TRANSIENT HEAT TRANSFER IN SOLIDS

Thesis by

Edward R. H. McDowell

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1964

ACKNOWLEDGEMENTS

Grateful acknowledgement is given to Dr. P. A. Longwell for initiating and stimulating my interest in this project and for his constant, patient assistance and encouragement, and helpful suggestions in carrying out the work and writing the thesis. I was supported during this time variously by Institute Scholarships, Teaching Assistantships, and National Science Foundation Fellowships. The figures, tables, and rough draft were prepared by my wife Joyce, and the final draft was typed by Mrs. Lucille Lozoya.

ABSTRACT

The accuracy of approximate calculations of transient temperature distributions in solids is studied, emphasizing easily applied methods for both irregular and regular solids with a heat-transfer coefficient boundary condition. General stability criteria are shown which are easily found and used, and which apply for regular or asymmetric networks.

Analytic solutions of the one-dimensional partial difference equation are found by z-transforms. A method with points located half an increment inside the surface gives the most accurate solutions. A graphical method is shown analytically to give approximations of engineering accuracy. A procedure is also presented for finding the differencing parameters which give a solution of a specified accuracy with the minimum calculations.

Transient temperature distributions of satisfactory accuracy can be calculated for irregular solids using a relatively coarse network, by locating points that describe the boundary a short distance within, rather than on, the boundary, and by making the interior network as regular as possible.

CONTENTS

	Page
Acknowledgements	ii
Abstract	iii
Table of Contents	iv
List of Tables	ix
List of Figures	xii
Nomenclature	xv
Chapter I: Introduction	1
Chapter II: Mathematical Formulation of Transient Temperature Distribution in Both Continuous and Difference Form	9
(A) Continuous or Exact Formulation	10
(B) Continuous Solution	15
(C) Approximate Formulation	18
(D) Solution of Approximate Formulations	32
(E) Conclusions and Summary	38
Chapter III: General Stability Criteria	43
(A) Stability--Review and Definitions	45
(B) Necessary and Sufficient Conditions for Stability and Non-Oscillatory Behavior	49
(C) Selection of $\Delta\tau$ and γ Based on Stability	57
(D) Numerical Example	68
(E) Alternate Calculation for Implicit Methods	78
(F) Conclusions	82
Tables	84
Figures	88

CONTENTS (continued) Page

Chapter IV: Solutions of the One-Dimensional Partial Difference Equation for Diffusion in Cartesian Coordinates Using z-Transforms . 92

(A) Introduction 92

(B) Difference Formulation 98

1. Continuous Problem and Initial Condition. 99

2. Meshes 101

3. Mesh $\Delta\xi$ 103

Method G 105

Method A 107

4. Mesh $\Delta\xi/2$ 110

Method C 112

Method F 113

5. Other Methods 114

6. Graphical Solution 114

7. Averaged Methods 116

Interpolation 120

Modifications to Step Changes in Boundary Conditions . . . 122

Conclusions and Averages to be Associated with Methods G, A, C, and F 125

(C) z-Transforms 127

(D) z-Transforms and Partial Difference Equations--Example Problems 138

1. Problem I--Finite H, Method G . . 139

Application of Initial Conditions . 143

Solution of Transformed Difference Equation 147

Inversion of Transform . . . 150

CONTENTS (continued)

	Page
2. Problem I, Averaged Method G--Ramp-Step Equivalence	159
3. Problem II--Infinite H, Method G	162
4. Problem V--Semi-Infinite Solid, Method G, Branch Cut Singularity	166
5. Conclusions	170
(E) Comparison of Methods--z-Transform Solutions	175
1. Particular Solutions	179
Conclusions and Summary--Particular Solutions	182
2. Transient Solution	182
3. Characteristic Equations and Roots	195
Conclusions and Summary--Trigonometric Roots	209
4. Damping Factors	210
Stability and Non-Oscillatory Behavior	212
Graphical Solution	216
Accuracy of Damping Factors	221
Conclusions and Summary--Damping Factors	233
5. Eigenfunctions and Eigenvectors	237
Conclusions and Summary--Eigenvectors	241
6. Initial Vectors	242
Unmodified Initial Vector Components	244
Graphical Solutions--Effective Initial Vector	250
Averaged Methods	254
Conclusions and Summary--Initial Vectors	258
7. Problem V--Complete Problem	263
8. Selection of Differencing Parameters	264

CONTENTS(continued)

	Page
(F) Conclusions and Summary--Chapter IV	279
(G) Suggestions for Future Work	290
Tables	295
Figures	333
Chapter V: Accuracy of Transient Solutions Using an Asymmetric Network of Nodes	350
(A) Introduction	350
(B) Definition of Discretization Error	357
(C) Propagation of Discretization Error	360
1. Conclusions	378
(D) Discretization Error	380
1. Division of Discretization Error with Time and Space Differencing Parameters	380
2. Time Discretization Error	387
(E) Space Discretization Error	389
1. Expansion for Space Discretization Error, Interior Node	389
2. General Comments--Space Discretization Error	398
3. Consistency	399
Practical Aspects of Consistency	408
Possible Ways of Eliminating Zero-Order Error Terms	410
(F) Accuracy of Numerical Results Based on an Asymmetric Network	414
1. Problems Studied and Network Used	414
2. Transient Problem	417
Time Differencing Parameters	418
Selection of Time Differencing Parameters for an Asymmetric Network	423
Space Differencing Parameters	426

CONTENTS (continued)

	Page
3. Steady-State Problem	431
4. Summary of Conclusions--Numerical Studies	436
(G) Rules for Locating a Limited Number of Points	438
1. Location of Adjacent Nodes	440
Fluid Boundary Condition, $0 \leq h \leq \infty$	442
Interface Boundary Condition	447
2. Location of Interior Nodes	451
Nodes	451
Networks	453
Other Rules	458
3. Summary--Location of Nodes	461
(H) Summary and Conclusions	464
(I) Suggestions for Future Work	468
Tables	472
Figures	490
References	508
Proposition I	511
Proposition II	516
Proposition III	520
Proposition IV	524
Proposition V	529

TABLES

		Page
III-1	Conductance/Capacitance Matrix, Y/A 1/8 - Square Cross-Section	84
III-2	Eigenvalues, Norms, and Stability Criteria 1/8 - Square Cross-Section	85
III-3	Symmetrized Y/A Matrix, W = D Y D 1/8 - Square Cross-Section	86
III-4	U Matrix, $U = F W F^{-1}$, Similar to W and Y/A Matrices 1/8 - Square Cross-Section	87
IV-1	Generalized Method G	295
IV-2	Generalized Method A	297
IV-3	Generalized Method C	299
IV-4	Generalized Method F	301
IV-5	Graphical Methods	303
IV-6	Averaged Methods	304
IV-7	z-Transform Operation Pairs	305
IV-8	z-Transform Function Pairs	307
IV-9	Problems to be Approximated	309
IV-10	Continuous, Analog, and Difference Solutions in Difference Form	310
IV-11	Particular or Steady-State Solutions	311
IV-12	Characteristic Equations - Problem I	313
IV-13	Characteristic Equations - Problems II, III, and IV	314
IV-14	Damping Factors	315
IV-15	Difference in Logarithms of Damping Factors	316
IV-16	Eigenvector Matrices	318
IV-17	Components of Initial Vector, Problem I	320
IV-18	Components of Initial Vector, Problem II	321

TABLES (continued)

		Page
IV-19	Components of Initial Vector, Problem III . .	322
IV-20	Components of Initial Vector , Problem IV. .	323
IV-21	Effective Initial Vector Components, Graphical Solutions, Problem II . . .	324
IV-22	Comparison of Graphical Method F . . .	325
IV-23	Comparison of Solutions, Graphical Method A .	326
IV-24	Series to Multiply by "Unaveraged" Initial Vector Components to Obtain Series Representation of Initial Vector Components for Averaged Methods	327
IV-25	Averaged Initial Vector Components--Problem II	328
IV-26	Averaged Initial Vector Components--Problem III	329
IV-27	Series to Multiply by Oscillatory Series in Effective Initial Vector to Obtain Oscillatory Series in Averaged Solution . .	330
IV-28	Averaged Effective Initial Vector Components--Problem II	331
IV-29	Complete Solution--Problem V	332
V-1	Conductance Matrix (Y) and Inverse Conductance Matrix (Y^{-1}) Transient Problem, Chapter III	472
V-2	Conductance Matrix (Y) and Inverse Conductance Matrix (Y^{-1}) Steady-State Problem, Chapter V . . .	473
V-3	Conductance Matrix (Y) and Inverse Conductance Matrix (Y^{-1}) Shape I, see Longwell (9)	474
V-4	Conductance Matrix (Y) and Inverse Conductance Matrix (Y^{-1}) Shape II, see Longwell (9)	475
V-5	Conductance Matrix (Y) and Inverse Conductance Matrix (Y^{-1}) Square Grid - Specified Surface Temperature .	476

TABLES (continued)

		Page
V-6	Conductance Matrix (Y) and Inverse Conductance Matrix (Y^{-1}) Regular Rectangular Grid - Specified Surface Temperature	474
V-7	Steady-State Study, Square Shape Continuous, Approximate Solutions, Errors, and Discretization Error	478
V-8	Taylor Series Expansion $(T_j - T_i)$	479
V-9	Space Discretization Error	480
V-10	Errors - Transient Problem - Chapter III .	487

FIGURES

		Page
III-1	Layout of nodes on square cross-section . .	88
III-2	Behavior of approximate solutions, $\gamma = 0$ . .	89
III-3	Behavior of approximate solutions, $\gamma = \frac{1}{2}$ and $\gamma = 1$	90
III-4	Calculation at limit of stability, $\gamma = 0$, $\Delta\tau = 0.001387$, with oscillatory boundary conditions	91
IV-1	Comparison of graphical methods G and C . .	333
IV-2	Meshes $\Delta\xi$ and $\Delta\xi/2$	334
IV-3	Other meshes	335
IV-4	Example of a sampled or difference signal .	336
IV-5	Contour integrations for inverting z-transforms	337
IV-6	Semi-logarithmic graph of term in transient solution versus time	338
IV-7	Characteristic equation, continuous solution, first ten roots	339
IV-8	Characteristic equation, method G, real roots for S=5, first five roots	340
IV-9	Characteristic equation, method C, S=5, five roots	341
IV-10	Characteristic equation, method G, imaginary root (sixth root) for S=5 . . .	342
IV-11	Characteristic equation, graphical method A, S=5, all six roots	343
IV-12	Error in first two roots, method G, graphical method A, and method C, S=5 . .	344
IV-13	Error in roots three, four, and five, S=5 . .	345
IV-14	Sufficient conditions such that $q_{min} \geq -Y$ for methods where minimum norm of Y/A matrix is equal to or less than $4S^2$, or all trigonometric roots α_j are real	346

FIGURES (continued) Page

IV-15 Semi-logarithmic graph of term in transient solution versus time showing effective initial vector component for graphical method A 347

IV-16 Coefficient of μ^4 term in series for the logarithm of $e^{-\nu^2}/q^{1/\Delta\tau}$ as a function of the differencing parameters 348

IV-17 Coefficient of μ^6 term in series for the logarithm of $e^{-\nu^2}/q^{1/\Delta\tau}$ as a function of the differencing parameters 349

V-1 MacNeal's approximation for surface node on curved boundary with finite non-zero heat-transfer coefficient 490

V-2 Relationship between nodal area and the product of the norm of the Y^{-1} matrix and the nodal area, one-dimensional problem, regular network; infinite h 491

V-3 Relationship between nodal area and the product of the norm of the Y^{-1} matrix and the nodal area, two-dimensional problem, regular rectangular network; infinite h 492

V-4 General asymmetric node, two-dimensional problem showing geometry in dimensionless coordinates 493

V-5 Space discretization error, node J (see Table V-9), sequence of geometrically similar nodes 494

V-6 Space discretization error, node L (see Table V-9), sequence of geometrically similar nodes 495

V-7 Comparison between annular network and network of trapezoidal nodes for cylindrical coordinates based on MacNeal's rules; annular network is usual approximation . . 496

FIGURES (continued) Page

V-8	Temperature at node 1 (Fig. III-1) versus dimensionless time; short time	497
V-9	Temperature at node 1 (Fig. III-1) versus dimensionless time; intermediate time . .	498
V-10	Temperature at node 1 (Fig. III-1) versus dimensionless time; long time	499
V-11	Temperature at node 6 (Fig. III-1) versus dimensionless time; short time	500
V-12	Temperature at node 6 (Fig. III-1) versus dimensionless time; long time	501
V-13	Modification to MacNeal's approximation for adjacent node on curved boundary with finite non-zero heat-transfer coefficient; analogous to mesh $\Delta\xi/2$	502
V-14	MacNeal node for interfacial boundaries between uniform solids of different thermal properties	503
V-15	Continuous temperature distribution across interfacial boundary of uniform solids with very different thermal properties .	504
V-16	Modified nodes for interfacial boundaries between uniform solids of different thermal properties	505
V-17	Network to approximate solid with curved boundary; uses only square interior nodes . .	506
V-18	Network to approximate solid with curved boundary, distorted triangular network of distorted hexagonal nodes	507

NOMENCLATURE

A	S x S diagonal matrix with dimensionless thermal capacity along diagonal
A_i	$\frac{1}{4L^2\sigma_0 C_{p\,0}} \sum_{\substack{j=1 \\ j\neq i}}^{S} r_{ij}\,\ell_{ij}\,\sigma_i\,C_{pi}$, dimensionless thermal capacity of ith node; also i^{th} diagonal element of A matrix
A_{f_i}	surface area between fluid and node i through which energy flows
a	vector of continuous Fourier coefficients determined to fit continuous initial condition $T(\xi, 0)$, $0 \leq \xi \leq 1$
a_j	j^{th} continuous Fourier coefficient in vector a
B	continuous eigenvector arranged in matrix
$b_j(x)$	j^{th} orthogonal function to fit continuous boundary conditions
C	matrix whose columns are eigenvectors of Y/A matrix
C_p	heat capacity, Btu/°F-lb
c_i	i^{th} column eigenvector of difference system $\lambda_i c_i = \frac{Y}{A} c_i$
D	S x S diagonal matrix with $1/A_i$ in diagonal position
D(z)	part of denominator of z-transform that goes to zero as z goes to a pole
D or D_j	replacement factor for $S\psi$ or υ in approximate characteristic equation

d	differential operator
d_n	discretization error vector at time n, defined by equation V-4
$d_{i,n}$	discretization error vector associated with i^{th} node at time n
E	finite difference operator, $E\, t_n = t_{n+1}$
e	base of napierian or natural logarithm
e_n	round-off error vector occurring at time n, defined by equation V-8
F	constant in undetermined parameter solution
F	diagonal transformation matrix
F_C	continuous characteristic equation
F_D	characteristic equation for approximate method
$f(\tau)$	function of time
$\bar{f}(z)$	z-transform of difference function, f_n
G	constant in solution of boundary value difference equation
g	$C^{-1}(t_0 - t_{P,0}) = C'A(t_0 - t_{P,0})$ difference initial vector that is vector of coefficients required to fit the initial condition
g_j	j^{th} component of initial vector
$\bar{g}(z)$	z-transform of difference function g_n, used only in z-transform table

H	hL/k = dimensionless heat transfer coefficient
h	heat transfer coefficient, Btu/hr, ft^2, °F
I	identity matrix which is a diagonal matrix with 1 in the diagonal positions
J	limit of the integrals as defined in equations IV-194 and IV-196
J	number of terms in transient solution
K	$k/\sigma C_p$ = thermometric conductivity, ft^2/hr
k	thermal conductivity, Btu/hr, ft, °F
L	reference length, thickness of slab for one-dimensional problem
ℓ_{ij}	distance between i^{th} and j^{th} node
ℓ_{i,f_i}	distance along normal to boundary from surface to i^{th} node
M	minimum of row norm of Y/A, column norm of Y/A, and norm of W
N	number of non-zero multiplications to calculate to τ_1
N	number of nodes with non-zero conductances to the ith node
N(z)	part of z-transform that remains analytic as z goes to a pole, i.e., as D(z) goes to zero
n	number of time increments; as a subscript occurring at end of nth increment

n	direction normal to boundary of solid directed toward interior
P	constant in solution of boundary value difference equation
p	index of summation, an integer
p	proportionality constant
Q	S x S diagonal matrix of difference eigenvalues dia $q_1\ q_2 \cdots\ q_S$
$\vec{\dot{q}}$	heat flux per unit area
$\vec{\dot{\underline{q}}}$	total heat flux
q_j	j^{th} difference eigenvalue
R	total number of temperature nodes in problem including fluid and boundary nodes
r	$K\Delta\theta/(\Delta x)^2 = \Delta\tau/(\Delta\xi)^2$ = difference modulus
r_{ij}	length of perpendicular bisector between i^{th} and j^{th} nodes
S	total number of temperature points whose temperatures are to be found as functions of time; dimensions of Y matrix and number of elements in temperature vector. Note that in Chapter IV, for methods based on mesh $\Delta\xi$, S is number of subdivisions, which does not always conform to number of active temperature points.
S_1	variable points in ξ direction, regular network, two-dimensional problem

S_2	variable points in η direction, regular network, two-dimensional problems
s	Laplace transform complex variable
T	continuous temperature
$T(\xi, \eta, \tau)$	exact continuous solution temperature, two-dimensional problem
$T(\xi, \tau)$	$T(m\Delta\xi, n\Delta\tau)$ = continuous solution, one-dimensional problem
$T(\tau)$	S-dimensional exact continuous solution vector for certain temperature points
$T_f(\tau)$	continuous fluid temperature
t_n	vector of difference temperatures at end of nth time increment
$t_{i,n}$ or $t_{m,n}$	difference temperature of ith or m^{th} node
$t_{f,n}$	temperature of fluid as used in difference calculation
U	S x S diagonal matrix used to reduce bound on $\lvert\lambda_{min}\rvert$
U	total number of off-diagonal elements above diagonal in Y/A matrix, $= \sum_{i=1}^{S} u_i$
$\bar{u}(z)$	$z^{p-1}\bar{f}(z)$ = z-transform modified so that z^n occurs in numerator
u_i	number of off-diagonal elements above diagonal in i^{th} row of Y/A matrix

$\bar{u}$ — average number of off-diagonal elements above diagonal in Y/A matrix, = U/S

V — volume or dimensionless volume

V — bound on elements in error vector

$v_{m,n}$ — $T_{m,n} - t_{m,n}$ = element in error vector v_n

v_n — $T_n - t_n$ = error vector, difference between continuous solution vector and approximate solution vector

W — $D \frac{Y}{A} D$ = S x S symmetric matrix with same norms, diagonal elements, and eigenvalues as Y/A; off-diagonal element is $(y_{ij}/\sqrt{A_i A_j})$; i^{th} diagonal element is $\frac{-1}{A_i} \sum_{\substack{j=1 \\ j \neq i}}^{R} y_{ij}$

x — x-direction in Cartesian coordinates

Y — S x S matrix of dimensionless conductances; off-diagonal element is y_{ij}; i^{th} diagonal element is $-\sum_{\substack{j=1 \\ j \neq i}}^{R} y_{ij}$

$\frac{Y}{A}$ or Y/A — $A^{-1} Y$ = S x S matrix used to approximate Laplacian operator; off-diagonal element in i^{th} row, j^{th} column (y_{ij}/A_i); i^{th} diagonal element is $-\frac{1}{A_1} \sum_{\substack{j=1 \\ j \neq i}}^{R} y_{ij}$

$\frac{Y_B}{A}$ or Y_B/A — $A^{-1} Y_B$ = matrix of thermal conductances to nodes of known temperature as a function of time; element in i^{th} row, j^{th} column (y_{ij}/A_i) $\quad i = 1, \ldots, S$ $\quad j = S+1 \ldots R$

y	y-direction in Cartesian coordinates
y_{ij}	dimensionless conductance between nodes i and j
y_{i,f_i}	dimensionless conductance between i^{th} adjacent node and i^{th} fluid node
z	complex transform variable for difference equations

Subscripts

Ave	referring to average
B	referring to boundary
E	effective initial vector or initial vector component for graphical solution
Ex	explicit calculation, $\gamma = 0$
f	referring to surrounding fluid, f_0 at boundary 0, f_S at boundary S
f_i	refers to fluid opposite adjacent node i
G	graphical solution, $\gamma = 0$, $r = \frac{1}{2}$
H	homogeneous solution
I	referring to interpolation
Im	implicit calculation, $\gamma \neq 0$
i	referring to i^{th} row in matrix; i^{th} temperature node or i^{th} diagonal position in diagonal matrix; index of summation; i^{th} component of vector
i	refers to interface in equations V-88 and V-89
J	last significant term in transient solution

Subscripts

j	refers to j^{th} column in matrix; j^{th} element in vector; j^{th} diagonal position in diagonal matrix; index of summation; and j^{th} temperature node adjacent to i^{th}
k	refers to averaged solution, $t_{n+k\ Ave}$, time at which applied is $(n+k)\Delta\tau$; $0 \leq k \leq 1$
m	refers to temperature point in one dimension in a regular mesh where m has relationship to spatial position, i.e. in Figure IV-2 $\xi = m\Delta\xi$, m = 0, $\frac{1}{2}$, 1, . . . , S-$\frac{1}{2}$, S depending on case
n	subscripted quantity applies at end of nth time increment
P	particular solution
p	index of summation, summation variable
R	the Rth point
RS	refers to ramp-step modification
S	the Sth point
SS	steady state
0, 1, 2	zero, 1st, 2nd, etc. point, or time increment
I	refers to solid I with thermal properties k_I and $(\sigma C_p)_I$
II	refers to solid II with thermal properties k_{II} and $(\sigma C_p)_{II}$

Greek Letters

α	trigonometric parameter in difference characteristic equation
α	geometric parameter for trapezoidal nodes
β	weighting given to energy in from fluid at time (n+1)
β_{ij}	angle between positive ξ -axis and the connector ℓ_{ij}
Γ	closed curve for integration in complex plane
γ	weighting given to temperature vector at t_{n+1} in implicit formulation, $0 \leq \gamma \leq 1$
γ_o	$\frac{(6r-1)}{12r}$ or $\frac{1}{2} - \frac{1}{12S^2\Delta\tau}$
Δ	difference operator
δ	radius of circles about branch points in complex plane, Figure IV-5
$\delta_{j\rho}$	Kronecker delta function = $\begin{cases} 1 & j = \rho \\ 0 & j \neq \rho \end{cases}$
ϵ	small number
ϵ	distance from branch cut in complex plane, Figure IV-5
$\epsilon_{i,n}$	component of round-off error vector at i^{th} node at time n
ζ	dimensionless z-coordinate = z/L
η	dimensionless y-coordinate = y/L
θ	time, hr
θ	angle coordinate for cylindrical coordinates
Λ	diagonal matrix of eigenvalues of Y/A matrix

λ_i eigenvalue of Y/A matrix

λ_{ij} distance between nodes i and j in dimensionless coordinates, $= \ell_{ij}/L$

μ multiplicity of a pole

μ_{ij} element in i^{th} row j^{th} column of Y/A matrix

μ_{00} diagonal element of Y/A matrix with minimum absolute value

$\mu_{\rho\rho}$ diagonal element of Y/A matrix with maximum absolute value

μ_j $S\alpha_j = \alpha_j/\Delta\xi$ = trigonometric parameter root in characteristic equation for approximate solution in continuous form

ν_j trigonometric parameter root of characteristic equation for continuous solution

ν_{ij} element in i^{th} row and j^{th} column of inverse conductance matrix Y^{-1}

$e^{-\nu^2}$ diagonal matrix with $e^{-\nu_j^2}$ as the jth diagonal element

Φ scaling term, $\underset{m_1, m_2}{\text{MAX}} \left| t_{m_1,0} - t_{m_2,0} \right|$ where m_1 and m_2 take on all values of mesh and fluid temperature flux

φ weighting given to boundary node heat loss to interior node, Chapter IV

φ_j elements in F diagonal matrix

$\varphi_{i,n}$ part of discretization error associated with time differencing

Greek Letters

Ψ	angle parameter for solution of steady-state problem, equation V-61
ψ_j	$\nu_j \Delta\xi = \nu_j/S$ = trigonometric parameter in characteristic equation for continuous solution in difference form
ξ	dimensionless x-coordinate = x/L
π	ratio circumference to diameter of circle = 3.1416 ...
ρ	dimensionless radial coordinate in cylindrical coordinates
ρ_j	residue at j^{th} pole
ρ_{ij}	length of perpendicular bisectors between nodes i and j in dimensionless coordinates, = r_{ij}/L
ρ'_{ij}	solution to system of equations V-77 to V-80, length to replace ρ_{ij} to eliminate zero-order error terms
σ	specific weight, lb/ft^3
$\sigma(\xi)$	weighting factor in orthogonal function relationship
τ	$K\theta/L^2$ = dimensionless time
Y	allowable limit on absolute value of q_{min} or upper bound on q_{min}
$\Omega(\xi,\tau)$	$\frac{4}{\pi}\sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)}\cos(2j+1)\pi\xi\, e^{-(2j+1)^2\pi^2\tau}$, function in continuous solution for specified surface temperatures
ω	angular frequency, radians/unit time
ω_0	frequency in difference notation = $\omega/\Delta\tau$

Miscellaneous Symbols

$\rightarrow$	designates a physical vector
$\vec{\nabla}$	operator del, gradient of scalar
$\vec{\nabla}\cdot$	divergence operator
$\vec{\nabla}'$	$L\ \vec{\nabla}\cdot$ = del operator in dimensionless coordinates
∇^2	Laplacian operator $= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$
∇'^2	$L^2\ \nabla^2$ = Laplacian operator in dimensionless coordinates
	ξ and η
∞	infinity
S'	applied to a matrix means its transpose
$\sum_{j=p}^{S}$	summation on j from p to S
$\int$	integral of
∂	partial differential operator
$\mathcal{J}$	imaginary part of a complex number
$\mathcal{L}$	Laplace transform of
$\mathcal{Z}$	z-transform of
$\leq$	less than or equal to
$\geq$	greater than or equal to
$\equiv$	is defined as
$\simeq$	approximately equal to
$\rightarrow$	approaches
$\not\rightarrow$	does not approach
$\lvert\ \rvert$	absolute value of
$\lVert\ \rVert$	norm of

Miscellaneous Symbols

cos	cosine of
ctn	cotangent of
det	determinant of
dia...	diagonal matrix = matrix whose elements are all zero except those along the diagonal
sin	sine of
tan	tangent of
ln	natural logarithm of
MAX	maximum of
min	minimum of

CHAPTER I

INTRODUCTION

One of the problems often encountered by the chemical engineer is to find the transient temperature distribution in a solid where the initial temperature distribution is known and the boundary conditions are specified. This involves finding either the exact solution to the governing partial differential equation, commonly called the diffusion equation, or obtaining a reasonably accurate estimate of its solution for the specific conditions of the problem. To obtain the transient temperature distribution numerically, using either a known exact solution or an approximate method, usually requires a great many tedious, time-consuming calculations. Since most problems of engineering importance are too complicated for analytic solution, an approximation is necessary. Even when the exact solution is available (1), or can be found, it often requires more calculations to find the numerical solution than does an approximate calculation. Therefore, as an engineer is interested in finding a numerical solution that is sufficiently accurate for his purposes with a minimum of effort, an approximate method can be a practical and economical tool for him to use.

