral Synthesis option, adapted for the IBM System/360 FORTRAN.

APPENDIX II

THE MACH/360 CODE

MACH/360 is a revision and extension of the MACH-1 code. Several errors existing in the CDC-6500 version of MACH-1 were corrected, and in addition an attempt was made to convert all the special CDC FORTRAN statements to equivalent IBM System/360 FORTRAN-IV level G. This conversion was not tested thoroughly in all the various MACH options. MACH/360 is as independent of special system features as possible: necessary interactions take place through easily identified (and replaceable) service subroutines.

MACH/360 allows the calculation of the finite difference multigroup diffusion theory flux (from the equations presented in Chapter VII), or alternatively the calculation of the Spectral Synthesis approximation to that flux (also as described in Chapter VII). To reduce the core storage requirements a synthesis calculation may use no more than 50 mesh intervals (the regular flux calculation allows 150); synthesis is performed with 1 to 5 sets of flux and adjoint trial spectra read as input, while the regular flux calculation may use 1 to 28 energy groups. Neither the reduction of the F-D M-G equations to obtain the synthesis equations nor the subsequent solution of these equations have been written in particularly efficient FORTRAN - the synthesis option is clearly only an experimental tool. An additional special option in MACH/360 allows the use of Wielandt's Method (described in Chapter VIII) to accelerate the convergence of the Spectral Synthesis solution.

Source decks for MACH/360, with appropriate documentation, will be made available to the Argonne Reactor Code Center in early 1972.

APPENDIX III

THE MACHLIB CODE

This code is available from the Argonne Reactor Code Center as part of the MACH-1 code - its purpose is to produce the cross section library tape required for MACH-1 (and MACH/360) execution.

The CDC-6500 version of MACHLIB has been converted to IBM System/360 FORTRAN and will be returned to Argonne with MACH/360. The 6500 version has a number of features which have not been documented before but which will be described in the 360 version.

MACHLIB is distributed with a card version of the "ANL Set 224" cross section library. This is a pre-ENDF-B set of cross sections suitable for the analysis of fast reactor critical mock-ups, and is in the format required for input to MACHLIB.