This study is on approximate methods of finding numerical solutions to the diffusion equation that can be applied to a wide range of problems, that are of reasonable accuracy, and that are relatively easy to use. Although many of the methods discussed here require a high-speed computer, for problems in one space dimension a graphical method is shown analytically to give reasonably accurate results. If a desk calculator is available, further significant improvements can be made without the cost in time and money of programming and running a digital computer. The methods that require a high-speed computer are for a very general problem and, in contrast to most methods, are based on a relatively coarse difference grid. In all cases, particular emphasis is placed on problems where the boundary conditions involve a finite heat-transfer coefficient.

For chemical engineers the diffusion equation is probably the most important partial differential equation. In addition to governing the transient temperature distributions in solids, it, or a closely related equation, governs certain situations in diffusional mass transport, fluid mechanics, and energy transfers in fluid systems. Although this study is concerned only with the linear diffusion equation without sources or sinks (equation II-4), generalizations to many of the related equations are possible.

To use an approximate numerical method, the derivatives are replaced with differences for a finite number of points within and on

the solid boundary. This approximation, together with similarly approximated boundary conditions, gives either a linear system of ordinary differential equations or a system of algebraic equations to be solved to determine the transient temperature distribution. The system of ordinary differential equations can be solved almost instantaneously with an electric analog computer; the algebraic equations can be solved rapidly for successive time increments on a digital computer. Complicated initial temperature distributions and/or complicated boundary conditions for which it is impossible to find the exact solution are usually easily handled by the approximate method.

The advent of the high-speed digital computer, because of its capability of carrying out an enormous amount of calculations in an extremely short time, has made practical many of the classical methods of approximation and has greatly stimulated studies of difference methods for obtaining approximate numerical solutions to the diffusion equation. Starting in 1950, several important articles about the approximate methods for solving the diffusion equation have been published. O'Brien et al. (2) showed, both by theoretical arguments using a method of von Neumann, and by a numerical check, that the unstable behavior of certain approximations to the diffusion equation is a property of the difference scheme used and is not caused, but can be aggravated, by round-off error. In 1956 Lax and Richtmyer (3) showed that, for a wide range of linear initial-value problems including the diffusion equation,

in principle the approximate solutions can be made as accurate as desired, by using smaller and smaller difference increments, if the approximate solution remains stable and if the approximation is "consistent."

Since 1955, results from matrix algebra have been used to attack the stability problems by Todd (4) and Franklin (5), and accuracy problems by Varga (6,7). In addition, numerous other articles have appeared and many reviews of approximate methods for the diffusion equations are found in recent texts. Most of these articles and discussions are concerned with problems where the geometry is such that it may be represented by temperature points located in a regular, repeated geometric pattern, and where the boundary conditions are either a specified surface temperature (infinite heat-transfer coefficient) or adiabatic (zero heat-transfer coefficient). The important published results can be discussed more appropriately in sufficient detail later.

The two major goals of this research have been (1) to devise a reasonably accurate approximate method for calculating transient temperature distributions in solids with irregular boundaries without using an extremely fine space grid; and (2) to develop good methods of approximating the boundary condition equations for the practical case of finite heat-transfer coefficient.

The first problem has been studied by MacNeal (8), who suggested using an asymmetric location of the temperature points so that they

could be located on the irregular boundary. He also presented a set of rules for approximating the Laplacian operator. Longwell (9) has adapted this scheme to an explicit method for stepping out the transient temperature distribution. The asymmetric location of temperature points as applied to the transient problem raises two difficulties. First, since the points are not located regularly, the usual methods of finding stability criteria do not work. Second, errors are introduced by asymmetric location of the points. For example, 20,000 to 100,000 points have been used (6) for the steady-state solution in two dimensions. This shows the practical importance of obtaining a good approximation for the Laplacian which does not require an extremely fine space grid for irregular solids. Using a grid with this number of points is impractical or uneconomic for the transient problem. Moreover, for problems in three space dimensions or for a simulation problem where the transient temperature calculation is only a small part of the total problem, a good approximation using a coarse grid would be very desirable.

The most realistic boundary conditions for solids are those which have a finite heat-transfer coefficient. In spite of this, essentially no theoretical studies have been made of good methods to approximate this "third" (10) boundary condition. There are few practical discussions of this type of boundary condition.

The next chapter contains mostly introductory material upon which the following chapters are based. The problem of transient

temperature distribution is formulated as the solution to the partial differential equation of diffusion with its initial and boundary conditions. The MacNeal-Longwell (9) method is used to formulate the approximate equations. The analytic solutions are shown in symbolic form for both formulations. The emphasis is on pointing out the many similarities between the continuous and the approximate formulations. These similarities are in the derivations of both the continuous and approximate equations, in the form of their solutions, and in properties of the continuous and difference operators.

Chapter III presents the stability criteria for an approximate solution based on a network of points located either regularly or irregularly. An example problem for transient heat transfer in two space dimensions is also shown to check these results. This problem shows that an asymmetric point distribution does give a reasonably accurate solution. In Chapter IV the accuracy of approximations based on uniform mesh spacing for a problem in one space dimension in Cartesian coordinates is discussed in detail. To do this a z-transform method is developed for the solution of the partial difference equation of diffusion. Several methods for the approximation for boundary conditions with finite heat-transfer coefficients are presented. The accuracy of these methods is compared analytically for both generalized methods and graphical methods. Chapter V gives a detailed discussion of errors introduced by the asymmetric distribution of points for problems in two

space dimensions, and gives a general set of rules for location of the points.

Before starting the actual study several general comments should be made about the nomenclature. Although every effort has been made to try not to use the same symbol in two ways; to have a representation for matrices, vectors, and scalars; and yet to use standard chemical engineering nomenclature; many conflicts can and do occur. Capital Roman letters are generally used to represent matrices, small Roman letters to represent vectors, and Greek letters for scalars. Sometimes in the sequence of matrices, vectors, and scalars, when this has not been possible, an element of a matrix or a vector may be designated with a small Roman letter with a double or single subscript, respectively.

The main exception to the sequence of capital and small Roman letters, and Greek letters, is that standard chemical engineering nomenclature has been maintained, which gives rise to some inconsistencies. Also, in matrices, i refers to the row and j to the column. The subscripts i and m are used to designate temperature points, the distinction being that m refers to a uniform network so that its numerical value has a physical meaning. The letter T is used for the exact continuous solution of the partial differential equation and t for an approximate solution. The two letters represent vectors evaluated at a time if they are shown, $T(\tau)$ and t_n with elements $T_i(\tau)$ and

$t_{i,n}$. In using these symbols, the first subscript indicates an increment of space and the second time an increment of time; if only one is used, the quantity represents a vector at the time corresponding to the subscript. In cases where ambiguity may result, the quantity is defined in the text before or just after its first use.

Several comments may also add to the clarity of the text. The formulation and solution of the partial differential equation are referred to as the continuous formulation and solution. The difference or analog solutions to the continuous problem are referred to as approximate solutions. Since numerical results may be obtained from either the continuous or approximate formulation, the terms "numerical method" and "numerical solution" have been avoided. "Similar" is used in its precise mathematical meanings, that is, similar matrices have the same eigenvalues, and similar nodes are those that have the same angles and proportionate lengths.

CHAPTER II

MATHEMATICAL FORMULATION OF TRANSIENT TEMPERATURE DISTRIBUTION IN BOTH CONTINUOUS AND DIFFERENCE FORM

In this chapter the transient temperature problem is formulated as the continuous solution to a partial differential equation. Also, a generalized version of the MacNeal-Longwell (9) approximate method is derived, and the equations suitable for stepping out the solution are shown. Further, the forms of an analytic solution of both the partial differential equation, and of the system of linear difference equations that result from the approximation, are shown. The dependence on the differencing parameters of the quantities in the solution to the difference approximation is indicated. Throughout this discussion the emphasis is on showing the likeness and analogies between the continuous and approximate methods. These analogies are in the derivation of the equations, in the properties of the operators occurring in the equations, in the form of the solutions for the equations, and in the properties of the terms in the equations.

(A) CONTINUOUS OR EXACT FORMULATION

The diffusion equation is simply derived, starting with Fourier's law for heat conduction in one direction

$$\dot{q} = - k \frac{\partial T}{\partial x} \tag{II-1}$$

or, in general, in three dimensions

$$\vec{\dot{q}} = - k \overrightarrow{\nabla T} \tag{II-2}$$

Applying the law of conservation of energy to a differential volume element which contains no sources or sinks, the net energy into the differential volume equals the accumulation, or, in vector notation, the time rate of change of temperature is proportional to the negative of the divergence of the heat flux.

$$\sigma C_p \frac{\partial T}{\partial \theta} = - \vec{\nabla} \cdot \vec{\dot{q}} \tag{II-3}$$

Combining equations II-2 and II-3 gives the diffusion equation,

$$\sigma C_p \frac{\partial T}{\partial \theta} = \vec{\nabla} \cdot k \vec{\nabla} T \tag{II-4}$$

which applies to the interior of the solid and is linear if k, σ, and C_p do not depend on T. For convenience, dimensionless variables are usually introduced which for three space dimensions are

$$\tau = \left(\frac{k}{\sigma C_p}\right)_0 \frac{\theta}{L^2} = K_0 \frac{\theta}{L^2} \tag{II-5}$$

$$\xi = x/L \tag{II-6}$$

$$\eta = y/L \tag{II-7}$$

$$\zeta = z/L \tag{II-8}$$

The diffusion equation in terms of these dimensionless variables becomes

$$\frac{\partial T}{\partial \tau} = \frac{1}{C_p \sigma K_0} \vec{\nabla}' \cdot k \vec{\nabla}' T \tag{II-9}$$

where $(1/L)\ \nabla' = \nabla$

For the one- and two-dimensional cases where the thermal conductivity and volumetric capacity are constant the equations are

$$\frac{\partial T}{\partial \tau} = \frac{\partial^2 T}{\partial \xi^2} \tag{II-10}$$

$$\frac{\partial T}{\partial \tau} = \frac{\partial^2 T}{\partial \xi^2} + \frac{\partial^2 T}{\partial \eta^2} \tag{II-11}$$

To uniquely define the solution temperature T, the initial temperature distribution and the boundary conditions must be specified. For the two-dimensional case the initial distribution would be of the form

$$T = T(\xi, \eta, 0) \qquad \tau = 0 \tag{II-12}$$

For many practical problems the boundary conditions are often of the type

$$h[T_f(\xi_B, \eta_B, \tau) - T(\xi_B, \eta_B, \tau)] = -k \frac{\partial T}{\partial n}(\xi_B, \eta_B, \tau) \qquad \text{(II-13)}$$

$$0 \leq h \leq \infty$$

where ξ_B and η_B are the coordinates of the boundary of the solid, and where n is normal to the boundary directed toward the interior. The heat-transfer coefficient is always positive and may vary from zero (adiabatic condition) to infinity (a specified boundary temperature). For the one-dimensional case the initial condition would be a function only of ξ, and the boundary condition would give equations like II-13 for two values of ξ.

The Laplacian operator* $\nabla \cdot k \nabla$ or $\nabla' \cdot k \nabla'$ itself and the combination of Laplacian operator together with its boundary conditions have several important mathematical properties. The Laplacian operator is a stable operator, or, equivalently, it tends to minimize the energy at a point. When the Laplacian operator and its boundary conditions are considered together, this property has the physical meaning that, for boundary conditions which represent a finite amount of energy, the temperature remains bounded. Mathematically this is expressed as a negative property meaning that, for homogeneous

* Although in this thesis the operators $\nabla \cdot k \nabla$ or $\nabla' \cdot k \nabla'$ are called Laplacian operators, the Laplacian operator is usually defined as ∇^2.

boundary conditions (equation II-13 with T_f of zero), the following volume integral is negative (5).

$$\iiint_V T\nabla \cdot k\nabla T \, dV < 0 \tag{II-14}$$

where T is any smooth function that satisfies a homogeneous boundary condition (equation II-13 with $T_f(\xi_B, \eta_B) = 0$) .

$$k \frac{\partial T}{\partial n}(\xi_B, \eta_B) - hT(\xi_B, \eta_B) = 0 \tag{II-15}$$

This means essentially that the characteristic roots of the operator are negative; and the transient part of the solution is a sum of negative exponentials. This result can be generalized, using theorems found in Carslaw and Jaeger (1), by superposition to nonhomogeneous boundary conditions where T_f is a non-zero function of the boundary coordinates, but constant with time, and by Duhamel's Theorem (1) to cases where T_f is a function both of boundary coordinates and time. The fluid temperature function can be considered to be analogous to a forcing function for a stable linear ordinary differential equation. The second important property is that the Laplacian operator is symmetric or self-adjoint. Physically this means that the thermal conductance in one direction is the same as in the reverse direction. For the operator and the nonhomogeneous boundary condition, equation II-13, this property is stated mathematically in terms of a volume integral.

$$\iiint_V T^{*} \nabla \cdot k \nabla T^{**} dV = \iiint_V T^{**} \nabla \cdot k \nabla T^{*} dV \qquad \text{(II-16)}$$

where T^{*} and T^{**} are functions that fit the homogeneous boundary condition, equation II-15. These are properties that the approximate formulation retains.

(B) CONTINUOUS SOLUTION

The solution for the linear one-dimensional problem (equation II-10) with linear boundary conditions (equation II-13) is written symbolically as:

$$T(\xi,\tau) = T_P(\xi,\tau) - \sum_{j=1}^{\infty} a_j b_j(\xi) e^{-\upsilon_j^2 \tau} \qquad \text{(II-17)}$$

$$0 \le \xi \le 1$$

$$\tau \ge 0$$

(Here and in the following discussion ξ need not be a Cartesian coordinate.) In this equation $T_P(\xi,\tau)$ is a true steady-state solution that is not a function of time if the fluid temperature is constant; or it is a quasi-steady-state solution if the fluid temperature changes with time and, in this case, is related to Duhamel's Theorem. The terms after the summation sign are the transient solution and this part of the solution is related to the problem of zero fluid temperature and an arbitrary initial temperature distribution. In the transient solution the υ_j's are an infinite number of real roots to the characteristic equation which depends upon the boundary equations. Since the υ_j's are real, $-\upsilon_j^2$ is negative, and the exponentials go to zero with increasing time. The $b_j(\xi)$ is the eigenfunction that corresponds to υ_j; these functions are orthogonal over the range of ξ with respect to a weighting function

$$\int_0^1 \sigma(\xi)\, b_j(\xi)\, b_\rho(\xi)\, d\xi = \delta_{j,\rho} \qquad \text{(II-18)}$$

$$j, \rho = 1, 2, \ldots \infty$$

The a_j's are the Fourier coefficients determined with the orthogonality relationship so that the expression fits the initial temperature distribution.

$$a_j = \int_0^1 \sigma(\xi)\, b_j(\xi)\, [T(\xi, 0) - T_P(\xi, 0)]\, d\xi \qquad \text{(II-19)}$$

Although the solution above is shown specifically for the heat-transfer coefficient boundary condition, equation II-13, the above form is also true for other problems, for which the solution may be found by the separation of variables technique, or for which the Laplace transform of the solution has only poles for singularities. Consequently the above form holds for many of the commonly encountered problems. The main problem it does not appear to apply to is where ξ extends to infinity (semi-infinite solid); however, by replacing the summation with an integration, together with other appropriate manipulations, a form very similar to equation II-17 is found which does apply.

For a two-dimensional problem of equation II-11 the above arguments can be used to show that the solution is in the form:

$$T(\xi, \eta, \tau) = T_P(\xi, \eta, \tau) - \sum_{j=1}^{\infty} \sum_{j=1}^{\infty} a_{ij}\, b_{ij}(\xi, \eta) e^{-\nu_{ij}^2 \tau} \qquad \text{(II-20)}$$

Either to compare the continuous solution with an approximate solution, or to calculate a continuous solution for a case where it is known, a vector-matrix form for the solution is sometimes preferred. Assuming that the temperature at S points within the solid is desired, an S-dimensional "continuous temperature vector" can be defined with its ith element found by evaluating equation II-17 or II-20 at the coordinate of the ith point. Also, because the infinite sum is over damped exponentials, to obtain a certain degree of accuracy for any time greater than zero. only a finite number of the largest exponential terms need be considered. If this number is J the continuous solution vector for a one-dimensional problem can be written as

$$T(\tau) = T_P(\tau) - B\, e^{-\upsilon^2 \tau}\, a \qquad \text{(II-21)}$$

where $T_P(\tau)$ is an S-dimensional vector of particular solutions; B is an S by J matrix whose j^{th} column is $b_j(\xi)$ evaluated for the appropriate ξ's; $e^{-\upsilon^2 \tau}$ is a diagonal matrix with the jth diagonal element $e^{-\upsilon_j^2 \tau}$; and a is a J-dimensional vector of Fourier coefficients a_j. The continuous solution equation II-20 for the two-dimensional problem also can be put into exactly the same vector-matrix form as above, equation II-21.

(C) APPROXIMATE FORMULATION

The difference equations that are used to approximate the partial differential equation are usually derived by either of two methods: first, directly from physical considerations, and second, from a Taylor series expansion of the temperature function. The series method is also very useful in the analysis of errors introduced by replacing the differential equation by a difference approximation, and is used in Chapter V. Here the approximations are derived directly from an energy balance for a two-dimensional problem using an asymmetric distribution of points following MacNeal (8) and Longwell (9). Although this system derived for the irregular distribution of temperature points might not be entirely satisfactory, as discussed in Chapter V, it results in a useful general matrix-vector formulation that can be considered to be independent of some of the rules for finding the actual matrix elements, and which reduces to the usual approximations for a regular arrangement of nodes. Further, although only the two-dimensional case is derived, the concepts of one- and three-dimensional cases are similar; however, in practice, the three-dimensional case is sometimes difficult to visualize.

Following MacNeal's (8) rules for two-dimensional flow, points are selected within the cross-section. Lines are drawn from each point to surrounding points in such a manner that the lines do not intersect. It must be possible to connect the points with physically

realizable resistors, which means that the internal angles must not be obtuse in triangles. If necessary, the points should be rearranged to meet these restrictions. Next, perpendicular bisectors are constructed for each line between the points; if the above rules are followed these bisectors will meet at a point, and the network will appear as in Figure III-1. A temperature node is defined for convenience as a point mass having a heat capacity and thermal conductances to surrounding nodes based on the physical and geometrical properties of the surrounding area.

Considering the ith node, the thermal conductance between this and an adjacent node j is given by:

$$\frac{\dot{q}_{ij}}{t_i - t_j} = \frac{k\, r_{ij}}{\ell_{ij}} \qquad i = 1, 2, \ldots, R \qquad \text{(II-22)}$$

The volume or area of unit thickness for the two-dimensional case for the ith node is

$$\text{Area}_i = \tfrac{1}{4} \sum_{\substack{j=1 \\ j \neq i}}^{R} r_{ij}\, \ell_{ij} \qquad i = 1, 2, \ldots, R \qquad \text{(II-23)}$$

where r_{ij} is zero if there is no connection between i and j; r_{ii} and ℓ_{ii} have no meaning; and the summation is over all nodes surrounding the ith node. Thus, setting the accumulation of energy at the node equal to the net flux in, we may write an ordinary differential equation for each node,

$$\frac{\sigma C_p}{4}\left(\sum_{\substack{j=1\\ j\neq i}}^{R} r_{ij}\,\ell_{ij}\right)\frac{dt_i}{d\theta} = k\sum_{\substack{j=1\\ j\neq i}}^{R}\frac{r_{ij}}{\ell_{ij}}(t_j - t_i) \qquad \text{(II-24)}$$

$$i = 1, 2, \ldots, R$$

Consequently, the partial differential equation has been replaced with a system of ordinary differential equations which are suitable for solution on an analog computer. Following Longwell's (9) derivation, a dimensionless heat capacity A_i and a dimensionless conductance are introduced.

$$A_i = \frac{\sigma C_p (\text{Area})_i}{(\sigma C_p)_0 L^2} = \frac{\sigma C_p}{4(\sigma C_p)_0}\sum_{j=1}^{R}\frac{r_{ij}\,\ell_{ij}}{L^2} \qquad \text{(II-25)}$$

$$i = 1\ 2\ .\ .\ ., R$$

$$y_{ij} = \frac{k}{k_0}\frac{r_{ij}}{\ell_{ij}} \qquad i,j = 1, 2, \ldots, R \qquad \text{(II-26)}$$

Substitution gives

$$\sum_{\substack{j=1\\ j\neq i}}^{R} y_{ij}(t_j - t_i) = A_i\frac{dT_i}{d\tau} \qquad i = 1, 2, \ldots, R \qquad \text{(II-27)}$$

Now departing from both MacNeal's (8) and Longwell's (9) derivations, the system of equations may be summarized in matrix form:

$$\left[\begin{array}{cccc|ccc} A_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & A_S & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & A_{S+1} & \cdots & 0 \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 0 & 0 & & A_R \end{array}\right] \frac{d}{d\tau} \left(\begin{array}{c} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_S \\ \hline t_{S+1} \\ \cdot \\ \cdot \\ \cdot \\ t_R \end{array}\right) - \qquad \text{(II-28)}$$

$$\left[\begin{array}{cccc|ccc} -\sum_{\substack{j=2 \\ j\neq 1}}^{R} y_{1j} & y_{12} & \cdots & y_{1S} & y_{1S+1} & \cdots & y_{1R} \\ y_{12} & -\sum_{\substack{j=1 \\ j\neq 2}}^{R} y_{2j} & \cdots & y_{2S} & y_{2S+1} & \cdots & y_{2S} \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ y_{1S} & y_{2S} & \cdots & -\sum_{\substack{j=1 \\ j\neq S}}^{R} y_{Sj} & y_{SS+1} & \cdots & y_{SR} \\ \hline y_{1S+1} & y_{2S+1} & \cdots & y_{SS+1} & -\sum_{\substack{j=1 \\ j\neq S+1}}^{R} y_{S+1j} & \cdots & y_{S+1R} \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ y_{1R} & y_{2R} & \cdots & y_{SR} & y_{S+1R} & \cdots & -\sum_{\substack{j=1 \\ j\neq R}}^{R} y_{Rj} \end{array}\right] \left(\begin{array}{c} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_S \\ \hline t_{S+1} \\ \cdot \\ \cdot \\ \cdot \\ t_R \end{array}\right)$$

Since some of the nodes represent boundary nodes, their temperatures are known as a function of time. Assuming that the boundary nodes are (S+1) to R, we may rewrite the matrix without the equations for nodes (S+1) to R. Carrying this out, and simplifying the equations by partitioning the matrices and temperature vector, gives

$$\begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & A_S \end{bmatrix} \frac{d}{d\tau} \begin{pmatrix} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_S \end{pmatrix} = \begin{bmatrix} -\sum_{\substack{j=2 \\ j\neq 1}}^{R} y_{1j} & y_{12} & \cdots & y_{1S} \\ y_{12} & -\sum_{\substack{j=1 \\ j\neq 2}}^{R} y_{2j} & \cdots & y_{2S} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ y_{1S} & y_{2S} & \cdots & -\sum_{\substack{j=1 \\ j\neq S}}^{R} y_{Sj} \end{bmatrix} \begin{pmatrix} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_S \end{pmatrix} \tag{II-29}$$

$$+ \begin{bmatrix} y_{1\,S+1} & \cdots & y_{1R} \\ y_{2\,S+1} & \cdots & y_{2R} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ y_{R\,S+1} & \cdots & y_{RR} \end{bmatrix} \begin{pmatrix} t_{S+1} \\ \cdot \\ \cdot \\ \cdot \\ t_R \end{pmatrix}$$

which is written equivalently in vector-matrix form as

$$A \frac{d}{d\tau} t = Yt + Y_B t_B \tag{II-30}$$

where the two equations serve to define the several matrices and vectors. The A matrix, because each diagonal element is a positive

non-zero element, is a positive definite matrix. The Y matrix is a symmetric matrix that represents a network of thermal resistors; thus, it is a negative definite matrix.

Although the above equation was derived specifically for points located on the boundary, whose temperature is known as a function of time, generalizations to other types of linear boundary conditions can be made. This can be done, for example, in the case of the solid surrounded by a fluid with a heat-transfer coefficient h, or for a specified flux per unit area, merely by modifying the energy balances appropriately for the effect on nodes on or close to the boundary. If the boundary condition equations involve a heat-transfer coefficient, the Y and Y_R matrices contain the heat-transfer coefficient. This condition is thoroughly discussed later. If the flux is specified, then there is no boundary conductance, the Y matrix is the same as for zero h or an adiabatic condition, and $Y_B\ t_B$ is replaced directly with the known flux in to each node per unit time. Also, for the equation in the above form, it is easy to see that the boundary temperature vector (or flux vector) is the nonhomogeneous part of the equation and the boundary temperature vector t_B can be considered a forcing function or input disturbance.

Now solving equation II-30 for $\frac{d}{d\tau} t$ to put it in a form similar to the partial differential equation, there is obtained after defining two new matrices:

$$\frac{d}{d\tau} t = \frac{Y}{A} t + \frac{Y_B}{A} t_B \tag{II-31}$$

where $$\frac{Y}{A} = A^{-1} Y \text{ and } \frac{Y_B}{A} = A^{-1} Y_B \tag{II-32}$$

At this point the Laplacian operator $(1/K_0 C_p \sigma) \nabla' k \nabla'$ and its boundary conditions which apply to all points within and on the boundary of the solid have been replaced with matrix operators Y/A and Y/A_B, which apply at a finite number of points within and on the solid boundary. For the ith point, we have

$$\frac{1}{K_0 \sigma C_p} \nabla' k \nabla' T_i \approx \frac{1}{A_i} \sum_{j=1}^{R} y_{ij}(t_j - t_i) \tag{II-33}$$

$$i = 1, \ldots, S$$

The matrix Y/A does retain both the stable or negative property and the symmetric property of the Laplacian operator. First the Y/A matrix as defined by the above rules can be shown to be a negative definite or negative semi-definite matrix and so has all negative eigenvalues, and a finite sum analogous to the volume integral is always negative. The transient solution to the system of ordinary differential equations is then a sum of damped exponentials giving a stable solution. As the Y/A matrix is not a symmetric matrix (except for the important case when all the A_i's are equal), it might appear that the symmetric property is lost; however, the following discussion shows that the matrix operator is symmetric in the same sense as the

Laplacian is symmetric. In the volume integral that defines this property, the value of $\nabla' . k \nabla' T$ is weighted by the differential volume for which it applies; consequently, an equivalent weighting is introduced when the Laplacian and the differential volume element, dV, are expressed in cylindrical or other orthogonal curvilinear coordinates. Thus, in the finite sum over the volume that is the discrete analog of the volume integral, the A_i's are used as the weighting for the ith point. This sum is then

$$\sum_{i=1}^{S} A_i t_i^* \left[\frac{1}{A_i} \sum_{j=1}^{S} y_{ij} (t_j^{**} - t_i^{**}) \right] = \sum_{i=1}^{S} t_i^* \sum_{j=1}^{S} y_{ij} (t_j^{**} - t_i^{**}) \quad \text{(II-34)}$$

where t^* and t^{**} correspond to T^* and T^{**} and the summation is on all the points within and on the boundary and thus includes the boundary conditions. Because the Y matrix is a symmetric matrix, this finite sum has the symmetric relationship:

$$\sum_{i=1}^{S} t_i^* \sum_{j=1}^{S} y_{ij} (t_j^{**} - t_i^{**}) = \sum_{i=1}^{S} t_i^{**} \sum_{j=1}^{S} y_{ij} (t_j^{*} - t_i^{*}) \quad \text{(II-35)}$$

which is analogous to equation II-16. However, the use of A_i as a weighting factor here suggests that the best location for the points might be such that the weighting A_i is related to the weighting introduced by dV for the continuous case. The symmetric property of Y/A also can be shown in an indirect way by finding a symmetric matrix W that has the same negative eigenvalues of Y/A. This matrix, as found by expanding on arguments in Perlis (11), Theorem 9-13, is:

$$W = D\ Y\ D \tag{II-36}$$

where D is a diagonal matrix with $1/\sqrt{A_i}$ in the ith diagonal position and the elements of W are:

$$w_{ii} = \mu_{ii} = -\frac{1}{A_i} \sum_{\substack{j=1 \\ i \neq j}}^{R} y_{ij} = \text{same diagonal element as } \frac{Y}{A}$$

$$w_{ij} = w_{ji} = \frac{y_{ij}}{\sqrt{A_i A_j}} = \mu_{ij} \frac{\sqrt{A_i}}{\sqrt{A_j}}$$

$$\mu_{ij} = \text{element of } \frac{Y}{A}$$

The algebraic approximation in equations II-31 and II-33 to the Laplacian could be used in other partial differential equations where the Laplacian occurs. For example, an approximate solution to Laplace's equation could be obtained by letting $\frac{d}{d\tau}(t) = 0$ in equation II-30 and solving the resulting system of algebraic equations.

The resulting system of linear ordinary differential equations are of an initial-value nature and may be solved analytically or on an analog computer. Also, any of the commonly used numerical integration methods could be used, such as one of Runge-Kutta's methods. The fourth-order Runge-Kutta method (12) in particular would give a very close approximation to the solution of the linear ordinary differential equations. However, this exact solution of the system of ordinary differential equations does not necessarily mean that a better approximation to the partial differential solution is obtained unless an extremely

fine space mesh is used, and thus the simple Euler's method or an implicit variation is usually used with adequate accuracy. This is true because, in the approximate solution of this partial differential equation of diffusion, the time-discretization error often compensates the space error and more accurate results are obtained. To formulate this procedure the temperature-time derivative is approximated by

$$\frac{d}{d\tau} t = \frac{1}{\Delta\tau}(t_{n+1} - t_n) \tag{II-37}$$

However, when this is substituted in equation II-31, a question arises whether the temperature vectors on the right side of equation II-31 should be evaluated at $n\Delta\tau$ or $(n+1)\Delta\tau$ or at an intermediate time. Assuming that a linear combination of vectors t_{n+1} and t_n is to be used, we may introduce weighting factors γ and $(1-\gamma)$. These factors are usually taken between zero and one, and are the weighting given to the temperature vector at $(n+1)\Delta\tau$ and $n\Delta\tau$, respectively. Thus, the temperature vector on the right hand side of equation II-31 is replaced with

$$t_{\tau_1} = (1-\gamma)t_n + \gamma t_{n+1} \tag{II-38}$$

$$n\Delta\tau \leq \tau_1 \leq (n+1)\Delta\tau$$

Introducing γ in our derived expressions gives not only the explicit forward-difference method when γ is zero, but also gives the implicit

type for γ not equal to zero; specifically, it gives central and backward differences when γ is $\frac{1}{2}$ or 1. More complicated methods of determining the vector $t_{\tau 1}$ for substitution on the right side of the difference equations have been suggested to improve the accuracy (6, 7) or the stability (5).

Continuing with the derivation, substituting $t_{\tau 1}$ for the temperature vector on the right side of equation II-31 and solving for t_{n+1}, we obtain

$$t_{n+1} = \left[I-\gamma\Delta\tau\frac{Y}{A}\right]^{-1}\left\{\Delta\tau\frac{Y_B}{A}(\gamma t_{B\,n+1}+[1-\gamma]t_{Bn})+\left[I+(1-\gamma)\Delta\tau\frac{Y}{A}\right]t_n\right\} \quad \text{(II-39)}$$

This equation is suitable for stepping out the approximate solution, although when γ is not zero a matrix inversion or equivalent operation is required. The solution of the equations for the implicit methods should not be considered a serious disadvantage, because the matrices usually contain many zeros, and are of a special form for which simplifications are possible (6). When $\gamma = 0$, the equation reduces to the explicit forward difference formulation used by Longwell (9) which requires no matrix inversion.

$$t_{n+1} = -\Delta\tau\frac{Y_B}{A}t_{B\,n}+\left[I+\Delta\tau\frac{Y}{A}\right]t_n \quad \text{(II-40)}$$

Here it should be strongly emphasized that these equations are those actually used in stepping out the numerical solution for successive time increments, starting with the known initial condition (t_0), and knowing how the boundary temperature changes with time.

There are several advantages to formulating the difference problem in this vector-matrix notation. The first is that this is a very general form that applies to one-, two-, or three-dimensional problems for an irregular, as well as a regular, distribution of points. Even though a specific set of rules, MacNeal's (8),was used in the derivation, using another system of rules would only change the elements in the Y/A matrix. As a matter of fact, Chapter V is a study of the accuracy of solutions using this matrix for an asymmetric distribution of nodes. Further, if a regular distribution of points is used, this formulation reduces to those represented by a partial difference equation. The second advantage is that for carrying out the calculations on a digital computer, matrix inversion, multiplication, and addition subroutines are generally available; consequently, the amount of programming necessary is greatly reduced. The third advantage is that a study of the vector-matrix equation itself gives a great deal of information concerning the usefulness, stability, and accuracy of the approximate method.

As mentioned above, for a regular location of the points the partial difference equation can be solved. This solution is found for the one-dimensional problem using z-transforms in Chapter IV. However, the formulation of the one-dimensional problem in matrix form is an example of the simplicity of this approach. In this case ℓ_{ij} becomes Δx for equally spaced points and r_{ij} is 1 for all i and j.

If the temperature vector is ordered either in increasing or decreasing subscripts m, the Y/A matrix is

$$\frac{Y}{A} = \frac{-1}{(\Delta\xi)^2} \begin{bmatrix} - & - & - & - & & - & - & - \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ - & - & - & - & & - & - & - \end{bmatrix} \qquad \text{(II-41)}$$

The entries of the top and bottom rows are not specified since these entries are a direct result of the type of boundary conditions and the method used for approximation; these corresponding equations are derived from energy balances in Chapter IV, section B. The system of difference equations represented by the vector-matrix equation II-39 with Y/A matrix from equation II-41 can also be represented by the following partial difference equation, plus the two boundary condition equations.

$$-r\gamma t_{m-1,n+1} + (1+2r\gamma)t_{m,n+1} - r\gamma t_{m+1,n+1} = r(1-\gamma)t_{m-1,n} + (1-2r\{1-\gamma\})t_{m,n} + r(1-\gamma)t_{m+1,n} \qquad \text{(II-42)}$$

$m \neq$ subscripts for boundary points

$n = 0, 1, \ldots, \infty$

Usually the computational procedure is constructed from this partial difference equation. Note that only when the points are located in some regular manner is it possible to have a partial difference equation, and that generally, for any asymmetric arrangement, the elements of the Y/A matrix cannot be expressed by a simple recurrence formula.

This constitutes a general formulation for the numerical solution of the diffusion equation, which applies both to a regular and an irregular network of nodes. This was accomplished by treating the space differencing and the time differencing as separate problems. In addition to the diffusion equation a difference scheme for steady-state equations is also implied and analog approximations are also shown.

(D) SOLUTION OF APPROXIMATE FORMULATIONS

The main advantage of using difference methods for approximating the solution of partial differential equations is that the numerical results can be found without finding an analytic solution for the approximation. Thus, as previously pointed out, the system of algebraic equations II-39 or II-40 can be solved for successive time increments, stepping out an approximate solution, or the solution of a system of ordinary differential equations can be estimated using an electric analog. However, as the partial differential equation is actually just the special case of an approximate method that is consistent with the partial differential equation where the time and space increments are zero, and n and S are infinite, the approximate methods also have an analytic solution. In order to study both the stability and accuracy of an approximate method, a comparison between the form of the analytic solution for the approximate method and the form of the analytic solution of the partial differential equation is extremely useful. In the following discussion, the form of the solution for the system of difference equations is shown; analogies to the solution of the partial differential equation are indicated; and important properties and characteristics of the difference system are discussed and summarized for future reference. These results are not derived here, as they involve fairly long matrix arguments. Chapter III uses these solutions and

matrix properties to derive a rigorous but simple stability criterion. Chapter IV uses some of these results together with z-transforms to find analytic expressions for solutions using several approximate methods as applied to specific problems. Again, it should be emphasized that, although analytic solutions to the system of difference equations can be found, in order to study their behavior and the accuracy of the approximation, the way to find the numerical approximate solution is by using equation II-39.

The solution for the linear system of ordinary difference equations II-37 can be found by classical methods to be the sum of a particular and complementary solution. The solution vector is:

$$t_n = t_{P\,n} - C\,Q^n\,g \qquad \text{(II-43)}$$

where the ith component of the vector is:

$$t_{i,n} = t_{P\,i,n} - \sum_{j=1}^{S} g_j\,c_{ij}\,q_j^n \qquad i = 1,\ 2,\ \ldots,\ S \qquad \text{(II-44)}$$

The particular solution $t_{P\,n}$ is an S-dimensional vector; it is found by the method of undetermined parameters or other methods, and it depends primarily on how the fluid temperatures in t_B, or forcing functions, change with space and time. This solution represents the steady-state solution if it does not change with time, or a quasi-steady-state if it is a function of time. The complementary solution $-C\,Q^n g$ is also an S-dimensional vector. It is the solution to the homogeneous system

of difference equations derived from equation II-39, which is:

$$\left[I - \gamma\Delta\tau\frac{Y}{A}\right]t_{n+1} - \left[I + (1-\gamma)\Delta\tau\frac{Y}{A}\right]t_n = 0 \qquad \text{(II-45)}$$

Both C and Q are S by S matrices and g is an S-dimensional vector. The matrix Q is a diagonal matrix with q_j in the jth diagonal position. These S q_j's are the S real roots of the characteristic equation which is an Sth order polynomial.

$$\det\left[(q-1)I - (q\gamma + 1 - \gamma)\Delta\tau\frac{Y}{A}\right] = 0 \qquad \text{(II-46)}$$

The S q_j's are actually the eigenvalues of the matrix $[I-\gamma\Delta\tau(Y/A)]^{-1}[I+(1-\gamma)\Delta\tau(Y/A)]$. Since the matrix Q^n is the quantity in the approximate solution which corresponds to the damped exponential matrix $e^{-\upsilon^2\tau}$ in the continuous solution where the υ_j's are real, it can be seen that each q_j must be real, and between 0 and 1, to give the same type of behavior (or at least between -1 and +1 to be bounded); and the matrix arguments only show that the q_j's are real. Thus, to obtain a solution that is to be an approximation to the continuous solution we must select Y/A, $\Delta\tau$, and γ so that the q_j's meet this restriction; this is the topic of stability and is thoroughly discussed in Chapter III. However, for purposes of discussion, it will be assumed here that the q_j are all less than one in absolute value.

The jth column of matrix C is the eigenvector of the matrix $[I-\gamma\Delta\tau(Y/A)]^{-1}[I+(1-\gamma)\Delta\tau(Y/A)]$ which corresponds to the jth eigenvalue

q_j; it is defined by:

$$\left[(q_j-1)I - (q_j\gamma+1-\gamma)\Delta\tau \frac{Y}{A}\right]c_j = 0 \qquad j = 1, 2, \ldots, S \qquad \text{(II-47)}$$

The eigenvectors c_j are real and the set of S eigenvectors are linearly independent; the inverse of the eigenvector matrix C can be found explicitly, by using and expanding on matrix arguments in Perlis (11), Theorem 9-13, and using the relationship between Y/A and the symmetric W matrix, to be

$$C^{-1} = C'A \qquad \text{(II-48)}$$

where C' is the transpose of C. From this, the relationship analogous to the orthogonality relationship for the continuous solution is

$$\sum_{i=1}^{S} A_i c_{ij} c_{i\rho} = \delta_{j\rho} \qquad \rho, j = 1, 2, \ldots, S \qquad \text{(II-49)}$$

The vector g is found using the above inverse relationships so that the solution fits the initial condition:

$$g = -C'A (t_0 - t_{P\,0}) \qquad \text{(II-50)}$$

Its components are

$$g_j = -\sum_{i=1}^{S} A_i c_{ij} (t_{i,0} - t_{P\,i,0}) \qquad j = 1, \ldots, S \qquad \text{(II-51)}$$

The solution for the linear system of ordinary differential equations can be found either as the limiting case of the above solution as the time increment $\Delta\tau$ goes to zero, or directly from equation II-31 by classical methods. It is, in vector form:

$$t(\tau) = t_P(\tau) - C\, e^{\Lambda\tau}\, g \qquad \text{(II-52)}$$

where a component of $t(\tau)$ is

$$t_i(\tau) = t_{P\,i}(\tau) - \sum_{j=1}^{S} g_j\, c_{ij}\, e^{\lambda_j \tau} \qquad i = 1, 2, \ldots, S \qquad \text{(II-53)}$$

In this case the λ_j are the S negative real eigenvalues of the Y/A or W matrices, and are the S roots of the Sth order determinantal characteristic polynomial, which can be written in any of the following ways:

$$\det[\lambda I - \frac{Y}{A}] = \det[\lambda I - W] = [\det D]^2\, (\det[\lambda A - Y]) = 0 \qquad \text{(II-54)}$$

(each of the determinantal equalities are actually identical for all values of λ) where the roots are

$$\lambda_j < 0 \qquad j = 1, 2, \ldots, S \qquad \text{(II-55)}$$

In this case $(C\, e^{\Lambda\tau}\, g)$ does indeed represent a transient solution as the λ_j's are negative. Comparison of equations II-54 and II-46 gives us an important relationship between the eigenvalues of the Y/A matrix, λ_j, and of the difference system:

$$\lambda_j = \frac{q_j - 1}{(q_j + 1 - \gamma)\Delta\tau} \qquad j = 1, 2, \ldots, S \qquad \text{(II-56)}$$

or, solving for q_j:

$$q_j = \frac{1 + \lambda_j (1-\gamma)\Delta\tau}{1 - \lambda_j \gamma \Delta\tau} \qquad \text{(II-57)}$$

Also from equation II-47 and the relation between λ and q, equation II-56, and the fact that the Y/A matrix is not a function of the time increment, it is seen that the eigenvectors c_j can be taken as the eigenvectors of the Y/A matrix and are then the same for both the analog and difference approximations. Thus, if the particular solutions evaluated at time zero are the same for both the difference and analog approximations, as they are when the fluid temperatures are constant with time, the initial vectors are also identical. The particular solution represents the condition at infinite time for the analog solution, and it depends on how the boundary temperature changes with time.

(E) CONCLUSIONS AND SUMMARY

The transient temperature distribution in a solid, with a known initial temperature distribution and known boundary conditions, has been shown to be the continuous solution of a partial differential equation, the diffusion equation. This equation is derived by making an energy balance on a differential volume element, for a differential element of time, and it applies to the infinite points within the solid boundaries.

A difference approximation to the partial differential equation also has been presented. In this approximation a finite number of points is located within and on the solid boundaries, but not necessarily in a regular fashion. An energy balance based on a finite volume and a finite time increment for each point then gives a system of algebraic equations. By solving the algebraic equations for successive time increments, a numerical approximation of the transient temperature distribution can be stepped out, starting with the known initial conditions and using known forcing functions as a function of time.

Even though the points are not located regularly, the system of difference equations possesses several important properties that are analogous to properties of the partial differential equation. The matrix operator Y/A which replaces the Laplacian operator is negative definite, corresponding to the negative or stable character of the Laplacian. Although this matrix is not symmetric, it has a symmetric property

directly analogous to that of the Laplacian, and is similar to a symmetric matrix W. For the difference system, the negative and symmetric property is defined in terms of summations, which can be considered to be discretizations of the volume integrals which define the properties for the partial differential equation.

In addition to these analogies in the derivations and in the equations themselves, the system of difference equations has an analytic solution with the same form as the solution of the partial differential equation. The main advantage of the system of difference equations is that its solution can be stepped out numerically merely using algebraic techniques, but the analytic solution is not found in practice. However, a study of the form of the analytic approximate solution, and a comparison of it with the analytic continuous solution, are extremely important and useful in determining the stability criteria, and in analyzing the accuracy of an approximate method. As much of this thesis is concerned with an analysis of these solutions, it is appropriate to summarize and review here not only the analogies between the several parts of the solutions, but also to state which parts of both solutions are affected by changing the problem parameters, boundary conditions, and initial conditions, and which parts of the approximate solution are affected by changing the differencing parameters.

Both the continuous and difference solutions are made up of a transient solution and a particular solution (assuming stability of the

difference system). The particular solution governs the solution at long or infinite time and depends upon how the forcing functions change with time and the space variables. In general, if the forcing function is a constant, both particular solutions are constant; if the forcing function varies only with space variables, both particular solutions are functions of space variables; if the forcing functions are a function of time, the particular solutions are a function of space and time. In addition, the particular solution for the difference problem is affected by the discretization used for the variables upon which its particular solution depends. That is, for example, when the forcing function varies only with space variables, the approximate particular solution is also a function of the number and location of the points which fix the Y and Y_B matrices.

The transient solution of the continuous solution is an infinite sum of terms of the form $a_j b_j e^{-\nu_j 2_\tau}$ and of the difference solution is a finite sum of terms of the form $g_j c_j q_j^n$. In these solutions the exponential term $e^{-\nu_j 2_\tau}$ and the exponential form q_j^n correspond and determine the rate of change of the jth term of the solution. Both the ν_j and the q_j are determined by a characteristic equation which depends upon the geometric and thermal properties of the solid and the boundary equations which are physical system variables. The eigenfunction b_j and eigenvector c_j determine the weighting of the jth exponential (or exponential form) at each point where the solution applies. These

quantities are dependent upon the physical system variables. Although the eigenvectors c_j are not orthogonal vectors, according to the strict mathematical definition, a finite summation relationship exists that is directly analogous to the orthogonal integration of the eigenfunctions. The difference initial vector components g_j are found by fitting the initial condition with this finite summation relationship in the same way that the Fourier coefficients a_j are found with the orthogonal integration. These quantities then depend upon the initial temperature distribution and the nature of the particular solution, in addition to the physical system variables.

The difference transient solution also depends upon the space discretization parameters, the number and location of the points, and the time discretization parameters, the time increment and weighting γ. For the formulation in which the points are located and the equations calculated by MacNeal's rules, the Y/A and Y_B/A matrices are then fixed by the number and location of the points in addition to the properties of the physical system. Consequently, the eigenvectors c_j and eigenvalues λ_j depend upon the number and location of the points, but are independent of the time differencing parameters and of the forcing function or initial condition. The exponential forms q_j^n depend upon both the time differencing parameters $\Delta\tau$ and γ, in addition to the space differencing, as it affects λ_j. For forcing functions that are constant, or a function only of space, the only differencing parameter that affects

the g_j is the number and location of points. If the particular solution evaluated at zero is a function of the time differencing, then the g_j's also depend upon the time differencing.

Because of the additional dependence of the approximate solution on the differencing parameters, nothing can be inferred about the accuracy of the approximate solution. As mentioned previously, we are not even assured of the stability of the difference system. A useful stability criterion is developed in the next chapter, and accuracy is discussed in the final two chapters.

CHAPTER III

GENERAL STABILITY CRITERIA

Finite difference approximations to the diffusion equation can give anomalous results, as just mentioned. For example, depending upon the values of r and γ selected in equation II-42, undamped oscillations, damped oscillations, or smooth curves with time, can result for cases when the exact continuous solution is known to be an infinite sum of decaying exponentials. Although this problem has been the subject of many studies, almost all of them consider only cases of uniform mesh spacing where a partial difference equation exists and can be studied, and usually their results only apply to problems where the boundary temperatures are specified functions of time (infinite heat-transfer coefficient). For the asymmetric network or other non-uniform mesh spacing the partial difference equation does not exist, and this is one of the reasons that the matrix formulation and solutions just shown constitute a more general method of studying these difference approximations. Although the stability definitions and criteria based on the matrix solutions are almost trivial, in order to put them in proper perspective, a brief consideration of earlier important work is presented. The definitions used in this study are then presented and the criteria derived. The practical use of these criteria is discussed,

followed by a numerical example using an asymmetric network to check the criteria and also to show that the asymmetric distribution of temperature points leads to reasonable and fairly accurate results.

(A) STABILITY--REVIEW AND DEFINITIONS

In describing the numerical approximation of partial differential equations, the words stability and convergence are often used. Neither of these terms has a generally accepted precise definition, although certain concepts usually are associated with each. An unstable solution is usually taken to mean one where the numerical results go to infinity with time or the number of time increments, i.e.,

$$\lim_{n\to\infty} \| t_n \| \to \infty \qquad \text{(III-1)}$$

and is attributed by some authors to round-off error, or the computer's inability to carry an infinite number of decimals. A convergent solution is usually defined as one where the difference solution goes to the continuous solution as both the time and space increments are refined to zero; and thus convergence is often associated with the discretization error. Most of the recent studies have shown that round-off error is usually not significant and that the solution would oscillate with increasing amplitude even if an infinite number of decimals were carried along. If a q_j is greater than one in absolute value, the solution becomes infinite, even when infinite decimals are carried.

One of the most general discussions of stability and convergence is that of Richtmyer (3). He has precisely defined stability, convergence, and consistency on the basis of making a series of approximate

calculations. For each calculation in the series, the approximation is carried from zero to a fixed time. And, for each successive approximate calculation in the series, the time and space increments are reduced so that a sequence of approximate calculations is defined; the last limiting approximate calculation uses a zero time increment and zero space increment. A consistent approximation* is one where, in the limit of this sequence, the difference between the continuous operator and the approximate operator operating on the continuous function goes to zero. A stable approximation requires, for all approximate calculations in the sequence including the limiting final calculation at zero time increment, that an upper limit exists which is never exceeded in the calculations to the constant time. Note that, in the limit when the time increment is zero, n must be infinite. A convergent approximation is one which, in the sequence of approximate calculations, equals the continuous solution exactly in the limit of the sequence. The basic result is that if the initial-value problem is linear and fulfills certain other very general conditions, and further, if the approximation is consistent and stable, it is then convergent, and arbitrarily accurate approximations can be obtained by using very fine space-time increments. This consequence applies not only to linear partial differential equations of an initial-value nature but also to linear integral equations.

---

* These definitions are stated in a very precise mathematical form in terms of norms of Banach spaces.

The above result can be applied directly to approximations of the linear diffusion equation based on the regular location of temperature points. In these cases space and time increments can be reduced simultaneously and a more accurate approximation is obtained if the solution remains bounded. For a rectangular network, this could be done by halving the space increment for each new calculation. However, for the more general irregular or asymmetric network, the refinement of the space network has an ambiguous meaning. It is theoretically possible to have a sequence of geometrically similar nodes; i.e., each successive node in the sequence has the distance to its neighbors reduced in proportion while the angles remain constant; but it is usually not possible to define a sequence of networks based on refining a general asymmetric node in that way. Further, in practice, we are not usually concerned with such a sequence of calculations, each using smaller increments, but points have been located based on the above rules to fit the geometry and thermal properties of the solid, taking into account the purpose of the calculation and the capacity of the computing equipment or the cost of solution. It is desired then to select a time increment $\Delta\tau$ and a weighting factor γ that give useful results. Two definitions that are found to be convenient from this point of view, and which are based on the form of the solution of the difference system (equation II-43) are given below for a stable solution and a non-oscillatory solution. The definitions are used throughout the remainder of this study.

Stable Approximation: A stable approximation is defined as one where, in the numerical stepping-out of the solution for a fixed network of points, no calculated temperature goes to infinity with time for all problems where the forcing functions represent a finite amount of energy into or out from the solid from zero to infinite time. Examples of such forcing functions included in this category are step functions and equal sinusoids. This behavior must be independent of the initial temperature distribution. The above definition is equivalent to the usual stability definition used for linear feedback control systems (13, 14).

Non-Oscillatory Approximation: A non-oscillatory approximation gives a stable solution that contains no oscillations except those caused by the oscillations of the boundary forcing functions.

Both these definitions are based on the behavior of the approximate solution and do not insure accuracy. However, the continuous solution of the diffusion equation with no sources (equation II-4) gives both a stable solution and a non-oscillatory solution, according to the above definition, so that in order that an approximation be accurate, it must be stable, if not both stable and non-oscillatory. Further, the stability definition is like Richtmyer's definition in that they both are concerned with the boundedness of the calculated temperatures. In Chapters IV and V the accuracy is studied in much more detail, from a practical standpoint, for cases where the time and space increments are not zero.

(B) NECESSARY AND SUFFICIENT CONDITIONS FOR STABILITY AND NON-OSCILLATORY BEHAVIOR

These definitions together with the solution of the system of difference and ordinary differential equations II-39 and II-31 facilitate the derivation of the necessary and sufficient conditions for stable or non-oscillatory solutions. First, since the Y/A matrix is a negative definite matrix, its eigenvalues are all real and negative and the analog system of differential equations is, therefore, always stable and non-oscillatory. Also, if modifications are made in the rules used to compute Y/A, it is necessary and sufficient to insure stability of the analog that the negative definite property be retained.

For the complete difference approximation, the necessary and sufficient condition for stability is that all the difference eigenvalues q_j's be less than one in absolute value, or, more precisely,

$$-1 < q_j \leq 1 \qquad j = 1, 2, \ldots, S \qquad \text{(III-2)}$$

The reason that a q_j of -1 is not allowed is that, if a q_j is -1 and if the fluid temperature has equal oscillations with a frequency of $1/2\Delta\tau$ or a multiple of $1/2\Delta\tau$, a term of the form $n(-1)^n$ occurs in the particular solution; when a term of this form occurs in the solution the q_j of -1 is said to be "excited." Thus the solution would give undamped oscillations for a forcing function representing a finite amount of energy. A q_j of +1 must be allowed for cases where the boundary is adiabatic

(Y_B is zero), as a q_j of 1 is necessary to represent a constant non-zero steady-state solution. Also, if a q_j falls outside these limits, carrying an infinite number of decimals cannot make the solution stable. However, since it can be shown that any error introduced into the calculation is propagated in such a manner that it satisfies a system of difference equations similar to equation II-39, any round-off error introduced into a calculation where a q_j less than -1 exists "increases" or "aggravates" the instability, or, if the solution were such that the g_j which corresponds to this q_j were zero, this round-off error would "activate" this eigenvalue (make g_j non-zero) and the solution would become unbounded. Therefore, if the matrix $[I-\gamma\Delta\tau(Y/A)]^{-1}[I+(1-\gamma)\Delta\tau(Y/A)]$ contains an eigenvalue outside the range shown in equation III-2, the solution is unstable, even if in a particular calculation the coefficient g_j that corresponds to this eigenvalue is zero.

However, as mentioned previously, the exponentials of the continuous transient solution go to zero with no oscillation, and for the difference solution to have no oscillations, the necessary and sufficient conditions are more restrictive in that all the eigenvalues must also be non-negative.

$$0 \leq q_j \leq 1 \qquad j = 1, 2, \ldots, S \qquad \text{(III-3)}$$

In order to relate the necessary and sufficient conditions in equations III-2 and III-3 to the location of the nodes, Y/A matrix, and the time differencing parameters, equation II-57 is used. First,

remembering that as the λ_j's are negative they can be ranked in the following order:

$$\lambda_{min} < \ldots\ldots < \lambda_{MAX} \leq 0 \qquad \text{(III-4)}$$

where λ_{min} is the smallest λ_j, as they are negative, but has the largest absolute value,

$$|\lambda_{min}| > \ldots\ldots > |\lambda_{MAX}| \geq 0 \qquad \text{(III-5)}$$

The λ_{min} determines the minimum q_j, q_{min}, or the q that would be negative; λ_{MAX} determines q_{MAX}, and λ_{MAX} is equal to or less than zero. Combining these considerations with equation II-57, the inequalities, equations III-2 and III-3, and the fact that $\Delta\tau$, $(1-\gamma)$, and γ are positive, only the lower bound on q can limit the selection of $\Delta\tau$ and γ. The necessary and sufficient conditions are

For Stable Solution:

$$|\lambda_{min}|(1-2\gamma)\Delta\tau < 2 \qquad \text{(III-6)}$$

For Non-Oscillatory Solution:

$$|\lambda_{min}|(1-\gamma)\Delta\tau \leq 1 \qquad \text{(III-7)}$$

These then are the rigorous, necessary, and sufficient conditions to obtain the desired exponential behavior, since when they are fulfilled all q_j's must fall in the required range.

Directly from the criteria, stable solutions are obtained for all values of $\Delta\tau$ if

$$\gamma \geq \tfrac{1}{2} \tag{III-8}$$

and further the solution is never oscillatory if

$$\gamma = 1 \tag{III-9}$$

This agrees with the stability and oscillatory behavior for the one-dimensional problem with equally spaced nodes (15).

Although the eigenvalues for a given Y/A matrix can be found numerically by iteration or Jacobi's method (16), this procedure is probably as time consuming as carrying out several trial combinations of $\Delta\tau$ and γ. However, two easily determined matrix norms are known to be greater than or equal to the absolute value of the eigenvalue which is greatest in absolute value, $|\lambda_{min}|$ (Faddeva (16), pages 55-59). These norms are found by computing the sum of absolute values of the elements in each row and column. The norms are then the maximum of the individual row sums and the maximum of the individual column sums. Defining the element in the ith row and j^{th} column of the Y/A matrix by μ_{ij}, and of the W matrix by w_{ij}, the norms are

Row Norm:

$$\left\|\frac{Y}{A}\right\|_I = \underset{i}{MAX} \sum_{j=1}^{S} |\mu_{ij}| = \underset{i}{MAX} \left[\frac{1}{A_i} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{R} y_{ij} + \sum_{\substack{j=1 \\ j\neq i}}^{S} y_{ij} \right\} \right] \tag{III-10}$$

Column Norm:

$$\left\|\frac{Y}{A}\right\|_{II} = \underset{j}{MAX} \sum_{i=1}^{S} |\mu_{ij}| = \underset{j}{MAX} \left[\frac{1}{A_j} \sum_{\substack{i=1 \\ i\neq j}}^{R} y_{ij} + \sum_{\substack{i=1 \\ i\neq j}}^{S} \frac{y_{ij}}{A_i} \right] \tag{III-11}$$

Row and Column Norm:

$$\|W\|_I = \|W\|_{II} = \underset{i}{\text{MAX}} \sum_{j=1}^{S} |w_{ij}| = \underset{i}{\text{MAX}} \frac{1}{A_i}\left[\sum_{\substack{j=1 \\ i\neq j}}^{R} y_{ij} + \sum_{\substack{j=1 \\ j\neq i}}^{S} \frac{y_{ij}}{\sqrt{A_i A_j}}\right] \quad \text{(III-12)}$$

The first equality in these three equations is the definition of the norms and represents computing the sum of the absolute values for each row and column of the Y/A and W matrices, followed by selecting the maximum of the row sums for the row norm I and maximum column sum for the column norm II. The second equality shows these sums in terms of the individual elements of the A, Y, and Y_B matrices. The first summation over all the nodes, R, represents the diagonal element of the Y and W matrices; the second summation over the unknown temperature nodes, S, is the sum of the off-diagonal elements in the row or column. The second equality is useful in understanding the effect of the location of the temperature points on the nodes, and also since the W matrix is not usually calculated explicitly, its norm can be found directly from the second equality in equation III-12.

For a lower bound on $|\lambda_{min}|$ we use the fact that for a symmetric matrix such as W $|\lambda_{min}|$ is greater than or equal to the absolute value of the maximum diagonal element. Since W has the same eigenvalues and the same diagonal elements as Y/A, this is a satisfactory lower bound for $|\lambda_{min}|$.

Now defining M as the minimum of the three norms defined by equations III-10, III-11, and III-12, the largest eigenvalue must fall in

the range

$$\underset{i}{\text{MAX}}\ \frac{1}{A_i}\ \sum_{\substack{j=1 \\ i\neq j}}^{R} y_{ij} = \underset{i}{\text{MAX}}\ |\mu_{ii}| = |\mu_{\rho\rho}| \leq |\lambda_{min}| \leq M = \text{min of} \quad \begin{matrix} \left\|\frac{Y}{A}\right\|_I \\ \left\|\frac{Y}{A}\right\|_{II} \\ \|W\|_{II \text{ and } I} \end{matrix} \qquad \text{(III-13)}$$

Therefore, useful and simply found sufficient conditions for stable and non-oscillatory solutions are:

Stable Solution: $M(1-2\gamma)\Delta\tau < 2$ (III-14)

Non-Oscillatory Solution: $M(1-\gamma)\Delta\tau \leq 1$ (III-15)

Again, in the stability criterion, if the value of $M(1-2\gamma)\Delta\tau$ is allowed to equal 2, possibly a q of -1 may occur.

As an example of the usefulness and simplicity of these stability criteria, we will apply the criteria to the problem of one-dimensional heat flow in a solid of constant thermal properties with a regular mesh spacing $\Delta\xi$ starting from both boundaries, Figure IV-2, mesh $\Delta\xi$. The boundary conditions are a specified temperature on one boundary and a zero gradient at the other. Under these conditions the Y/A matrix is

$$\frac{Y}{A} = \frac{-1}{(\Delta\xi)^2} \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -2 & 2 \end{bmatrix} \qquad \text{(III-16)}$$

The three norms in this case are equal and M becomes by inspection

$$M = 4/(\Delta\xi)^2 \qquad \text{(III-17)}$$

Defining the modulus r in the usual manner,

$$r = \frac{\Delta\tau}{(\Delta\xi)^2} = \frac{K\Delta\theta}{(\Delta x)^2} \qquad \text{(III-18)}$$

We obtain directly the familiar sufficient conditions

Stable Solution: $r(1-2\gamma) < \frac{1}{2}$ (III-19)

Non-Oscillatory Solution: $r(1-\gamma) \leq \frac{1}{4}$ (III-20)

Normally these results are derived by a complicated procedure involving separation of variables and Fourier analysis (2, 10, 15, 17) which actually determines the q_j's. However, as shown in Chapter IV, section E-4, the matrix procedure above is more satisfactory, even when the q_j's can in principle be found. If, on the sufficient condition for stability, the value of $r(1-2\gamma)$ is allowed to equal $\frac{1}{2}$, a -1 root can occur. Although this lack of equality would indicate graphical methods

where γ is zero and r is $\frac{1}{2}$ are unstable, as the above condition is merely sufficient, most graphical solutions are stable for many problems; for problems with two adiabatic boundaries or specified flux conditions a -1 root does occur in some graphical methods. It is also known that with a large finite heat-transfer coefficient, some graphical and other approximate solutions become unstable. This result is easily predicted from the matrix norms, as a diagonal element increases linearly with the heat-transfer coefficient, and further, the norm allows an r to be selected that eliminates the undesired behavior. These examples show the inherent simplicity and usefulness of the rigorous criteria in equations III-14 and III-15.

(C) SELECTION OF $\Delta\tau$ AND γ BASED ON STABILITY

At this time, it is convenient to discuss the selection of the differencing parameters $\Delta\tau$ and γ based on the stability criteria and matrix relationships. Obviously, because of the wide variety of problems possible, the different purposes for which the results are to be used, and the capacity of available computing equipment, this discussion must be general so that when one is confronted with a problem a reasonable selection can be made.

In applying the sufficient criteria to the general Y/A matrix for an asymmetric location of nodes, one might ask how much smaller a time increment would be found by using one of these inequalities than the actual minimum necessary time increment. This would be particularly important when the calculations are to be carried to a fixed time where an explicit method or low γ method is desirable, possibly for ease of calculation in the first case or accuracy in the second. From the bounds on λ_{MAX}, equation III-13, and the structure of the specific Y/A matrix, as defined by MacNeal's rules, the largest range for λ_{MAX} is such that the upper bound is at most twice the lower bound. This would occur only when

$$\left\| \frac{Y}{A} \right\|_1 = M = 2\left|\mu_{\rho\rho}\right| = \frac{2}{A_\rho} \sum_{j=1}^{S} y_{\rho i} \qquad \text{(III-21)}$$

Thus the time increment found from the sufficient criteria, equation

III-14 and equation III-15, would never be less than half that necessary. Although in the worst case this could mean doubling the amount of calculations to a constant time τ_1, the criteria do give the correct order of magnitude for $\Delta\tau$. If they show the calculation to be infeasible for the desired γ, consideration should be given to a different arrangement of nodes, or to other approximations that may be made, such as neglecting the heat capacity of nodes in high conductance regions (9). However, for any node arrangement, $\Delta\tau$ can be made as large as desired, and a stable or even a non-oscillatory solution can be obtained, by selecting γ large enough.

In some cases, the time increment can be increased over that shown in the sufficient conditions, equations III-14 and III-15, and useful solutions are obtained. Short of actually numerically determining $|\lambda_{min}|$, one way to do this would be to reduce the range for $|\lambda_{min}|$ and, in particular, to lower M. This can be done by performing a similarity transformation on either the Y/A or W matrices, forming similar matrices with the same eigenvalues, but lower norms. These transformations are

$$U_1 = F_1 \frac{Y}{A} F_1^{-1} \qquad \text{(III-22)}$$

or

$$U_2 = F_2 W F_2^{-1} \qquad \text{(III-23)}$$

and since either of the U's would have the same eigenvalues, $|\lambda_{min}|$

would have to be more than the absolute value of the largest diagonal element and smaller than the smallest norm. A particularly simple and useful form for F is a diagonal matrix with the elements in the ith row designated by φ_i and the jth column by φ_j ; its inverse is then known explicitly, and the new U is also known explicitly. Its diagonal elements are the same as those of Y/A; the off-diagonal element of the ith row and j^{th} column is the corresponding element in the old matrix times φ_i/φ_j. That is, for a transformation on Y/A the elements of U_1 are given by

$$u_{ii} = \mu_{ii} \tag{III-24}$$

$$u_{ij} = \frac{\mu_{ij}\,\varphi_i}{\varphi_j} \tag{III-25}$$

By properly selecting the φ_j's the maximum row (or column) sum can be reduced, while the other row or column sums are increased. Although the lowest sum would occur when all row (or column) sums are equal, this is not practically possible. Also, it is probable that because of the many different matrices possible no one method of selecting the φ's is satisfactory. However, by remembering that φ_j multiplies each off-diagonal element in the jth column and $1/\varphi_j$ multiplies each off-diagonal row element of the jth row, it is feasible to increase the lower row (or column) sums in such a manner that the larger row (or column) sums are significantly reduced. Often it is useful to assume

an approximate sum that falls about midway in the range given by equation III-13 and try to make all row (or column) sums lower than this number.

Another more complicated transformation is the one used in the Jacobi method for eigenvalues of a real symmetric matrix. This method finds the eigenvalues by successive transformations that eliminate the off-diagonal elements. By applying this type of transformation to the largest off-diagonal elements in the W matrix, the range within which the maximum eigenvalue falls (equation III-13) is reduced by increasing the maximum diagonal element and/or reducing the norms. This transformation changes the elements in the two rows and two columns for each element eliminated and equations for the new matrix are given in the references (18). If only a few very large off-diagonal elements were in the matrix this method might prove very useful in reducing the range for $|\lambda_{min}|$.

Since either of the above transformations attacks the rows (or columns) whose sums are much larger than the other sums, neither could be applied to a matrix such as in equation III-16 for a one-dimensional regular mesh spacing. Further, since we can show (Chapter IV, section E-3) that the maximum eigenvalue of the type of matrix in equation III-16 goes to the norms as the dimension of the matrix is increased to infinity, it is concluded that the minimum norm obtained for a Y/A matrix, whose row and column sums are nearly equal, is very probably not much larger than its maximum eigenvalue.

Since the effect of a q_{min} that is negative but small in absolute value is negligible as n becomes large, under some circumstances an oscillatory solution might be tolerated. This would allow a larger $\Delta\tau$ to be used. Consequently we could specify that q_{min} be less than Y in absolute value so that at the lowest time of interest τ_0, or after $\tau_0/\Delta\tau$ time increments, the oscillatory component would be less than Y^{n_0} or $Y^{\tau_0/\Delta\tau}$ and would no longer be significant. The disregarding of the approximate solution for the first few n's must usually be done in any case because the approximate solution has replaced the infinite sum of exponentials of the continuous solution with a finite sum of exponential forms. The sufficient condition to obtain q_{min} less in absolute value than Y is

$$(\Delta\tau)M[1-\gamma(Y+1)] \leq (1+Y) \qquad \text{(III-26)}$$

In general, the value of Y which makes $Y^{\tau/\Delta\tau}$ insignificant at the time of interest is satisfactory for use in equation III-26 to select γ and $\Delta\tau$. For many purposes Y can be as high as 0.6. However, since increasing the time increment reduces all the q_j's, care should be taken that q_{MAX} is significantly larger than Y or the solution will probably show excessive oscillations. Although for most reasonable selections of the time differencing parameters this will not occur, it is possible with a γ of $\frac{1}{2}$ and a very large time increment $\Delta\tau$ to have a stable solution with all the q_j's including q_{MAX} negative. In cases where it is necessary q_{MAX} can be estimated in one of the following

three ways. First a lower bound for q_{MAX} can be derived from the fact that

$$\lambda_{MAX} \geq \mu_{00} \tag{III-27}$$

where μ_{00} is the largest diagonal element of Y/A (smallest in absolute value and negative). This gives for q_{MAX}:

$$q_{MAX} \geq \frac{1 + \mu_{00}(1-\gamma)\Delta\tau}{1 - \mu_{00}\gamma\Delta\tau} \tag{III-28}$$

This lower bound is usually not close to q_{MAX} and consequently it might not be useful. The second way of estimating q_{MAX} for a given time increment is to find the smallest υ_j root for a problem like the one to be approximated; $e^{-\upsilon_1^2\Delta\tau}$ is then an estimate of q_{MAX}. Third, a calculation using a small number of points and time increment $(\Delta\tau)_1$ can be stepped out until the logarithm of the approximate temperature at a point is linear with time. From this the q_{MAX} for the specific calculation can be obtained. Then $(q_{MAX})^{(\Delta\tau)_2/(\Delta\tau)_1}$ is the required approximation to the q_{MAX} for time increment $(\Delta\tau)_2$.

It should be noted that limiting q_{min} makes no allowances for the weighting or coefficients that multiply q_{min}^n in the solution. These coefficients are the associated eigenvector and the coefficient in the initial vector corresponding to the q_{min}. Indeed, the initial motivation for the study in Chapter IV was to find why, under the same conditions, one of two graphical methods with the same difference eigenvalues had

an oscillatory component so large as to make the solution useless, but the second method gave useful approximate solutions.

To eliminate a large oscillatory component, without having to use either a very small time increment or a large weighting factor for all the calculations to very long times, the following scheme can be used. First, the time increment $(\Delta\tau)_1$ and weighting factor γ are selected so that q_{min} is not negative and the contribution after n_1 time increments from $q_{min}^{n_1}$ is small. The estimate of $(\Delta\tau)_1$ and γ_1 can be based on the upper bound M. Then at step n_1 the temperature distribution should be smooth and a larger $(\Delta\tau)_2$ and smaller or zero γ_2 can be used for the rest of the calculation with a reasonable assurance that any oscillatory component is probably insignificant; obviously the selection of $(\Delta\tau)_2$ and γ_2 is still limited as no q can be smaller than -1. To see that this procedure reduces the weighting on the oscillatory component, consider the calculated approximate temperature distribution at increment n_1 as the initial condition for the second choice of $(\Delta\tau)_2$ and γ_2. The solution for this difference calculation becomes

$$t_n = t_{P\,n} - C\,Q_1^{\,n} g \qquad 0 \le n \le n_1 \qquad \tau = n(\Delta\tau)_1 \qquad \text{(III-29)}$$

$$t_n = t_{P\,n} - C\,Q_2^{\,n-n_1}\,Q_1^{\,n_1} g \qquad n \ge n_1 \qquad \text{(III-30)}$$

$$\tau = n_1(\Delta\tau)_1 + (n-n_1)(\Delta\tau)_2$$

where Q_1 is the diagonal matrix of q_{j1}, the difference eigenvalues based on $(\Delta\tau)_1$ and γ_1, and Q_2 is the matrix of q_{j2} based on $(\Delta\tau)_2$ and

γ_2. Now the weighting of $q_{min\ 2}$, which is negative for the second $(\Delta\tau)_2$ and γ_2, is $g_S\, q_{min\ 1}^{n_1}$, which is small, so the oscillations should be negligible for the remaining calculations. In the above equations, III-29 and III-30, no change in the particular solution was shown when the γ and $\Delta\tau$ were changed; thus, the conclusion applies rigorously only for problems where $t_{P\ n}$ is not a function of n but is a true steady-state solution. Although problems where the particular solution does change with time and is a quasi-steady-state solution have not been studied, it is very probable that the above scheme would eliminate spurious oscillations for all but a few types of forcing functions.

A second way of reducing oscillations is by averaging the calculation for two successive times n and $(n+1)$ and then applying the result at n, $(n+1)$, or an intermediate value of n. By examining the form of the solution equation it is seen that this has the effect of multiplying each weighting g_j of the jth eigenvector-eigenvalue product by $(q_j+1)/2$:

$$g_{j\ \text{averaged}} = \left(\frac{q_j+1}{2}\right) g_{j\ \text{unaveraged}} \qquad \text{(III-31)}$$

If q_{min} is close to -1, and q_{MAX} is close to +1, averaging significantly reduces the weighting q_{min} of the negative eigenvalue-eigenvector product without changing the weighting g_{MAX} of the q_{MAX} eigenvalue-eigenvector product. However, the weighting for the moderate sized

q's is radically changed. The effect of averaging is discussed more precisely for one-dimensional problems in Chapter IV, sections B-7, D-2 and E-6. It is shown to be mostly useful in graphical methods in which the q_{min} has the same absolute value as q_{MAX}.

The criteria shown in equations III-14 and III-15 should be compared and related to another criterion often suggested (9) for the explicit method. This other criterion is that none of the coefficients of $[I + \Delta\tau(Y/A)]$ be negative to avoid wild oscillations. For cases of uniform mesh spacing simple arguments show that this is a sufficient requirement (3). For the asymmetric network and the explicit method this amounts to the condition that

$$\Delta\tau \leq \frac{1}{|\mu_{\rho\rho}|} \qquad \text{(III-32)}$$

Referring to equations III-13 and III-15, this is seen to be equivalent to assuming that the absolute value of the minimum eigenvalue, $|\lambda_{min}|$, is less than, or equal to, $2|\mu_{\rho\rho}|$ and equivalent to substituting $2|\mu_{\rho\rho}|$ into the stability ctieria. Although this condition is sufficient for stability, if the time increment is selected by the equality, the solution probably contains an oscillatory component, and also it does not eliminate the possibility of a -1 root. In graphical methods for one dimension the time increment used conforms to the equality in equation III-32 and, as mentioned previously, oscillations that make the solution useless are sometimes obtained. The criteria derived in this section

have the advantage of being almost as simple and yet rigorous and, with proper modifications, they allow the selection of γ and $\Delta\tau$ to obtain the type of solution required with almost the largest time increment possible.

In conclusion, to obtain a satisfactory approximate solution, the time increment and weighting γ must be selected to give a stable solution. The precise selection of the time increment and γ must then be based on the specific purpose of the approximation. In general, the combination of time increment and γ should be selected which gives an accurate representation of how the continuous exponential varies with time, and simultaneously minimizes the calculations necessary, and keeps any oscillations from destroying the usefulness and accuracy of these results. Although it is probably not possible to derive a general set of rules that would apply to all asymmetric distributions, this question is considered in detail in Chapter IV for one-dimensional regular networks in Cartesian coordinates and some generalizations are possible.

For some problems the selection of the time increment may be fixed or limited by considerations other than stability or oscillatory effects. For example, if the boundary temperature contains oscillatory components, the time increment must be smaller than one-half the smallest period of oscillation of the oscillatory components. This can be demonstrated rigorously from Shannon's sampling theorem (12) or

can be seen by considering that, if the value of the oscillating function is not sampled at least this often, a wholly inadequate description of the oscillations results.

(D) NUMERICAL EXAMPLE

In order to numerically check the necessary and sufficient criteria for stable and non-oscillatory solutions, a transient heat transfer problem was solved using an asymmetric distribution of temperature points for several combinations of time increments and weighting factors. In addition, the norms of Y/A and W were studied to show the usefulness of the sufficient conditions and the closeness of the bounds for $|\lambda_{min}|$. The presentation of an example problem also affords an opportunity to mention some of the practical points concerning the calculation of the Y/A matrix, and procedures for solving the equations for the implicit methods. The problem selected is one where the solution for the partial differential equation is known, so that it also can be used to study the accuracy.

The problem is to find the transient temperature distribution in a solid with a unit square cross-section of uniform thermal properties, and a uniform initial temperature distribution of unity, which has its surfaces maintained at zero for all times greater than zero. Taking the center of the solid as the origin, the solution of the two-dimensional partial differential equation can be found as the product of solutions of one-dimensional problems:

$$T(\xi,\eta,\tau) = \Omega(\xi,\tau)\Omega(\eta,\tau) \qquad -\tfrac{1}{2} \le \xi \le \tfrac{1}{2} \qquad \text{(III-33)}$$

$$-\tfrac{1}{2} \le \eta \le \tfrac{1}{2}$$

$$\tau \ge 0$$

where

$$\Omega(\xi,\tau) = \frac{4}{\pi} \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)} e^{-\pi^2(2j+1)^2\tau} \cos(2j+1)\pi\xi \qquad \text{(III-34)}$$

This problem is important because its solution has the same eigenvalues and eigenfunctions as the transient solution for problems where the temperatures of the surfaces of the unit square are specified as any function of space and/or time. In these other cases, only the particular solution $T_P(\xi,\eta,\tau)$, which is zero above, and the Fourier coefficients a_{ij}, which are $16(-1)^{i+j}/\pi^2(2j+1)(2i+1)$ above, would be changed.

For the approximate solution, points need be located only on the minimum area necessary, considering the symmetry of the problem. This area is the 1/8 of the cross-section as shown by the triangle in Figure III-1. Nine points where the temperature is an unknown function of time are located within the solid, and five points where the temperature is known as a function of time are located along the surface. The variable temperature points are located away from the axes of symmetry, which can be considered as adiabatic boundaries. By locating the points away from the boundary, the number of nodes required to be calculated for the small cross-section is not raised, but the equivalent of the total number of nodes in the whole cross-section is

significantly raised. Moreover, since the whole area of the surface nodes A, B, C, D, and E is assumed to be at the boundary temperature of zero, the heat capacity of the solid corresponding to the total heat capacity of these nodes is neglected. The network of points was made irregular so that the effects of asymmetric location on the accuracy could be studied. However, this could be considered as a practical arrangement for a problem where only the corner temperatures are important.

The Y/A matrix was then calculated using exact geometric relationships for all the lengths, and thus errors are probably less than 0.01 per cent in the elements of Y/A. In a practical case, sufficient accuracy probably is obtained by making a large-scale drawing and measuring the lengths. A simple digital computer program could be devised to determine the Y/A matrix from arrays of the r_{ij} and ℓ_{ij} in a matrix form, together with the effective thermal conductivities for each connector, and heat capacity data for each node. A program to find Y/A from just the location of the temperature points, the specification of the connected points, and the effective thermal properties, probably could be devised although it would be very complicated.

The Y/A matrix for the arrangements of nodes in Figure III-1, as computed using a desk calculator, is in Table III-1. The first row in this matrix contains the coefficients in the ordinary differential equation for node 1, the second row for node 2, etc. Since a numerical

check was to be made of the necessary and sufficient criteria, the eigenvalues of the matrix were determined using a method based on the general numerical method of iterative multiplications of the Y/A matrix (16), and these are in Table III-2. From the $|\lambda_{min}|$ of 1442, the limiting time increments for stable and non-oscillatory solutions have been calculated for γ's of 0, $\frac{1}{2}$, and 1 from equations III-14 and III-15, and are summarized in the table. In order to check these criteria, the approximate solution was stepped out using equation II-39 for these γ's and for time increments on both sides of the critical values. The results are shown as graphs relating the calculated approximate temperature at node 1 and dimensionless time in Figure III-2 for the explicit case of γ of zero, and in Figure III-3 for the implicit γ's of $\frac{1}{2}$ and 1. Node 1 was selected as it is the most sensitive to instability and oscillatory behavior. Also, for comparison, the continuous solution of the partial differential equation for that point is shown.

For the explicit calculation an unstable solution is expected for a time increment $\Delta\tau$ greater than 0.001387, equal undamped oscillations for a $\Delta\tau$ of 0.001387, damped oscillations for a $\Delta\tau$ between 0.001387 and 0.000694, and smooth damped exponentials for a positive $\Delta\tau$ less than 0.000694. These expected types of solutions were obtained as shown in Figure III-2. Referring to Figure III-2, the midpoint of equal oscillations shown for the limiting time increment for stability, 0.001387, (b) decreases and approaches zero for large times.

The oscillatory solution for a $\Delta\tau$ of 0.001, although useless for the first four increments, does give a fair approximation for long times. The solution found by using the maximum time increment which gives no negative coefficients in the calculation matrix, $[I + \Delta\tau(Y/A)]$ (Longwell's sufficient criteria, equation III-32), although oscillatory, gives a better approximation to the continuous solution shown in (f). Approximate solutions using this criterion for similar problems also show oscillations, but they are so rapidly damped (19) that they are not significant. The smallest increment shown, $\Delta\tau = 0.0001$ (e), is seen to give an accurate approximation to the continuous solution. Not shown are good approximations with time increments of 0.0005 and 0.000694.

The implicit calculations were calculated by finding the inverse matrix and then stepping out the solution explicitly using equation II-39. For a γ of $\frac{1}{2}$ the solution is always expected to be stable, and should be non-oscillatory for positive time increments less than 0.001387; for a γ of 1, neither instability nor oscillatory behavior is expected. The graphs, together with other numerical results in Figure III-3, show that the expectations are fulfilled. However, for the large $\Delta\tau$ of 0.01 and a γ of $\frac{1}{2}$ the solution oscillates (damped) so badly that it is useless as an approximation. The corresponding calculation for a γ of 1 at a $\Delta\tau$ of 0.01, although a smooth damped exponential, is not an accurate approximation as discussed later. Comparison of the

non-oscillatory solutions for a small time increment of 0.0005 with the continuous solution, Figure III-3 (b), (d), and (e), shows that the approximation with γ of $\frac{1}{2}$ appears to be more accurate than with a γ of 1 or zero.

Although a detailed discussion of accuracy is deferred until the last chapter, several brief comments can be made. First, the approximate solutions that are stable and which have only small oscillations are reasonably accurate. More accurate results are obtained for the smaller nodes, 1, 2, and 3; less accurate results for the large nodes 6 and 7, as would be expected. For the large time increments the most accurate results are with a γ of $\frac{1}{2}$. As the time increments are reduced, the errors appear to approach a constant for each node indicating that the approximate solution approaches the analog as the time increment goes to zero.

To show that, when $\Delta\tau$ and γ are selected so that a q of -1 occurs, the calculated temperature can go to infinity even though the surface temperatures are bounded, the explicit calculation for the limiting time increment for stability was repeated but with the surface temperature of the nodes A, B, C, D, E oscillating between +1 and -1.

$$t_f = (-1)^n \tag{III-35}$$

The temperatures of the surface nodes and of node 1 are shown in Figure III-4. The temperature of node 1 oscillates with a linearly increasing amplitude indicating that a term of the form $n(-1)^n$ must be

in its solution. Consequently, the solution is unbounded and the limiting $\Delta\tau$ and γ combination, although bounded for some boundary conditions, is unstable according to the definition of stability used here. Also, this shows that with an oscillatory forcing function the time increment should not be a multiple of the period of oscillation or the oscillation might excite an oscillatory difference eigenvalue which does not occur in the continuous solution.

Thus, the necessary and sufficient criteria for stable and non-oscillatory behavior have been checked. However, practically, the eigenvalue $|\lambda_{min}|$ probably would not be obtained, but the sufficient conditions given by equations III-14 and III-15 in terms of the matrix norms would have been used. The calculated matrix norms for the Y/A matrix are in Table III-2. The row norm $\| \frac{Y}{A} \|_I$ is the smallest at 1618.8 and the largest (in absolute value) diagonal element is 1266.7, which defines the range for $|\lambda_{min}|$. If M had been used as 1618.8 the sufficient time increment computed for a constant γ would be about 12 per cent lower than the necessary and sufficient increment. Also, from the diagonal element of Y/A that is largest in absolute value, maximum of $|\mu_{ii}|$ of 1266.7, a time increment based on the minimum norm of Y/A would not be more than 22 per cent lower than necessary. The symmetrized Y/A matrix, or W matrix, can be calculated directly or from the elements of Y/A, and is shown in

Table III-3. Its norm is lower than those of Y/A and is 1583.3, which, if used in the sufficient condition, would give a $\Delta\tau$ ten per cent smaller than necessary. A simple diagonal similarity transformation on the W matrix as suggested in equation III-23 gives a U matrix which is shown together with the transformation matrix in Table III-4. The U matrix has a minimum norm (column) of 1501.8 which is only four per cent greater than the $|\lambda_{min}|$. The trial and error calculation of the F matrix and the simultaneous calculation of U required about an hour's time using a slide rule and adding machine; the transformation F was selected so that it reduced the off-diagonal elements in the fourth and fifth rows and second and third columns without increasing the other column sums.

A numerical study was also made of the norms of Y/A, maximum diagonal element $\underset{i}{MAX}\,|\mu_{ii}|$, and $|\lambda_{min}|$ for the two Y/A matrices used by Longwell in his rocket motor study (9). For his shape 1, fourteen equations were used. The minimum norm of Y/A itself was only 9 per cent greater than $|\lambda_{min}|$ and about 13 per cent greater than the maximum $|\mu_{ii}|$. Sixteen equations were needed for his shape 2, and the minimum norm was 8 per cent larger than $|\lambda_{min}|$ and 35 per cent larger than the maximum $|\mu_{ii}|$.

From the limited experience of these three problems, the norm of Y/A appears to be a good estimate of the absolute value of the eigenvalue which determines the stability and oscillatory behavior of

the difference solution. Replacing $|\lambda_{min}|$ with the norm to compute the limiting time increment gives, on the average, a time increment only about 10 per cent too small. However, in these problems the minimum norm of Y/A does not approach being 100 per cent greater than the maximum diagonal element, as is theoretically possible. The only case when this occurs is when a node with the largest diagonal element, $\sum_{j=1}^{R} y_{ij}/A_i$, is connected only to interior temperature nodes where the temperature is to be calculated or when it is bounded by an adiabatic boundary. However, it is probable that in the most frequently encountered problems, the maximum diagonal element occurs for a point which has connectors to a known fluid temperature and thus one of the conductances appears in the Y_B/A matrix. Consequently, the row sum is less than twice the diagonal element which is the sum over all conductances. In these cases, the norm of the Y/A matrix probably can be used satisfactorily without further calculation; for the problems where the norm and diagonal element give too large a range for $|\lambda_{min}|$, the calculation of the norm for W and/or a diagonal transformation may be used.

The number and location of the points also affects the limiting time increment. Increasing the number of points reduces the dimensionless heat capacity A_i's of some or all of the points which in turn raises the norm, diagonal elements, and $|\lambda_{min}|$ requiring a lower time increment or an implicit method with a relatively high γ. From the

previous consideration of the Y/A matrix, the elements associated with the node or nodes having the smallest dimensionless heat capacities A_i's are the ones that usually determine the norm of the matrix and the maximum diagonal element. Consequently, these nodes can greatly reduce the allowable time increment. Thus, to prevent one or a few nodes from restricting the time increment too severely, the points should be located, in so far as practical, so that the elements in Y/A are of the same order of magnitude.

Locating S points in a regular fashion for a solid of uniform thermal properties gives a Y/A matrix with smaller norms and smaller maximum diagonal element than the Y/A matrix for locating the S points so that some nodes have smaller heat capacities. The Y/A for the asymmetric case also very probably has a large $|\lambda_{min}|$, and thus a smaller time increment must be used.

(E) ALTERNATE CALCULATION FOR IMPLICIT METHODS

In calculating the approximate solution for the implicit methods, the inverse matrix $[I-\gamma\Delta\tau\,(Y/A)]^{-1}$ was first computed with an elimination method; then the calculation matrices of $[I-\gamma\Delta\tau(Y/A)]^{-1}[I+(1-\gamma)\Delta\tau(Y/A)]$ and, if necessary, $[I-\gamma\Delta\tau(Y/A)]^{-1}[\Delta\tau(Y_B/A)]$ were found. The approximate solution was then stepped out using equation II-39 explicitly as a matrix-vector product, using the known temperature vector at the beginning of the increment and known boundary temperature vector. Although this method is simple, it suffers from two disadvantages. First, as the elements of the inverse matrix $[I-\gamma\Delta\tau(Y/A)]^{-1}$ contain round-off error, significant additional errors might be introduced into the approximate solution by the successive multiplications by that matrix. Of more importance is that under certain circumstances methods of solution of the system of equations II-39 are available that require fewer of the time-consuming multiplication operations than does a vector-matrix product. Thus, for these cases, the system of equations II-39 can be solved for each time increment for t_{n+1} more rapidly than the vector-matrix product can be calculated.

The equation to be solved is

$$[I-\gamma\Delta\tau\,\frac{Y}{A}]t_{n+1} = [I+(1-\gamma)\Delta\tau\,\frac{Y}{A}]t_n + \frac{\Delta\tau\,Y_B}{A}(\gamma\,t_{B\,n+1} + \left\{1-\gamma\right\}t_{B\,n}) \qquad \text{(III-36)}$$

where the quantity on the right side of the equality represents a known

vector and t_{n+1} is to be found. The method for solution of the one-dimensional case is based on simplifications in the elimination method which are possible because only the three major diagonals of $[I-\gamma\Delta\tau(Y/A)]$ and $[I+(1-\gamma)\Delta\tau(Y/A)]$ contain non-zero elements. For two- or three-dimensional problems, where the temperature points are located on the corners of a rectangle or rectangular solids, line iterations and alternating direction methods (6) are useful. However, they do not give the exact solution of the system of equations III-36, but actually represent a slightly different method, and further it appears that they cannot be modified for an asymmetric distribution of points. A modification of the elimination solution for the one-dimensional case can be devised which can solve the system of equations for the irregular network more rapidly than the vector-matrix product can be formed, providing the multiplication by zero is avoided, or providing it is significantly faster than multiplication by non-zero elements. This method for the general asymmetric network is done by using an elimination technique to convert the system of equations III-36 to an upper triangular system, which can then be solved explicitly. After determining the constants on the right side of the equation (note that because of the many zeros in both matrices on the right side the vector-matrix products here are formed much more rapidly than would be the case with the inverse matrix which probably has no zeros) the conversion to an upper triangular system requires the following steps:

(1) Solve the first equation for the first variable, $t_{1,n+1}$. As $[I+\gamma \Delta\tau(Y/A)]$ is sparse, it should only require three or four divisions by the diagonal element μ_{11}.

(2) Eliminate $t_{1,n+1}$ from all the following equations by substituting the expression found in step (1). Since $[I+\gamma \Delta\tau(Y/A)]$ is a sparse matrix, this should require substitution in only two or three equations.

(3) Repeat steps 1 and 2 for the next equation and variable, eliminating the next temperature point from all succeeding equations. Continue until all the coefficients below the diagonal in $[I+\gamma \Delta\tau(Y/A)]$ are zero. Now the last variable, $t_{S,n+1}$, is determined explicitly by the last equation. Knowing $t_{S,n+1}$, $t_{S-1,n+1}$ is calculated explicitly from the (S-1) equation. Following this procedure back through the equations each component of t_{n+1} is found. This procedure, including the calculation of the known vector on the right side of the equation, requires a number of multiplicative operations involving non-zero elements directly proportional to S; on the other hand, the inverse matrix-vector product has S^2 multiplicative operations. Consequently, the advantage of using the modified elimination method depends upon the proportionality constant and the number of points used.

The proportionality constant depends upon exactly where the non-zero elements are located in the $[I-\gamma \Delta\tau(Y/A)]$ matrix. For the one-dimensional case where each node has two neighbors, 8S

multiplicative operations are needed to solve the system, including the determinations of the constants on the right side of the equation. For a two-dimensional problem where each node is surrounded by an average of four nodes, the solution requires about 16S multiplicative operations. Thus, if S is greater than 8 or 16 for the one- or two-dimensional system, the system of equations III-36 can be solved at each time increment more rapidly than the inverse matrix-vector product can be constructed.

(F) CONCLUSIONS

In Chapter III the criteria for stability and non-oscillatory approximations were derived as simple relationships involving the time increment $\Delta\tau$, the weighting factor γ, and the eigenvalue of the Y/A matrix with the largest absolute value for a fixed network of points. The necessary and sufficient criteria for both stable approximations and non-oscillatory approximations were then checked numerically by approximating a problem using an asymmetric network of points.

Under any practical conditions, it is too time consuming to determine this eigenvalue. Therefore, sufficient conditions were derived in terms of easily calculated norms of the Y/A matrices. These matrix norms were then shown to give the same stable and non-oscillatory criteria usually found by Fourier analysis for one-dimensional problems with a regular distribution of temperature points. It was also shown that these sufficient conditions can never give a time increment less than $\frac{1}{2}$ of the time increment actually required. Applications of these sufficient conditions to the example problem and to two other two-dimensional problems using an asymmetric distribution of points showed that these norms usually are only 4 to 15 per cent larger than the absolute value of the critical eigenvalue $|\lambda_{min}|$, which indicates that the time increment determined by the sufficient conditions is expected

to be just slightly smaller than its limiting value to obtain the required behavior.

In addition to these points, a weighting γ of $\frac{1}{2}$ was shown always to give a stable solution and a weighting γ of 1 always to give a non-oscillatory solution. Also a negative q_{min} can be tolerated providing q_{min}^{n} is negligible at the times of interest. Further, a proof was given that, after starting a solution using a small time increment for several time increments, a larger time increment could be used for the remaining calculations without excessive oscillations. It was also noted that by averaging the calculated temperatures for successive time increments, the weighting of the oscillatory components was reduced.

However, a solution which is stable or non-oscillatory is not necessarily an accurate approximation. A stable solution is bounded as is the continuous solution. A non-oscillatory solution only oscillates as does the continuous solution. But although the general behavior of the approximation might be like that of the continuous solution, nevertheless the eigenvectors, the large damping factors q_j, and the initial vector components might not be good approximations to the corresponding quantities in the continuous solution as would be necessary for an accurate solution. The question of accuracy for one- and two-dimensional problems is the subject of Chapters IV and V.

Table III-1

CONDUCTANCE/CAPACITANCE MATRIX, Y/A

1/8 - Square Cross-Section

$\frac{Y}{A} =$

$j \rightarrow$ / $i \downarrow$	1	2	3	4	5	6	7	8	9	Row Sum $\sum_{j=1}^{9} \lvert \mu_{ij} \rvert$
1	-1266.7	49.80	219.30	0	0	0	0	0	0	1535.80
2	29.35	-974.15	430.81	57.44	86.16	0	0	0	0	1577.91
3	160.30	541.10	-812.59	0	109.19	0	0	0	0	1623.18
4	0	19.26	0	-302.22	50.55	23.11	7.04	0	0	402.18
5	0	34.46	34.78	60.30	-170.94	0	41.36	0	0	341.84
6	0	0	0	16.51	0	-99.01	46.37	15.48	0	177.37
7	0	0	0	63.92	31.16	58.30	-121.75	25.96	0	301.09
8	0	0	0	0	0	25.35	33.80	-87.28	28.13	174.56
9	0	0	0	0	0	0	0	320.00	-320.00	640.00
$\sum_{i=1}^{9} \lvert \mu_{ij} \rvert$	1456.35	1618.77	1497.48	500.39	448.00	205.77	250.32	448.72	348.13	

Column Sum

Table III-2

EIGENVALUES, NORMS, AND STABILITY CRITERIA

1/8 - Square Cross-Section

a. Negative of Eigenvalues

	$-\lambda$
$\lvert\lambda_{min}\rvert$	1442
	1241
	401
	354
	323
	175
	121
	80
	19

b. Norms

$$\left\|\frac{Y}{A}\right\|_I = \max_i \sum_{j=1}^{9} \left|\mu_{ij}\right| = 1618.8 \qquad i=3$$

$$\left\|\frac{Y}{A}\right\|_{II} = \max_j \sum_{i=1}^{9} \left|\mu_{ij}\right| = 1623.18 \qquad j=2$$

$$\max_i \left|\mu_{ii}\right| = 1266.7$$

c. Stability and Oscillatory Criteria

γ =	0	1/2	1
Limiting $\Delta\tau$			
$\lvert\lambda_{min}\rvert(1-2\gamma)\Delta\tau \leq 2^*$	0.001387	∞	∞
$\lvert\lambda_{min}\rvert(1-\gamma)\Delta\tau \leq 1^{**}$	0.0006937	0.001387	∞

* Stable

** Non-Oscillatory

Table III-3

SYMMETRIZED Y/A MATRIX, W = D Y D

1/8 - Square Cross-Section

$W =$

j → / i ↓	1	2	3	4	5	6	7	8	9
1	-1266.7	38.2	187.4	0	0	0	0	0	0
2	38.2	-974.2	482.8	33.3	54.6	0	0	0	0
3	187.4	482.8	-812.6	0	61.6	0	0	0	0
4	0	33.3	0	-302.2	55.2	19.5	21.2	0	0
5	0	54.6	61.6	55.2	-170.9	0	35.9	0	0
6	0	0	0	19.5	0	-99.0	52.0	19.8	0
7	0	0	0	21.2	35.9	52.0	-121.8	29.6	0
8	0	0	0	0	0	19.8	29.6	-87.3	94.9
9	0	0	0	0	0	0	0	94.9	-320.0
Column Sum $\sum_{i=1}^{9} \lvert w_{ij} \rvert$	1492.3	1583.1	1544.4	431.4	378.2	190.3	260.5	231.6	414.9

Table III-4

U MATRIX, $U = F\,W\,F^{-1}$, SIMILAR TO W AND Y/A MATRICES

1/8 - Square Cross-Section

$U =$

j → / i ↓	1	2	3	4	5	6	7	8	9
1	-1266.7	38.2	187.4	0	0	0	0	0	0
2	38.2	-974.2	482.8	1000	545	0	0	0	0
3	187.4	482.8	-812.6	0	616	0	0	0	0
4	0	1.1	0	-302.2	18.4	130	21.2	0	0
5	0	5.5	6.2	165.6	-170.9	0	107.7	0	0
6	0	0	0	2.9	0	-99 .0	7.8	1	0
7	0	0	0	21.2	12.0	348	-121.8	9.9	0
8	0	0	0	0	0	396	88.8	-87.3	9.5
9	0	0	0	0	0	0	0	94.9	-320
Column Sum $\sum_{i=1}^{9} \lvert u_{ij} \rvert$	1492.3	1501.8	1489.0	1491.9	1362	973	348	1047	330

F^{-1} = dia 1 1 1 30 10 200 30 10 1

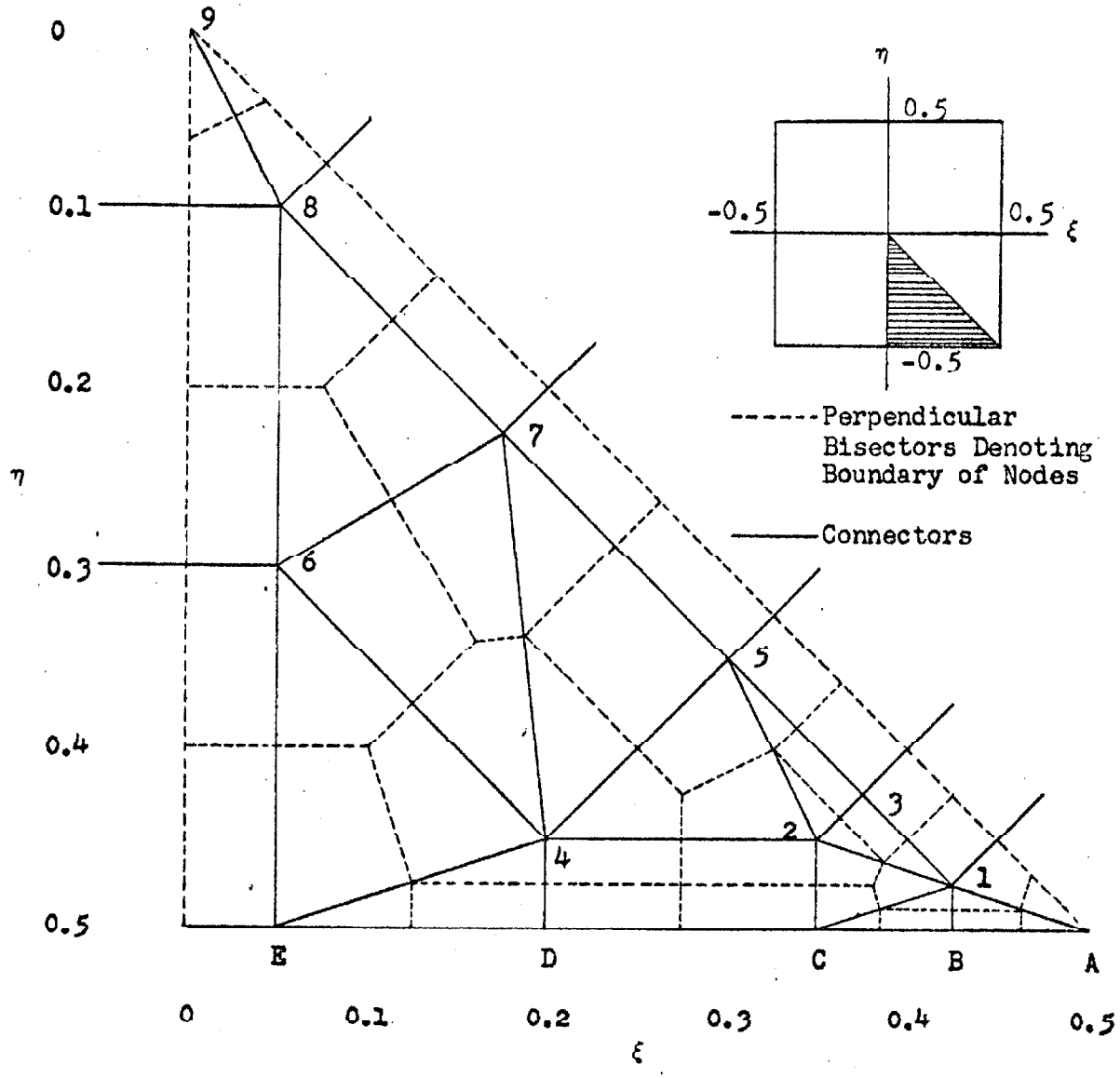


Figure III-1. Layout of nodes on square cross-section. Nodes 1-9, interior points where temperature is a known function of time. Nodes A-E, temperature nodes in boundary vector.

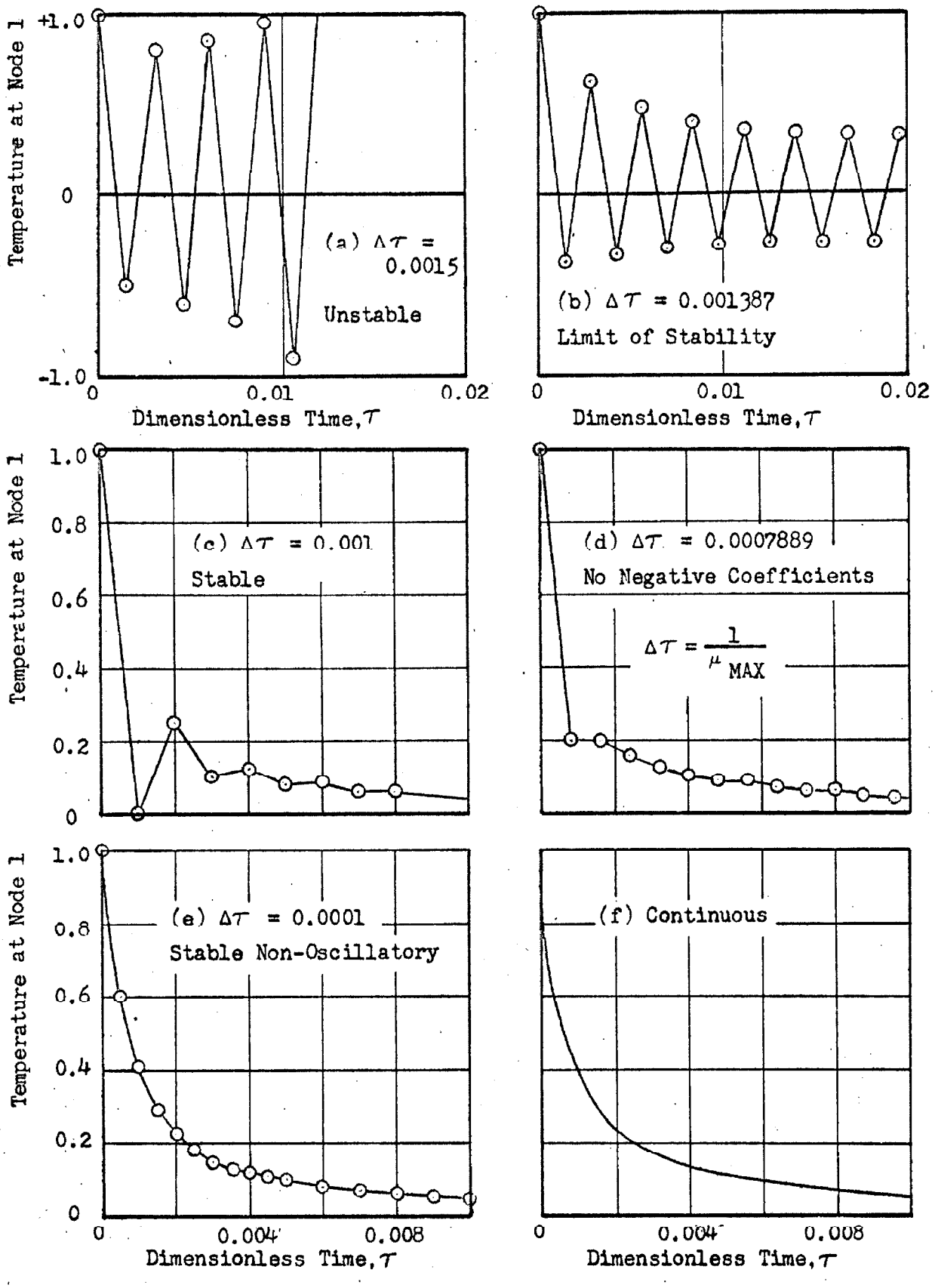


Figure III-2. Behavior of approximate solutions, $\gamma = 0$.

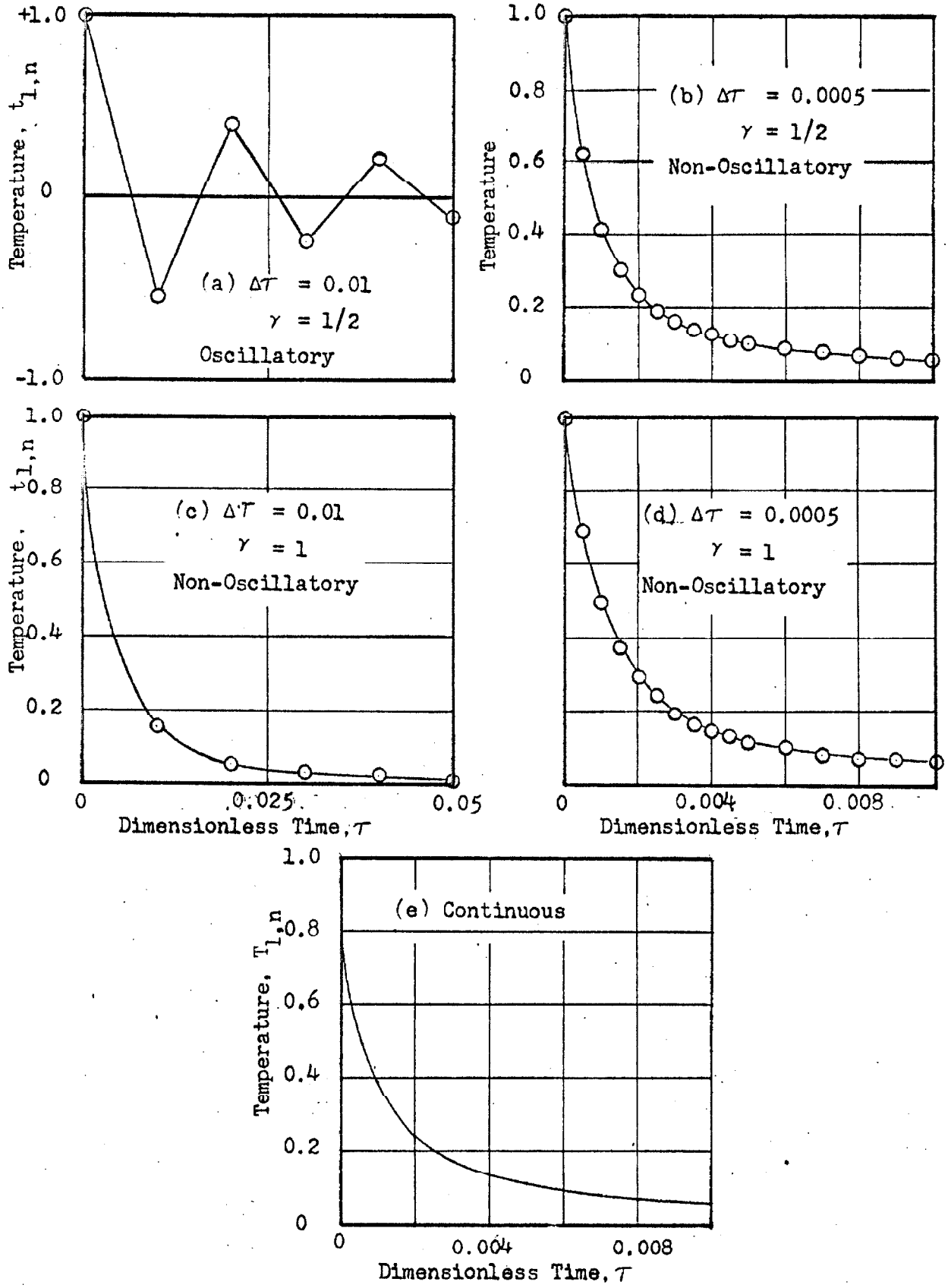
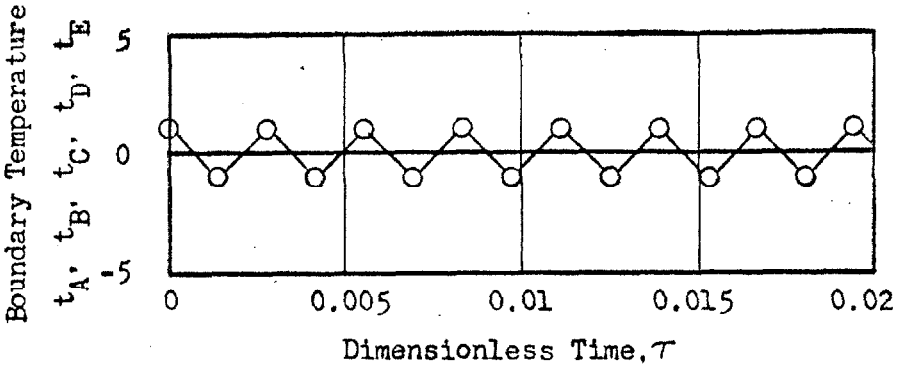


Figure III-3. Behavior of approximate solutions, $\gamma = 1/2$ and $\gamma = 1$.



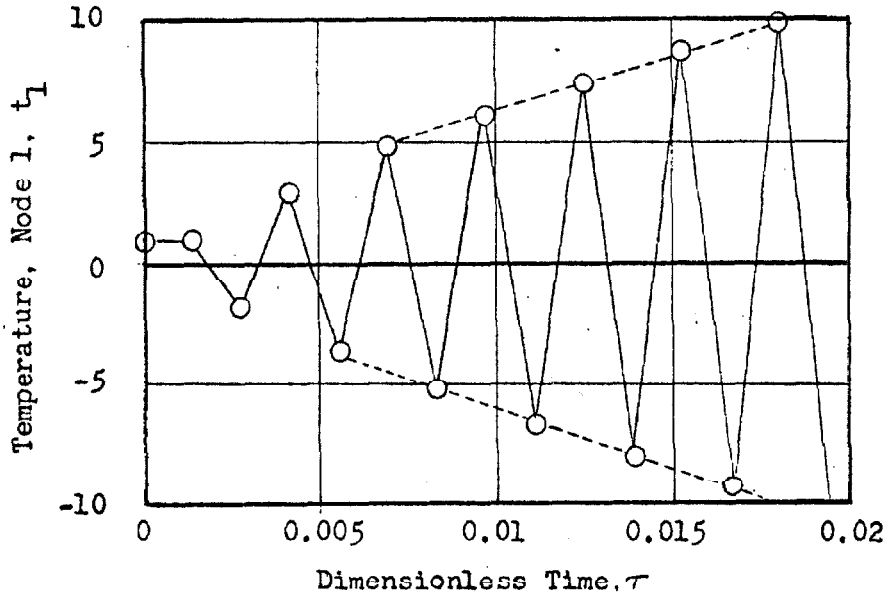


Figure III-4. Calculation at limit of stability, $\gamma = 0$, $\Delta\tau = 0.001387$, with oscillatory boundary conditions.

CHAPTER IV

SOLUTIONS OF THE ONE-DIMENSIONAL PARTIAL DIFFERENCE EQUATION FOR DIFFUSION IN CARTESIAN COORDINATES USING z-TRANSFORMS

(A) INTRODUCTION

In this chapter a z-transform method is presented for finding the complete analytic solution of linear partial difference equations of an initial-value type. This method is then used to find the solution of the partial difference equation II-42 which approximates the one-dimensional diffusion equation in Cartesian coordinates for several methods of approximating typical initial and boundary conditions. A comparison of the analytic difference and continuous solutions is then made.

The discussion in Chapter II has shown the solutions to the continuous partial differential equation and the difference formulation to be in the same form. A useful criterion has been derived in Chapter III for selecting the time differencing parameters so that the transient difference solution must have a behavior like that of the transient continuous solution. However, no mention was made of accuracy, or how well the difference eigenvectors, eigenvalues, initial vector, and particular

solution approximate the corresponding continuous eigenfunctions, eigenvalues, initial vector, and particular solution, or how well the total expression approximates the continuous expression. One way of making this comparison is to numerically step out an approximate solution for a problem where the continuous solution is known, and then compare the calculated continuous solution with the stepped-out approximate solution. Although this method of comparison does give results that show the accuracy for the specific problem with specific mesh size and other approximations and which reflect the accuracy and interrelationships between the various quantities mentioned above, it does not give any understanding or insight as to why the approximate method behaved as it did. Further generalizations made from such a comparison might be questionable. As a matter of fact, such a comparison of two possible graphical methods, for the case of specified boundary conditions, initially stimulated the search for a way to find the complete analytic difference solution or analytic expressions for all the quantities shown in equation II-43.

The specific problem that initiated this study was a comparison between graphical methods C and G for a specified boundary temperature. (Details of these methods are discussed later.) The results for a calculation of the temperature of the point nearest the boundary versus dimensionless time are shown in Figure IV-1 along with the continuous solution. The results for graphical method C oscillate so badly that

while the solution is stable it is useless; graphical method G on the other hand gives results which, although oscillatory, are useful. These results would not be anticipated as the norms of the Y/A matrix for both methods were equal and, further, a preliminary calculation of the eigenvalues for each method for two by two and three by three matrices indicated that both methods had identical eigenvalues. The complete solution not only explains this behavior, but it gives a much better understanding of the methods, shows ways of improving them, and allows us to make valid generalizations.

As demonstrated in Chapter II, for a regular mesh the approximate formulation for the one-dimensional diffusion equation in Cartesian coordinates can be shown as a partial difference equation, equation II-42, or by the matrix vector equation II-39, where the Y/A matrix is in equation II-41. In that matrix, the top and bottom rows are not shown. In the next section two general derivations for these equations are given which are approximations to the continuous boundary condition equation:

$$\pm h\left[T_f(\tau) - T(\xi_b \tau) \right] = \frac{k}{L} \frac{\partial T}{\partial \xi}(\xi_b \tau) \qquad \text{(IV-1)}$$

From the resulting general equations, several generalized methods are defined based on the selection of the formulation parameters that correspond to graphical methods A, C, G, and F (20). Although these define the calculation procedure, modifications are sometimes used

such as averaging, interpolating, and using slightly different initial temperatures and/or fluid temperatures, other than those directly suggested by the continuous problem. Specification of the calculation procedure and the modifications to the boundary and initial temperature functions then complete the difference formulation as a partial difference equation, subject to two boundary conditions and an initial distribution. Its solution is unique.

Next the z-transforms are defined from Laplace transforms and their important properties derived. Several related example problems are solved for approximate method G. These solutions show that the z-transform procedure is exactly analogous to the Laplace transform procedure for partial differential equations. The partial difference equation and its boundary equations are transformed. The resulting boundary-value ordinary difference equation is solved. The transformed temperature is then inverted to find the complete analytic expression for the difference solution which fits the initial and boundary conditions and includes the particular solution. Further, this solution is in a form that can be directly compared with the continuous solution.

This method gives the complete solution directly, in contrast to the classical methods of direct attack on matrix eigenvalue-eigenvector problems and the separation of variables or substitution technique suggested by von Neumann, as quoted in references (2, 21). The analytic solution for matrix eigenvalues is possible using the work of

Wolstonholne as quoted in Muir (22) and extended by Rutherford (23, 24). This is based on simple yet sophisticated use of a determinantal equation like equation II-46. A difficult analysis is required to determine the eigenvectors. Further, it is only applicable when the Y/A matrix is symmetric, so method G could not be studied in this way. The results have been used by Todd (4) for an approximation to the problem where the boundary temperature is specified. The technique of von Neumann is directly analogous to the separation of variables technique for solution of partial differential equations, and this method is usually used to derive the stability criteria for cases where a boundary temperature is specified (10, 21, 17, 15, 2, 18). The eigenvectors and eigenvalues are obtained directly upon fitting the boundary equations.

Although both the separation of variables technique and the matrix methods could be used for cases of finite non-zero heat-transfer coefficient, no references have been found where this has been done. Also, after the eigenvectors and eigenvalues have been found, to obtain the complete solution the particular solution must be determined and the coefficients required to make the solution fit the initial conditions must be found. This determination of the initial vector requires finding and using the orthogonal relationships of the eigenvectors, equations II-49 and II-51, and the analytic evaluation of several finite sums, essentially a finite harmonic or Fourier analysis. Only for one

approximation in the case of the infinite heat-transfer coefficient is this complete solution reported (2). These finite Fourier coefficients or the initial vector have an important effect on the accuracy of an approximate solution.

The last section of this chapter presents a comparison of the analytic expression of the solutions for the difference methods mentioned for several problems with the continuous solution of these problems (each of the defined methods is not necessarily derived or compared for each problem). These problems for the comparison are usually taken with a zero initial condition for convenience and they are summarized in Table IV-9.

(B) DIFFERENCE FORMULATION

In the formulation of the difference equations in Chapter II, the equations for interior points were fully developed from MacNeal's rules for both regular and irregular distributions of temperature points. No mention was made specifically (except by implication) of how the boundary conductances for cases of finite heat-transfer coefficient would be calculated or what assumptions could be made about the heat capacity of the boundary points. However, Longwell (20) has used a method for finite heat-transfer coefficient in association with his numerical solutions using an asymmetrical network. Also, he and his co-workers have pointed out that several different approximations to the boundary conditions may be used for graphical solutions for one-dimensional problems which give different results (20).

Here, the difference equations which approximate boundary conditions are derived in a general fashion for the one-dimensional problem for a node on the surface or $\Delta\xi/2$ inside the surface. These are the equations that would give the top and bottom rows in the Y/A matrix in equation II-41. For the point locations shown in Figures IV-2 and IV-3, the derivations are made in such a way that generalizations of the more accurate graphical methods mentioned in the references are obtained. This then completes the Y/A matrix for the one-dimensional problem.

If the initial temperature distribution, $t_{m,0}$, and the fluid temperature, $t_{f,n}$, are found directly from the corresponding continuous conditions, the method is specified. However, for more accuracy the initial temperature distribution and/or the fluid temperature function to be used in the approximate calculation are often modified slightly from the continuous conditions. Averaging or interpolation of the approximate results is also used sometimes. Thus, to completely specify a method, not only the grid and calculation equations must be specified, but also whether and how the initial temperature distribution and fluid temperature functions are to be modified, and if interpolation or averaging is to be used. In this study, the general equations are derived for each of the two meshes first. From these derivations certain logical assumptions follow concerning the initial conditions. Next, averaged methods are defined which can be applied to the methods. These averaged methods are shown to give the identical results as interpolation in some cases, and as certain modifications to the initial distribution and/or the fluid temperature function.

1. Continuous Problem and Initial Condition

The problem to be approximated in one dimension is

$$\frac{\partial T}{\partial \tau} = \frac{\partial^2 T}{\partial \xi^2} \qquad 0 < \xi < 1 \qquad \tau > 0 \qquad \text{(IV-2)}$$

with the known initial condition

$$T = T(\xi, 0) \qquad 0 < \xi < 1 \qquad \tau = 0 \qquad \text{(IV-3)}$$

and the two boundary conditions found by energy balances on the surface

$$H_0 \left[T_{f_0}(\tau) - T(0,\tau) \right] = -\frac{\partial T}{\partial \xi}(0,\tau) \qquad \xi = 0 \qquad \tau \geq 0 \qquad \text{(IV-4)}$$

$$H_L \left[T_{f_L}(\tau) - T(1,\tau) \right] = +\frac{\partial T}{\partial \xi}(1,\tau) \qquad \xi = 1 \qquad \tau \geq 0 \qquad \text{(IV-5)}$$

$$0 \leq H \leq \infty$$

where $H = hL/k$ = dimensionless heat-transfer coefficient

$T_f(\tau)$ = known function of time

The above initial condition, as specified, might not be known at the boundaries and in using an approximate solution with a point on the boundary, one must use a temperature for the boundary point to start the calculation. For this reason a brief discussion of the initial distribution follows. First, the exact continuous solution evaluated at time zero fits the boundary conditions, equations IV-4 and IV-5, evaluated at time zero. This means that, for the case of an infinite heat-transfer coefficient or specified boundary temperature initially, the boundary temperature does not equal the initial distribution found by extending equation IV-3 to the boundary. Two initial distributions can then be defined which apply for all ξ, including the boundaries. The first is designated $T(\xi, 0+)$ which results from the Fourier series for the continuous solution evaluated at a time of zero. It can be considered to be

the temperature distribution the instant after the boundary conditions are imposed. In most cases, this is known even if the exact solution is not. If h is less than infinity, then no discontinuity occurs at the boundary; this can be seen from Laplace transforms. If h is infinity, then the boundary temperature is equal to the fluid temperature. In some cases, either because of the method being used, or because $T(\xi, 0+)$ is not obvious, a second distribution is designated as $T(\xi, 0-)$ and is the distribution for the interior extended to the boundary. It can be considered the temperature distribution the instant before the boundary conditions are applied. For a finite heat-transfer coefficient it is the same as $T(\xi, 0+)$; for infinite heat-transfer coefficient initially the surface temperature would not be the fluid temperature, but would be the temperature of the material an infinitesimal distance inside the surface.

2. Meshes

The regular difference meshes that may be used for the one-dimensional diffusion equation are in Figures IV-2 and IV-3. A point which has two points as neighbors located $\pm\Delta\xi$ from it is called an interior point. The other points, which are located on or adjacent to a boundary with a conductance to a known fluid temperature, or with a known heat flux input, are called adjacent nodes. The main difference between the four meshes is that, in some cases, an adjacent node is

located on the boundary while in other cases it is located at a distance of $\Delta\xi/2$ from the boundary. Only the meshes with both boundary points located on the boundary or at a distance $\Delta\xi/2$ inside, as shown in Figure IV-2, are studied here.

The partial difference equation for diffusion in one dimension for the mth interior point as derived from Taylor expansions or Mac-Neal's rules is

$$\frac{t_{m,n+1} - t_{m,n}}{\Delta\tau} =$$

$$\frac{1}{(\Delta\xi)^2}\left[\gamma\left\{t_{m-1,n+1} - 2t_{m,n+1} + t_{m+1,n+1}\right\} + (1-\gamma)\left\{t_{m-1,n} - 2t_{m,n} + t_{m+1,n}\right\}\right]$$

$$1 \le m \le S-1 \qquad \text{(IV-6)}$$

Rearranging this equation gives equation II-42 and defines all but two rows of the Y/A matrix. In mesh $\Delta\xi$, Figure IV-2, this equation is written for $m = 1, 2, \ldots, S$; for mesh $\Delta\xi/2$ in Figure IV-2, it is written for $m = 3/2, 5/2, \ldots, S-3/2$. It serves to define all but the upper and lower rows of the Y/A matrix in equation II-33, where the two equations for the adjacent points are not given.

To derive the boundary equations, an energy balance at an adjacent point is solved for the temperature at that point. We will do this for both the $\Delta\xi$ and $\Delta\xi/2$ type adjacent points, and for the left adjacent point ($\xi = 0$ or $\xi = 1/2S$); the result can easily be applied to the opposite end. The derivations made are sufficiently general to obtain

a generalization of the graphical methods, including Longwell's backward difference type approximation (graphical method A), in addition to a general implicit formulation.

3. Mesh $\Delta\xi$

First consider mesh $\Delta\xi$ and the point 0. The energy in from a fluid is

$$h(t_f - t_0) = \dot{q}_{in} \tag{IV-7}$$

and out is

$$\frac{k(t_0 - t_1)}{L\,\Delta\xi} = \dot{q}_{out} \tag{IV-8}$$

Assuming an area associated with this point as $(\Delta\xi/2)$ we may write the accumulation term as

$$\sigma C_p \frac{L\,\Delta\xi}{2}\left[\frac{t_{0,n+1} - t_{0,n}}{L^2\Delta\tau/K}\right] = \text{volume rate of accumulation} \tag{IV-9}$$

Now the problem again occurs as to when to take the energy in, $\dot{q}_{in}$, and out, $\dot{q}_{out}$. Two weighting factors are defined, β and φ, such that β, $(1-\beta)$, φ, and $(1-\varphi)$ lie between zero and one. The weighting β is given to the energy in from the surrounding fluid at time $(n+1)$, and $(1-\beta)$ to that quantity at time n. Analogously, φ is the weighting given to the energy out from the interior node at time $(n+1)$, and $(1-\varphi)$ at time n. The energy balance then becomes (in dimensionless variables)

$$\frac{\sigma C_p L \Delta\xi}{2} \quad \frac{t_{0,n+1} - t_{0,n}}{\sigma C_p L^2 \Delta\tau / k} = h\left[(t_{f,n+1} - t_{0,n+1})\beta + (t_{f,n} - t_{0,n})(1-\beta)\right] \tag{IV-10}$$

$$- \frac{k}{L\Delta\xi}\left[(t_{0,n+1} - t_{1,n+1})\varphi + (t_{0,n} - t_{1,n})(1-\varphi)\right.$$

(If h is not constant it should be included in the β weighting.)

Combining and rearranging so that unknown temperatures at (n+1) are on the left side of the equation, the calculation equation for the adjacent point is:

$$t_{0,n+1}\left[1 + 2r\left(\frac{H\beta}{S} + \varphi\right)\right] - 2r\varphi t_{1,n+1} = t_{0,n}\left[1 - 2r\left\{\frac{H}{S}(1-\beta) + 1 - \varphi\right\}\right]$$

$$+ t_{1,n}\left[2r(1-\varphi)\right] \tag{IV-11}$$

$$+ \frac{2rH}{S}\left[t_{f,n+1}\beta + (1-\beta)t_{f,n}\right]$$

where $\Delta\xi = 1/S$

and $r = \Delta\tau/(\Delta\xi)^2$

Using a similar boundary equation for $t_{S,n+1}$ then specifies the calculation procedure, but does not specify the initial distribution or how the boundary temperature changes with time. This equation is convenient for all H less than infinity. By taking limits as H goes to infinity, which means neglecting the capacity of the surface node, the boundary condition equation for a specified boundary temperature is:

$$(t_{0,n+1} - t_{f,n+1})\beta = (1-\beta)(t_{f,n} - t_{0,n}) \qquad n \geq 0 \qquad \text{(IV-12)}$$

If β is not one and the above expression is to hold for all n, it is necessary in order to avoid oscillations that

$$t_{f,0} = t_{0,0} \qquad \text{(IV-13)}$$

otherwise a root q_j of $(\beta-1)/\beta$ would be added to the transient solution. For $\beta = 1$ the initial conditions $t_{0,0}$ need not be equal to $t_{f,0}$ as equation IV-10 gives

$$t_{0,n+1} = t_{f,n+1} \qquad n \geq 0 \qquad \text{(IV-14)}$$

In this case $T(m\Delta\xi, 0-)$ is used as the difference initial condition.

Two specific methods of the many defined by equation IV-10, and which are to be studied, are generalized methods A and G. These methods are defined only for mesh $\Delta\xi$; thus the equations for both boundary equations are defined by equation IV-11 where a different H_S and $t_{f_S,n}$ and different subscripts define the right hand ($\xi = 1$) boundary condition.

Method G. First method G is defined and derived. For this method, β and φ in the boundary equations are set equal to γ in the partial difference equation IV-6. This gives for the left boundary:

$$t_{0,n+1}\left[1 + 2r\gamma\left(\frac{H}{S} + 1\right)\right] - 2r\gamma t_{1,n+1} = t_{0,n}\left[1 - 2r(1-\gamma)\left(\frac{H}{S} + 1\right)\right] + 2rt_{1,n} + \frac{2rH}{S}\left[t_{f,n+1}\gamma + t_{f,n}(1-\gamma)\right] \qquad \text{(IV-15)}$$

This is a generalization of the implicit case of Longwell's (9) approximation for finite heat-transfer coefficient and it gives a Y/A matrix with all the properties described in Chapter II. However, this method suffers from a serious defect. The sufficient stability criterion from an examination of the Y/A matrix for all H less than infinity is

$$r(1 - 2r\gamma)M < 2 \tag{IV-16}$$

$$\text{where } M = \min \left(\left\| \frac{Y}{A} \right\|_{I}, \left\| \frac{Y}{A} \right\|_{II} \right) \tag{IV-17}$$

$$\left\| \frac{Y}{A} \right\|_{I} = 2\left(2 + \frac{H_{MAX}}{S}\right)$$

$$\text{and} \tag{IV-18}$$

$$\left\| \frac{Y}{A} \right\|_{II} = \text{MAX}\left(5 \text{ or } 3 + \frac{H_{MAX}}{S}\right)$$

This means that for an explicit method and a non-zero H a more severe restriction than r less than one-half must be used. One way to reduce this limit would be to neglect the capacity of the surface node $LC_p\sigma\Delta\xi/2$; (this is essentially what is done when H is increased to infinity) and the temperature at the surface is then for any time increment an explicit function of $t_{f,n}$ and $t_{1,n}$ at the same value of n.

$$t_{0,n} = \frac{\frac{H}{S} t_{f,n} + t_{1,n}}{\frac{H}{S} + 1} \tag{IV-19}$$

This equation can then be used to eliminate $t_{0,n}$ from the partial difference equation written for an m of 1. Thus, $t_{0,n}$ need not be computed; the Y/A matrix is reduced by one row and one column and the largest eigenvalue of Y/A is eliminated. However, as method G is defined, only for the case of a specified boundary temperature (infinite h) is the capacity of node 0 neglected. This method normally uses $T(m\Delta\xi, 0+)$ for its initial condition and $T_f(n\Delta\tau)$ for the difference fluid temperature. The details of this method are summarized in Table IV-1.

Method A. To avoid instability with a large, but finite, heat-transfer coefficient, Longwell (20) has suggested using a backward difference approximation for the energy into the solid at the surface, for a graphical method. A generalization of this principle is seen from the above equation IV-11; for if φ is zero an explicit equation for $t_{0,n+1}$ results.

$$t_{0,n+1} = \frac{t_{0,n}\left[1-2r\left\{\frac{H}{S}(1-\beta)+1\right\}\right] + 2rt_{1,n} + \frac{2rH}{S}\left[t_{f,n+1}\,\beta + t_{f,n}(1-\beta)\right]}{\left[1 + 2r\,\frac{H\beta}{S}\right]} \qquad \text{(IV-20)}$$

Although using an equation of the form of equation IV-20 for each boundary and using a non-zero γ in the partial difference equation is in principle possible, the main advantage of using the backward difference is to have a stable explicit method where the heat capacity of the

surface node is not neglected. In this explicit case, the Y/A matrix is not defined directly, but since $[I - Y/A]$ is defined as the matrix that operates on t_n, an equivalent Y/A may be found by difference. The first few rows are:

$$(\Delta\xi)^2 \frac{Y}{A} = - \begin{vmatrix} \frac{2(\frac{H}{S} + 1)}{1 + 2r\frac{H\beta}{S}} & - \frac{2}{1 + 2r\frac{H\beta}{S}} & 0 \ \ldots \\ -1 & 2 & -1 \ \ldots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{vmatrix} \qquad \text{(IV-21)}$$

From the norms, the sufficient stability criteria are

$$r\left[\frac{H}{S}(1 - 2\beta) + 2\right] \leq 1 \qquad \text{(IV-22)}$$

and $$r \leq \tfrac{1}{2} \qquad \text{(IV-23)}$$

If β is equal to or greater than $\frac{1}{2}$ the second inequality is the more restrictive, thus avoiding the much more restrictive stability requirement in equation IV-16. Although the Y/A matrix must be found by difference for the above family of methods where φ and γ are zero, but not equal to β which is not zero, the resulting Y/A matrix retains all the properties mentioned in Chapter II, except that it is now also a function of the time increment $\Delta\tau$ or r. This is in contrast to method

G where the Y/A matrix has all the properties described in Chapter II and is not a function of the time differencing parameters.

Longwell (19) has shown the advantageousness of a graphical method A based on equation IV-20 with a β of 1 and an r of $\frac{1}{2}$. In this study, a generalized method A is defined on the same equation with a β of 1 but for all r that give stable solutions. The boundary equation is

$$t_{0,n+1} = \frac{t_{0,n}(1-2r) + 2r\, t_{1,n} + \frac{2r\, H_0}{S}\, t_{f_0,n+1}}{1 + 2r\, \frac{H_0}{S}} \qquad \text{(IV-24)}$$

Note that if H is not constant its value should be taken at the end of the time increment. This equation is used to calculate the boundary temperature for all finite H. If H is zero, it reduces to the same equation as explicit method G. For infinite H, taking limits gives directly,

$$t_{0,n+1} = t_{f,n+1} \qquad n \geq 0 \qquad \text{(IV-25)}$$

The difference initial distribution for this method is taken as $T(m\Delta\xi, 0-)$ for all n including the points on the surface. Thus for infinite h the surface temperature point does not change until an n of one is reached. Then this method results in delaying the change in $t_{0,n}$ only for the first time increment. For problem II where one surface temperature is constant, the other adiabatic, an implicit method A is defined, which

gives the same results as implicit G, but delayed one time increment. However, for all other problems method A is an explicit method. Details of this method are summarized in Table IV-2.

4. Mesh $\Delta\xi/2$

Next consider mesh $\Delta\xi/2$ where there are no points on the boundary, but a boundary point is located a distance of $\Delta\xi/2$ in from the surface. An energy balance is made on the left adjacent node, $\frac{1}{2}$. The energy rates in and out are:

$$\dot{q}_{in} = \frac{t_f - t_{\frac{1}{2}}}{\frac{1}{h} + \frac{\Delta x/2}{k}} = \frac{h(t_f - t_{\frac{1}{2}})}{1 + \frac{H}{2S}} \qquad \text{(IV-26)}$$

$$\dot{q}_{out} = \frac{k(t_{\frac{1}{2}} - t_{3/2})}{\Delta x} = \frac{k(t_{\frac{1}{2}} - t_{3/2})}{L\Delta\xi} \qquad \text{(IV-27)}$$

and the rate of accumulation is:

$$\text{rate of accumulation} = L\Delta\xi\sigma C_p \left[\frac{t_{\frac{1}{2},n+1} - t_{\frac{1}{2},n}}{L^2\sigma C_p \Delta\tau/k} \right] \qquad \text{(IV-28)}$$

Again using β and $(1-\beta)$ for the energy from the fluid and φ and $(1-\varphi)$ for the energy to the adjacent interior point the energy balance gives

$$t_{\frac{1}{2},n+1}\left[1+r\,\frac{\beta\,H/S}{1+H/2S}\right]-r\varphi t_{3/2,n+1}=t_{\frac{1}{2},n}\left[1-r\left\{\frac{(1-\beta)H/S}{1+H/2S}+1-\varphi\right\}\right]$$

$$+r(1-\varphi)t_{3/2,n}+\left[\frac{r\,H/S}{1+H/2S}\right]\left[(1-\beta)t_{f,n}+\beta t_{f,n+1}\right] \qquad \text{(IV-29)}$$

This equation is suitable for use for all H less than infinity. Taking limits as H goes to infinity, which does not require neglecting any heat capacity, gives

$$t_{\frac{1}{2},n+1}[1+r(2\beta+\varphi)]+r\varphi t_{3/2,n+1}=t_{\frac{1}{2},n}[1-r\{2(1-\beta)+1-\varphi\}]$$

$$r(1-\varphi)t_{3/2,n}+2r[(1-\beta)t_{f,n}+\beta t_{f,n+1}] \qquad \text{(IV-30)}$$

These equations and the derivations are somewhat unique in that no temperature point is located on the boundary. If it is necessary to compute a surface temperature, one may do so using thermal resistances, and no thermal capacity associated with the surface point.

$$t^{*}_{n}=\frac{t_{\frac{1}{2},n}+\frac{H}{2S}t_{f,n}}{1+\frac{H}{2S}} \qquad \text{(IV-31)}$$

where t^{*}_{n} = surface temperature at n ($t_{0,n}$ is not used, as m is restricted to full increments; also, the difference solution does not give this temperature by letting m = 0 for this mesh)

or, rearranged,

$$H\left[t_{f,n} - t^{*}_{n}\right] = \frac{(t^{*}_{n} - t_{\frac{1}{2},n})}{\Delta\xi/2} \qquad \text{(IV-32)}$$

which is a direct difference approximation of the continuous boundary condition, equation IV-4. In several other ways, mesh $\Delta\xi/2$ leads to approximations that appear to be closer mathematically to the continuous problem than does mesh $\Delta\xi$. First of all, as each node has the same thermal capacity, the Y/A matrix is symmetric for all values of H when β and φ are taken equal to γ. In this case, its eigenvectors are orthogonal with respect to a weighting factor of one as the eigenfunctions of the continuous solution are orthogonal with respect to a weighting factor of one. Moreover, even for infinite h no heat capacity is neglected. Further, as no point is selected on the surface, the initial condition need not be specified on the surface, as is also the case for the continuous problem.

Method C. First, method C is defined which is analogous to method G in that β and φ are taken equal to γ in the partial difference equation. The boundary equations for finite and infinite H are, respectively,

$$t_{\frac{1}{2},n+1}\left[1+r\gamma\left(\frac{1+3H/2S}{1+H/2S}\right)\right] - r\gamma t_{3/2,n+1} = t_{\frac{1}{2},n}\left[1-r(1-\gamma)\left(\frac{1+3H/2S}{1+H/2S}\right)\right] \qquad \text{(IV-33)}$$

$$+ r(1-\gamma)t_{3/2,n} + \left[\frac{rH/S}{1+H/2S}\right]\left[(1-\gamma)t_{f,n} + t_{f,n+1}\right]$$

$$t_{\frac{1}{2},n+1}(1+3r\gamma) - r\gamma t_{\frac{1}{2},n+1} = t_{3/2,n}[1-3r(1-\gamma)]$$
$$+r(1-\gamma)t_{\frac{1}{2},n} + \left[(1-\gamma)t_{f,n} + \gamma t_{f,n+1}\right] \qquad \text{(IV-34)}$$

The Y/A matrix for this method has all the properties described in Chapter II and from its norms method C is stable for

$$r(1 - 2\gamma) \leq \tfrac{1}{2} \qquad \text{(IV-35)}$$

Graphical method C reduces to Schmidt's graphical method, as given in reference (19), except that his calculation for an n of zero does not follow equation IV-32 but does follow a construction for finite h that gives a t_0 that corresponds to the continuous surface temperature initially (note equation IV-31 does change the surface temperature for an n of zero). Also, although the graphical method is technically stable for all H, it is not used for cases where H/S is greater than 2, because of large oscillations. Details of method C are summarized in Table IV-3.

Method F. In order to avoid these oscillations for infinite h Schmidt proposed an alternate construction, called method F in the Longwell manuscript (20), for infinite h only. If in equation IV-31 we let $\beta = \frac{1}{2}$, $\varphi = 0$, $r = \frac{1}{2}$, and $\gamma = 0$, for the case of a constant fluid temperature the equation reduces to Schmidt's construction. Thus a generalized explicit method F may be defined by making $\beta = \frac{1}{2}$ and φ and γ zero. Details of this generalized method are in Table IV-4; however, here only the graphical method for infinite h is studied.

5. Other Methods

Although these are the basic four methods studied here, two other methods probably should be investigated. These are an explicit method for mesh $\Delta\xi$ with a β of $\frac{1}{2}$ and an explicit method for mesh $\Delta\xi/2$ with a β of 1. Also combination methods for the meshes shown in Figure IV-3 could be studied.

6. Graphical Solution

The system of difference equations given by the partial difference equation IV-6 and the boundary equations for one of the methods, G, A, C, or F, can usually be solved by a graphical construction when the differencing parameters are selected by

$$r = S^2\Delta\tau = \tfrac{1}{2} \qquad \text{(IV-36)}$$

$$\gamma = 0 \qquad \text{(IV-37)}$$

The simple graphical constructions for both the interior points and the boundary equations are described in the study by Longwell (20). However, simple graphical constructions that solve the boundary equations for methods G and F for H's that are neither zero nor infinite are not available. (Also method G is unstable for these H's and the combination of differencing parameters given in equations IV-36 and IV-37.) However, the selection of the r and γ according to equations IV-36 and IV-37 is important, even if the solution is calculated numerically rather than graphically. This is true because under these conditions

the number of non-zero multiplications required per step is S compared to 2S for any other explicit calculation ($\gamma = 0$) or 7S for an implicit calculation ($\gamma \neq 0$). This comparison of calculation counts assumes that advantage is taken of the fact that the off-diagonal elements in each row of the $[I-(1-\gamma)\Delta\tau(Y/A)]$ matrix are equal, and that use is made of the elimination method of solving the equations for an implicit calculation, as described in Chapter III. If a general vector-matrix multiplication is used to calculate $\left[I-(1-\gamma)\Delta\tau\left(\frac{Y}{A}\right)\right]t_n$, S multiplications must be added to each number giving 2S for graphical, 3S for explicit, and 8S for implicit calculations. Because of the ease of stepping out the solution either graphically or numerically a special designation is used for the solution when γ is zero and r is $\frac{1}{2}$. Since either the graphical or numerical stepping out of the solution in the absence of errors gives the same approximation, in this study selection of the parameters according to equations IV-36 and IV-37 for any method is called a graphical solution or graphical method. If the number of points to be used is greater than about eight, the graphical construction to step out the solution becomes inaccurate and the approximate solution must be stepped out numerically.

The general definition of graphical methods together with specific notes on the graphical solutions associated with each method is in Table IV-5. For more details concerning construction and a comparison of numerical results for specific problems the Longwell (20) reference

should be consulted.

7. Averaged Methods

Although the equations presented for each method, together with a straightforward use of the initial temperature distribution and the forcing functions, serve to completely define an approximate method, several modifications can be applied to these methods without changing the equations or matrix elements. In this section, the technique of averaging the approximate solution as mentioned in Chapter III is more fully defined and discussed. The important relationship between an averaged method and an unaveraged method is developed. Also brief sections are included on interpolation as suggested by Longwell (20) and on the modification of the fluid temperature function for step changes in fluid temperature. In these sections these techniques are defined and they are shown to be equivalent to an averaging procedure. Then, based on these equivalences, a specific averaged method is associated with the equations for methods G, A, C, and F.

In an averaged method the approximate solution of one of the defined methods is stepped out according to equation II-39 in the usual manner. The temperature vector t_n found for the beginning of the $(n+1)$ time increment and t_{n+1} for the end of that time increment are averaged and assumed to apply at the beginning, at the end, or at some time within the $(n+1)$ time increment. The defining equation for the averaged temperature vector $t_{n+k\ Ave}$ assumed to apply at time $(n+k)\Delta\tau$

is:

$$t_{n+k\,Ave} = \tfrac{1}{2}\,(t_n + t_{n+1}) \qquad \text{(IV-38)}$$

where t_n and t_{n+1} are the unaveraged temperature vectors found by successive uses of equation II-39. The parameter k is analogous to the weighting factor γ and tells at what fraction of the time increment the average applies; it is taken between zero and one inclusive:

$$0 \le k \le 1 \qquad \text{(IV-39)}$$

In this study k is usually 0 or 1, although a brief study of a k of $\frac{1}{2}$ is also made. For a k of zero, the average is called a forward average because it is analogous to the forward difference calculation when the weighting factor γ is zero as both involve t_{n+1} in the calculation at time n.

Forward Average, k = 0:

$$t_{n\,Ave} = \tfrac{1}{2}\,(t_n + t_{n+1}) \qquad n \ge 0 \qquad \text{(IV-40)}$$

Likewise the central and backward averages are defined by

Central Average, $k = \frac{1}{2}$:

$$t_{n+\frac{1}{2}} = \tfrac{1}{2}\,(t_n + t_{n+1}) \qquad n \ge 0 \qquad \text{(IV-41)}$$

Backward Average, k = 1:

$$t_{n+1\,Ave} = \tfrac{1}{2}\,(t_n + t_{n+1}) \qquad n \ge 0 \qquad \text{(IV-42)}$$

The backward average can be written for time n for convenience in comparing with the forward average as:

Backward Average, k = 1:

$$t_{n\,Ave} = \tfrac{1}{2}\left(t_{n-1} + t_n\right) \qquad n \geq 1 \qquad \text{(IV-43)}$$

Of these averages only the forward average can be used to calculate a temperature vector at a time of zero, and hence only the forward average (or only when k is zero) can change the initial temperature distribution.

Although the averaging technique is defined as averaging the approximate solution vectors for successive times n and (n+1) as calculated directly from the stepping out of equation II-39, identical results can be obtained by using an averaged temperature vector $t_{n+k\,Ave}$ and an averaged boundary temperature vector $t_{B\,Ave}$ in equation II-39 to step out the solution. This can be proved by writing equation II-39 for t_{n+1} in terms of t_n and t_{n+2} in terms of t_{n+1} and substituting these quantities in equation IV-38 which defines an averaged method. This gives, using the definition of $t_{n+k\,Ave}$:

$$t_{n+k+1\,Ave} =$$

$$\left[I-\Delta\tau\frac{Y}{A}\right]^{-1}\left(\left[I+(1-\gamma)\Delta\tau\frac{Y}{A}\right]t_{n+k\,Ave}+\Delta\tau\frac{Y_B}{A}\left[\gamma t_{B\,n+2}+t_{B\,n+1}+(1-\gamma)t_{B\,n}\right]\right) \qquad \text{(IV-44)}$$

Thus, to calculate an average solution an average need only be computed once, and then used successively in equation IV-44, or the regular solution stepped out and averages calculated only for the points at the times where necessary.

The solution for an averaged method can be found by substituting in the defining equation IV-38 the solution as found from equation II-43 for the unaveraged method for times n and (n+1). For a problem which has a true steady-state solution we obtain

$$t_{n+k\ Ave} = t_P - C\ Q^n\left[\tfrac{1}{2}(Q+I)g\right] \qquad \text{(IV-45)}$$

where t_P, C, Q, and g are the identical quantities that are in the solution for the unaveraged method. By comparing equation IV-45 and equation II-43, the effect of averaging is to replace the initial vector g for the unaveraged method with the vector $\left[\tfrac{1}{2}(Q+I)g\right]$. If an eigenvalue q_j occurs that is close to (but greater than) -1 the component of $\left[\tfrac{1}{2}(Q+I)g\right]$ in the averaged method is significantly smaller in absolute value than that in g (see equation III-32). Consequently, the major advantage of averaging, regardless of the time at which the averaged solution is applied, is that the amplitude of the oscillations is significantly reduced. Therefore averaging is particularely useful for graphical methods the solution for which contains a negative q_{min} close to -1. Indeed, for some graphical methods applied to certain problems a q_{min} of -1 exists, and averaging is necessary to obtain a stable solution.

Although the form of the solution in equation IV-45 is useful to show the effect of averaging on the amplitude of the oscillations, to compare an averaged solution with the continuous solution a more convenient form is

$$t_{n+k\ Ave} = t_P - C\,Q^{n+k}\left[\tfrac{1}{2}Q^{-k}(I+Q)g\right] \qquad \text{(IV-46)}$$

which is used to examine positive damping factors only. The Q^{n+k} matrix corresponds directly to the $e^{-\nu^2(n+k)\Delta\tau}$ matrix in the continuous solution, and the effect of using the averaged temperature vector $t_{n+k\ Ave}$ at time $(n+k)\Delta\tau$ is to change the initial vector to $\left[\tfrac{1}{2}Q^{-k}(I+Q)g\right]$. The jth component of this vector is $(q_j+1)g_j/2q_j^{\,k}$, which corresponds to a_j for the continuous solution and to g_j for the unaveraged approximate solution. For values of k of zero, or one, averaging changes components of the initial vector corresponding to q_j's of about 0.9 or larger from 2 to 5 per cent of the components in g_j; however, for the q_j's of about 0.6 to 0.7 the g_j's are significantly changed. When $k = \frac{1}{2}$ the components of the initial vector for the q_j larger than 0.9 are not significantly changed from those in the unaveraged vector g.

Interpolation. The interpolation technique was suggested by Longwell (20) for use with graphical methods to eliminate a certain type of oscillatory behavior. Some graphical methods including graphical method A, when applied to problems with a linear initial distribution and with either constant fluid temperature or adiabatic conditions, give a change in the temperature point only at alternate values of n. This gives a temperature-time graph that is a sequence of ramps and plateaus as in Figure IV-1(b). This condition is expressed, for example, by the equation for the temperature at a point as

$$t_{m,n-1} = t_{m,n} \neq t_{m,n+1} = t_{m,n+2} \qquad \text{(IV-47)}$$

For cases when this occurred it was suggested that the temperature at the end of a time increment during which no change occurred (at times n and (n+2) above) be replaced with a temperature linearly interpolated between those at the beginning of the previous time increment and the end of the next time increment. This gives for the above case at time n:

$$t_{m,n\,I} = \tfrac{1}{2}\left[t_{m,n-1} + t_{m,n+1}\right] \qquad \text{(IV-48)}$$

This then gives a sequence of ramps eliminating the plateaus.

Since $t_{m,n-1}$ and $t_{m,n}$ are equal the interpolated temperature is identical to a forward averaged temperature

$$t_{m,n\,I} = \tfrac{1}{2}\left[t_{m,n} + t_{m,n+1}\right] = t_{m,n\,Ave} \qquad k = 0 \qquad \text{(IV-49)}$$

Moreover, because $t_{m,n-1}$ and $t_{m,n}$ are equal and because interpolation does not change $t_{m,n-1}$, the forward average can be applied at the beginning of a plateau,

$$t_{m,n-1\,I} = t_{m,n-1} = \tfrac{1}{2}\left[t_{m,n-1} + t_{m,n}\right] = t_{m,n-1\,Ave} \qquad k = 0 \qquad \text{(IV-50)}$$

Consequently, if throughout the calculation each component of the temperature vector changes only at alternate time intervals, then using a forward average gives results equivalent to linear interpolation for those components which do not change.

Two points should be mentioned concerning the above discussion. First, although each component of the calculated temperature vector t_n changes only at alternate time increments, the vector t_n itself usually changes at each time increment. For example, in graphical method A applied to problem II, the odd-subscripted (m odd) components of t_n change only when n goes from odd to even and the even-subscripted (m even) components change only when n goes from even to odd; thus the vector t_n changes for each value of n, although some of its components remain constant for alternate values of n. The major conclusion is that the forward average is identical with interpolation for this case,

$$t_{n\,I} = \tfrac{1}{2}(t_n + t_{n+1}) = t_{n\,Ave} \qquad k = 0 \qquad \text{(IV-51)}$$

Secondly, a backward average applied to an approximate solution with a ramp-plateau temperature-time plot gives an approximation identical to using a linear interpolation for the temperature at the beginning of the plateau.

Modifications to Step Changes in Boundary Conditions. As noted in the Longwell paper (20), for problems with a constant initial temperature subjected to a constant different surface temperature, it has been suggested by many authors (6, 17) that a temperature half way between the initial temperature and the surface temperature can be used for the calculation from an n of zero to one. Then for all succeeding steps the actual surface temperature is used. For example, consider

a solid initially at zero temperature throughout, and at time zero. One surface temperature is held constant at unit temperature, the other boundary is insulated (problem II). The surface temperature for the continuous problem is:

$$T(0, n\Delta\tau) = 1 \qquad n \geq 0 \qquad \text{(IV-52)}$$

But instead of using in the difference calculation

$$t_{0,n} = 1 \qquad n \geq 0 \qquad \text{(IV-53)}$$

a ramp-step function defined by

$$t_{0,n} = \begin{cases} \frac{1}{2} & n = 0 \\ 1 & n \geq 1 \end{cases} \qquad \text{(IV-54)}$$

can be used to step out the solution. In the Longwell (20) work it was mentioned that this procedure for graphical method G gives the same results as interpolation for interpolated graphical method A for this problem. Hence we can conclude that for this case it is equivalent to an averaging technique.

This principle involved in modifying a constant boundary temperature can be generalized and applied to any problem where the fluid temperatures or the vector t_B are constant with time and where a boundary temperature vector $t_{B\ 0-}$ can be found that gives the initial temperature distribution as a steady-state solution. The last condition

can be expressed in difference form as either

$$Y\, t_0 = -Y_B\, t_{B\,0-} \tag{IV-55}$$

or

$$\left[I-\gamma\,\Delta\tau\,\frac{Y}{A}\right]t_0 = \left[I+(1-\gamma)\Delta\tau\,\frac{Y}{A}\right]t_0 + \Delta\tau\,\frac{Y_B}{A}\,t_{B\,0-} \tag{IV-56}$$

Since Y_B is a rectangular matrix the existence and uniqueness of $t_{B\,0-}$ is not guaranteed and depends upon the vector t_0. For a one-dimensional problem in Cartesian coordinates in a solid with uniform thermal properties $t_{B\,0-}$ exists only if the vector t_0 represents a linear distribution. Now the ramp-step procedure is to replace the vector $[(1-\gamma)t_{B\,n} + t_{B\,n+1}]$ that multiplies the matrix $\Delta\tau\, Y_B/A$ in equation II-39 with a ramp-step type function

$$(1-\gamma)t_{B\,n} + \gamma\, t_{B\,n+1} = \begin{cases} \frac{1}{2}(t_{B\,0-} + t_B) & n=0 \quad \text{(first step)} \\ t_B = \text{constant} & n \geq 1 \quad \text{(second and all other steps)} \end{cases} \tag{IV-57}$$

The equation for the first step becomes

$$t_{1,RS} = \left[I-\Delta\tau\,\frac{Y}{A}\right]^{-1}\left(\left[I+(1-\gamma)\Delta\tau\,\frac{Y}{A}\right]t_0 + \Delta\tau\,\frac{Y_B}{A}\left[\frac{t_{B\,0-} + t_B}{2}\right]\right) \tag{IV-58}$$

And using equation IV-55 to eliminate $t_{B\,0-}$ the equation is

$$t_{1,RS} = \left[I-\Delta\gamma\,\frac{Y}{A}\right]^{-1}\left(\left[I+(\tfrac{1}{2}-\gamma)\Delta\tau\,\frac{Y}{A}\right]t_0 + \Delta\tau\,\frac{Y_B}{A}\,\frac{t_B}{2}\right) \tag{IV-59}$$

This $t_{1,RS}$ is used to continue the stepping calculation. The averaged temperature vector $t_{0+k\ Ave}$ is found substituting t_1 from equation II-39 and t_0 into the definition giving:

$$t_{0+k\ Ave} = \tfrac{1}{2}\left\{\left[I-\gamma\Delta\tau\frac{Y}{A}\right]^{-1}\left(\left[I+(1-\gamma)\Delta\tau\frac{Y}{A}\right]t_0 + \Delta\tau\frac{Y_B}{A}t_B\right) + t_0\right\} \qquad \text{(IV-60)}$$

Combining the coefficients for t_0 and multiplying through by $\frac{1}{2}$ gives

$$t_{0+k\ Ave} = \left[I-\gamma\Delta\tau\frac{Y}{A}\right]^{-1}\left(\left[I+(\tfrac{1}{2}-\gamma)\Delta\tau\frac{Y}{A}\right]t_0 + \Delta\tau\frac{Y_B}{A}t_B\right) \qquad \text{(IV-61)}$$

which is the same expression as in equation IV-59 for $t_{1,rs}$; consequently,

$$t_{1,RS} = t_{0+k\ Ave} \qquad \text{(IV-62)}$$

Now, since using $t_{0+k\ Ave}$ to step out the rest of the solution gives an averaged solution according to equation IV-44, the same must be true when $t_{1,RS}$ is used. Further, if the averaged solution is assumed to apply at the end of the interval (k=1) then the backward average is identical to using the ramp-step modification for problems where the boundary temperature vector is constant and where the initial temperature distribution is a steady-state condition.

Conclusions and Averages to be Associated with Methods G, A, C, and F. The above discussion has shown that averaging the stepped-out solution for successive time increments greatly reduces the amplitude of the oscillations, but depending upon the time at which

the average is assumed to apply, the amplitude of the q_j's larger than 0.9 is not changed more than 5 per cent. Also, interpolation for approximate solutions with a ramp-plateau temperature-time plot and the replacement of a constant boundary vector with a ramp-step vector for problems with an initial condition, which fits a steady-state condition, are shown to be averaged methods. Since averaging can be applied to any problem or method it is a generalization of these two modifications.

Although the time at which an averaged solution applies is arbitrary, in order to be consistent with previous use of interpolation and the ramp-step modification, the backward average is used for methods G and C ; the forward average is used for methods A and F. Consequently, interpolation of graphical methods A and F is identical to averaging methods A and F for appropriate problems; and averaged methods G and C are identical to the ramp-step modification for appropriate problems. This also means that for problem II averaged methods A and G give identical results as do interpolated A and method G with a ramp-step modification as mentioned in reference (20).

(C) z-TRANSFORMS

The z-transform method of analysis of difference equations dates back to Heaviside and Laplace. Recently, it has been used in analyzing linear sampled-data control systems in electrical engineering and most of the properties derived here are shown in the texts of Aseltine (25), Jury (12), and Ragazzini and Franklin (13) on transform methods and sampled-data systems. In the following discussion the z-transforms are developed from the Laplace transforms and important properties are developed. This is not meant to be a rigorous mathematical discussion, but is meant to develop an understanding of the transforms. For purposes of reference, tables of z-transforms after Aseltine are reproduced (Tables IV-7 and IV-8).

A difference equation mathematically defines a function at only a point, and the function is not defined again until a certain increment of time or space has been traversed. (This obviously does not prevent us from interpolating between points.) For example, at a point m a temperature would only be defined at times $n\Delta\tau$, where n is an integer. This gives a sequence of impulses with zero width as shown in Figure IV-4.

The Laplace transform is defined by the equation (26)

$$\mathcal{L}\left\{f(\tau)\right\} = \int_0^\infty e^{-\tau s} f(\tau)d\tau \qquad \text{(IV-63)}$$

Considering that each impulse has an area associated with it equal to its magnitude, the Laplace transform integral of an impulse is, under these assumptions,

$$\int_{n\Delta\tau-}^{n\Delta\tau+} e^{-s\tau} f_n \, d\tau = (e^{-s\Delta\tau})^n f_n \tag{IV-64}$$

Therefore the Laplace transform of the sequence of impulses or of a difference signal is:

$$\mathcal{L}\left\{f_n\right\} = \int_0^{\infty} e^{-s\tau} f_n \, d\tau = \sum_{n=0}^{\infty} f_n e^{-sn\Delta\tau} \tag{IV-65}$$

Now, making the substitution

$$z = e^{s\Delta\tau} \tag{IV-66}$$

the expression becomes

$$\mathcal{L}\left\{f(n\Delta\tau)\right\} = \sum_{n=0}^{\infty} z^{-n} f(n\Delta\tau) = \sum_{n=0}^{\infty} f_n z^{-n} \tag{IV-67}$$

$$= f_0 + f_1 z^{-1} + f_2 z^{-2} + \ldots + f_n z^{-n} + \ldots$$

The z-transform is defined as the Laplace transform of a sampled signal, having an area equal to its magnitude, with the substitution z for $e^{s\Delta\tau}$ as given by equation IV-66. This definition is then

$$\mathcal{Z}\left\{f_n\right\} = \bar{f}(z) = \sum_{n=0}^{\infty} f_n z^{-n} \tag{IV-68}$$

The z-transform may be defined directly by the above relationship independently of the Laplace transform. As with the Laplace transform the difference signal is assumed to be zero for all n less than zero. The relationship between the Laplace transform and the z-transform is useful and the relationship in equation IV-66 is actually a mapping from the complex s-plane to the complex z-plane. For example, the imaginary axis of the s-plane is repeatedly mapped onto the unit circle in the z-plane; thus poles located in the area left of this axis in the s-plane for a stable solution are mapped inside the unit circle of the z-plane, which is the requirement of stability for the difference system. Also the z-transform of a signal sampled from a continuous signal may be found by a complex integration of the Laplace transform of the continuous signal times the transform of sequences of pulses (12, 13).

To show how the z-transform may be manipulated, a few of the important transform pairs will be briefly developed. For a unit step function that is zero for n less than zero, but one for all n equal to or greater than zero we would have

$$f_n = \begin{cases} 0 & n < 0 \\ 1 & n \geq 0 \end{cases} \qquad \text{(IV-69)}$$

$$\bar{f} = \mathcal{Z}\{f_n\} = \sum_{n=0}^{\infty} z^{-n} = 1 + z^{-1} + z^{-2} + \ldots + z^{-n}$$

$$= \frac{z}{z-1} \qquad \text{(IV-70)}$$

The last equality comes from Dwight (27), equation 9.04.

One of the most important Laplace transform relationships is that for a simple pole

$$\mathcal{L}\left\{ e^{\lambda \tau} \right\} = \frac{1}{s - \lambda} \tag{IV-71}$$

which includes the trigonometric interpretations if λ is considered to be complex. The corresponding z-transform is

$$\mathcal{Z}\left\{q^n\right\} = \sum_{n=0}^{\infty} q^n z^{-n} = 1 + q z^{-1} + q^2 z^{-2} + \ldots + q^n z^{-n} = \frac{z}{z - q} \tag{IV-72}$$

Again, if q is considered a complex number the above transform has corresponding interpretations in cosine and sine functions. Simple transforms are inverted using equation IV-72 and partial fractions. The third important relationship is the shifting property, which can be found starting from:

$$\mathcal{Z}\left\{f_{n+1}\right\} = \sum_{n=0}^{\infty} f_{n+1} z^{-n} = z \sum_{n=0}^{\infty} f_{n+1} z^{-(n+1)} \tag{IV-73}$$

Now change variable of summation by

$$p = n + 1 \tag{IV-74}$$

Then,

$$\mathcal{Z}\left\{f_{n+1}\right\} = z \sum_{p=1}^{\infty} f_p z^{-p} \tag{IV-75}$$

And adding and subtracting $(z\, f_0)$ on the right side of the equation:

$$\mathcal{Z}\left\{f_{n+1}\right\} = -z\, f_0 + z \sum_{p=0}^{\infty} f_p\, z^{-p} \qquad \text{(IV-76)}$$

which gives the shifting property:

$$\mathcal{Z}\left\{f_{n+1}\right\} = z\, \bar{f}_n - z\, f_0 \qquad \text{(IV-77)}$$

This relationship shows that the multiplying by z is analogous to multiplying by the advancing difference operator* E, as multiplying by s is analogous to multiplying by the operator $d/d\tau$.

To invert transforms not in the table an inversion theorem is used. This theorem is based on a Cauchey formula (12) for contour integration in the complex plane which is:

$$\frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} z^p\, dz = \begin{cases} 0 & p > -1 \\ 1 & p = -1 \\ 0 & p < -1 \end{cases} \qquad \text{(IV-78)}$$

where Γ is a closed curve enclosing the origin. To use this integral first we multiply the definition of the z-transform by z^{n-1} obtaining

$$z^{n-1}\, \bar{f}(z) = f_0 z^{n-1} + f_1 z^{n-2} + \ldots + f_{n-1} z^0 + f_n z^{-1} + f_{n+1} z^{-2} + \ldots \qquad \text{(IV-79)}$$

Now integrating both sides of equation IV-79 along a closed contour in the z-plane we obtain

---

* E defined by: $E\, f_n = f_{n+1}$

$$f_n = \frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} z^{n-1} \bar{f}(z)dz \qquad \text{(IV-80)}$$

The closed curve, Γ, must enclose all singularities of $z^{n-1} \bar{f}(z)$.

The usual circumstance is that all the singularities are poles and the integration path is taken as a circle enclosing them; this contour for a stable solution is the unit circle and is shown in Figure IV-5(a). In this case, the integral in equation IV-80 is evaluated directly from the residue theorem and the inversion is

$$f_n = \sum_{j=0}^{\infty} \text{residues of poles of } z^{n-1} \bar{f}(z) = \sum_{j=0}^{\infty} \rho_{j,n} \qquad \text{(IV-81)}$$

where $\rho_{j,n}$ = residues of $z^{n-1} \bar{f}(z)$ and the summation is over all the poles of $z^{n-1} \bar{f}(z)$.

The residue for simple poles (28) is

$$\rho_{j,n} = \lim_{z \to q_j} (z-q_j) z^{n-1} \bar{f}(z) \qquad \text{(IV-82)}$$

or

$$\rho_{j,n} = \lim_{z \to q_j} \frac{N(z)}{dD(z)/dz} \qquad \text{(IV-83)}$$

where

$$z^{n-1} \bar{f}(z) = \frac{N(z)}{D(z)}$$

$N(z)$ = numerator of $z^{n-1} \bar{f}(z)$; may contain elements of denominator as long as it is analytic as $z \to q_j$

$D(z)$ = denominator of $z^{n-1}\bar{f}(z)$ not contained in $N(z)$; goes to zero as $z \to q_j$

For a pole with multiplicity of μ the residue (28) is:

$$\rho_{j,n} = \lim_{z \to q_j} \left[\frac{1}{(\mu-1)!} \frac{\partial^{\mu-1}}{\partial z^{\mu-1}} \left\{ z^{n-1}(z-q_j)^{\mu} \bar{f}(z) \right\} \right] \qquad \text{(IV-84)}$$

Another form of equation IV-84 is found by using the formula for higher-order derivatives of a product.

$$\rho_{j,n} = \lim_{z \to q_j} \sum_{p=1}^{\mu} \frac{(n-1)! \; (z^{n-p}) \frac{d^{\mu-p}}{dz^{\mu-p}} \left\{ (z-q_j)^{\mu} \bar{f}(z) \right\}}{(n-p)! \; (\mu-p)! \; (p-1)!} \qquad \text{(IV-85)}$$

For the case of a pole with multiplicity of 2 an equation of the form of equation IV-83 is useful (28):

$$\rho_{j,n} = 2 \lim_{z \to q_j} \left[\frac{\frac{dN(z)}{dz}}{\frac{d^2 D(z)}{dz^2}} - \frac{N(z) \; \frac{d^3 D(z)}{dz^3}}{3 \left(\frac{d^2 D(z)}{dz^2} \right)^2} \right] \qquad \text{(IV-86)}$$

In some cases, the use of the above equations to invert $\bar{f}(z)$, summing only the residues of the singularities of $\bar{f}(z)$, gives a solution that agrees with the stepped-out calculations only for n's greater than some low number, n_1, where n_1 is usually zero. That is, the resulting solution fails to fit the assumed initial condition and the stepped-out

calculations up through increment n_1. The reason for this is that the inversion integration must be done on $z^{n-1}\bar{f}(z)$, not $\bar{f}(z)$, and that $z^{n-1}\bar{f}(z)$ may have a singularity at zero, or may have a higher-order singularity at zero than does $\bar{f}(z)$. But, since in this study of the partial difference equation of diffusion, we are not usually interested in the solution for $n = 0$ or 1, but only for the intermediate and large values of n, the residue of this singularity can be neglected. However, so that this apparent inconsistency does not cast doubt on the z-transform technique, the following two paragraphs show first, when this singularity occurs, and second, how to find the additional terms which when added to the residues of $\bar{f}(z)$ give the solution that is correct for all values of n.

Assuming that $\bar{f}(z)$ can be written as a ratio of polynomials in z, as is the usual case, then the behavior of $\bar{f}(z)$ as z goes to zero determines the existence and order of this singularity of $z^{n-1}\bar{f}(z)$. If $\bar{f}(z)$ goes to zero as z goes to zero, $z^{n-1}\bar{f}(z)$ must be analytic for all $n = 0, 1, \ldots, \infty$ at a z of zero. However, if $\bar{f}(z)$ goes to neither zero nor infinity as z goes to zero, $z^{n-1}\bar{f}(z)$ contains a simple pole at zero when n is zero. Here the sum of the residues of $\bar{f}(z)$ gives a solution valid for $n > 0$. If $\bar{f}(z)$ itself contains a pole at $z = 0$ of multiplicity n_1, $z^{n-1}\bar{f}(z)$ has a pole at $z = 0$ with multiplicity $(n_1 + 1)$ for an n of zero, and the sum of the residues of $\bar{f}(z)$ gives a solution valid for $n > n_1$.

In order to find the inversion relationship of a multiple pole at zero, the definition of the z-transform and the Kronecker delta function can be combined to give

$$\mathcal{Z}\left\{\delta_{n,p}\right\} = z^{-p} \qquad n = 0, 1, \ldots, \infty \qquad \text{(IV-87)}$$

where $\delta_{n,p} \equiv 0 \quad n \neq p$

$\qquad\qquad\quad 1 \quad n = p$

p = non-negative constant integer

Note that a constant term in the transform would give an impulse at an n of zero, $\delta_{n,0}$, because of the singularity introduced by z^{n-1}. If the transform can be written in terms of partial fractions the inversion is simply accomplished using this relationship and equations IV-87 and IV-72. For the complicated transforms that result from the diffusion problem this cannot be done conveniently. In these cases equation IV-82 may be used to find the residue of the zero pole. Since we are applying this to $z^{n-1}\bar{f}(z)$ the multiplicity of the zero pole is $(n_1 + 1)$ where n_1 is the multiplicity of zero pole for $\bar{f}(z)$; (n_1 must be zero or positive; if it is negative there is no singularity at zero). As an example, using equation IV-84 to invert the transform z^{-1} with a simple pole ($\mu = 2$), we obtain

$$\lim_{z \to 0} \frac{1}{1!} \frac{\partial}{\partial z}\left\{z^{n-1} z^{2} z^{-1}\right\} = \lim_{z \to 0} \frac{\partial}{\partial z} z^{n} = \lim_{z \to 0} n(z)^{n-1} = n(0)^{n-1} \qquad \text{(IV-88)}$$

$$n = 0, 1, \ldots, \infty$$

Although the term $n(0)^{n-1}$ is indeterminate for an n of zero, it is actually a short way of representing the limit of $n(z)^{n-1}$ as z goes to zero. Taking that limit using l'Hospital's rule for n of zero and usual definitions for other values of n this result is equivalent to equation IV-87

$$\delta_{n,1} = n(0)^{n-1} \qquad n = 0, 1, \ldots, \infty \qquad \text{(IV-89)}$$

As analogous arguments are valid for a higher-order pole at zero, it is concluded that this approach is valid and equivalent to using partial fractions to invert the transform for determining this residue.

A second type of singularity that can occur is a branch cut. The inversion of a transform of this type is like the procedure used when a branch cut occurs with Laplace transforms. The inversion integration is carried out on a closed curve that excludes the branch point and other singularities, but part of the path is a circle that would in a limit enclose the origin. ABCDA in Figure IV-5(b) is such a path for a branch cut from -1 to +1. The details of such an integration are shown in section D-4 of this chapter by an example.

A simple numerical method for the transform inversion is to generate the definition series equation IV-68. The coefficient of each $z^{-0}, z^{-1}, \ldots, z^{-n}$ is then $f_0, f_1, \ldots, f_n$. This may be simply done by arranging the transform as a ratio of polynomials in z^{-1} and carrying

out the long division. This procedure is often useful in checking the algebra used to derive the transform. Also by a simple examination of the definition the initial value theorem is derived,

$$\lim_{n \to 0} f_n = \lim_{z \to \infty} \bar{f}(z) \qquad \text{(IV-90)}$$

This theorem is used in checking derived transforms.

This brief discussion on the z-transform is not a presentation of rigorous proofs but is intended to show the development of z-transforms from the Laplace transforms, the important transform pairs, and the inversion. For a more rigorous treatment, any of the three texts mentioned should be consulted, particularly for such questions as uniqueness and restrictions on the types of functions. Texts (28) on complex variables are also useful.

(D) z-TRANSFORMS AND PARTIAL DIFFERENCE EQUATIONS--EXAMPLE PROBLEMS

The z-transform procedure for solving a partial difference equation is identical to that required for the Laplace transform procedure of solving a partial differential equation. The procedure is:

(1) Transform the partial difference equation with respect to the independent variable for which the initial values are known. This gives a difference equation with one less independent variable.

(2) Transform any boundary condition equations together with their initial conditions.

(3) Solve the transformed difference equation with the boundary conditions to obtain an explicit form for the transformed dependent variable.

(4) Invert the transform by using the tables or the inversion theorem to find the analytic solution of the partial difference equation. As z-transforms are not well known, the tables available are not as complete as those for the Laplace transform and usually the inversion theorem is required, particularly for the transforms obtained from the partial difference equation for diffusion.

Since the z-transform method of solving partial difference equations has not been found published, several related difference problems are now solved in detail to demonstrate the technique and the manipulations required. The problems selected are not the simplest possible,

but they show many of the difficulties encountered, how to overcome them, and where simplifications are possible. Most steps are carried out in some detail with discussion of special problems in addition to simplifications. In the earlier part of the problem the equations are kept in the most general form for as long as possible.

1. Problem I--Finite H, Method G

The difference problem to be solved involves a one-dimensional solid. Method G, Table IV-1, will be used to approximate a continuous problem with an arbitrary initial condition. The left boundary (x=0) is in contact with a fluid with a finite heat-transfer coefficient and whose temperature is specified as a function of time. The right boundary (x=L) is adiabatic. Method G uses mesh $\Delta\xi$ in Figure IV-2 with points located at both boundaries and uniformly spaced and it is defined in detail in Table IV-1. The arbitrary initial condition for the difference method is represented by $t_{m,0}$, where m goes from 0 to S corresponding to ξ from 0 to 1, and the fluid temperature is represented by $t_{f,n}$. Since to obtain a solution using z-transforms it is necessary to specify the initial distribution and how the fluid temperature changes with time, these conditions are taken as:

$$t_{m,0} = 0 \qquad 0 \leq m \leq S \tag{IV-91}$$

$$t_{f,n} = 1 \qquad n \geq 0 \tag{IV-92}$$

This is problem I as described in Table IV-9. However, these specific conditions are not introduced into the equations until necessary. After solving the specific problem above, the following related problems are shown based on method G. First, the solution is modified to an averaged method G. Second, the case for specified boundary condition or infinite heat-transfer coefficient is shown. Then the solution for an explicit method for a semi-infinite solid is shown.

The one-dimensional partial difference equation for diffusion in general form, equation II-42, is:

$$-r\gamma t_{m-1,n+1} + (1+2r\gamma)t_{m,n+1} - r\gamma t_{m+1,n+1} = r(1-\gamma)t_{m-1,n} + [1-2r(1-\gamma)]t_{m,n} + r(1-\gamma)t_{m+1,n} \qquad \text{(IV-93)}$$

$$m = 1, 2, \ldots, S-1$$

$$n = 0, 1, 2, \ldots, \infty$$

The boundary equations for the finite h at point zero and zero h at point S for method G are, respectively:

$$\left[1+2r\gamma\left(1+\frac{H}{S}\right)\right]t_{0,n+1} - 2r\gamma t_{1,n+1} = \left[1-2r(1-\gamma)\left(1+\frac{H}{S}\right)\right]t_{0,n} + 2r(1-\gamma)t_{1,n} + 2r\frac{H}{S}\left[(1-\gamma)t_{f,n} + \gamma t_{f,n+1}\right]$$

$$n = 0, 1, \ldots, \infty \qquad \text{(IV-94)}$$

$$(1+2r\gamma)t_{S,n+1} - 2r\gamma t_{S-1,n+1} = [1-2r(1-\gamma)]t_{S,n} + 2r(1-\gamma)t_{S-1} \qquad \text{(IV-95)}$$

$$n = 0, 1, \ldots, \infty$$

where $H = hL/k$.

A significant simplification in the adiabatic boundary equation IV-95 can be made by writing the partial difference equation IV-93 for point S, which requires the introduction of an imaginary point at (S+1). Subtracting the boundary equation IV-95 from the partial difference equation the result is simply:

$$r\gamma(t_{S-1,n+1} - t_{S+1,n+1}) = r(1-\gamma)(-t_{S-1,n} + t_{S+1,n}) \qquad \text{(IV-96)}$$

$$n = 0, 1, \ldots, \infty$$

Since the imaginary point (S+1) must have the same initial condition as point (S-1) we have simply

$$t_{S+1,n} = t_{S-1,n} \qquad n = 0, 1, \ldots, \infty \qquad \text{(IV-97)}$$

With this modification, the partial difference equation IV-93 is written for point S. The method for simplifying the adiabatic boundary condition could be used to modify the other heat-transfer coefficient condition, but as long as the coefficient is not zero no simplification is obtained.

Now that the partial difference equation and the boundary conditions have been simplified, they can be transformed using the relationships in the table. The transformed equations are, with no

simplification, in terms of the arbitrary initial condition and fluid temperature for equations IV-93, IV-94, and IV-97, respectively:

$$-r\gamma z[\bar{t}_{m-1} - t_{m-1,0} + \bar{t}_{m+1} - t_{m+1,0}] + (1+2r\gamma)z[\bar{t}_m - t_{m,0}] =$$

$$r(1-\gamma)[\bar{t}_{m-1} + \bar{t}_{m+1}] + [1-2r(1-\gamma)]\bar{t}_m \qquad \text{(IV-98)}$$

$$m = 1, 2, \ldots, S-1, S$$

$$z\left[1+2r\gamma\left(1+\frac{H}{S}\right)\right][\bar{t}_0 - t_{0,0}] - 2r\gamma z[\bar{t}_1 - t_{1,0}] =$$

$$\left[1-2r(1-\gamma)\left(1+\frac{H}{S}\right)\right]\bar{t}_0 + 2r(1-\gamma)\bar{t}_1 \qquad \text{(IV-99)}$$

$$+ 2r\frac{H}{S}\left[(1-\gamma)\bar{t}_f + \gamma z(\bar{t}_f - t_{f,0})\right]$$

$$\bar{t}_{S+1} = \bar{t}_{S-1} \qquad \text{(IV-100)}$$

A more convenient form for equations IV-98 and IV-99 is found by rearranging and combining terms so that the unknown transformed variables are on the left side and the known quantities and functions specified by the initial and boundary conditions are on the right.

$$\bar{t}_{m-1} - \left[\frac{z-1+2r(z\gamma+1-\gamma)}{r(z\gamma+1-\gamma)}\right]\bar{t}_m + \bar{t}_{m+1} =$$

$$\frac{z}{r(z\gamma+1-\gamma)}\left[r\gamma(t_{m-1,0} + t_{m+1,0}) - (1+2r\gamma)t_{m,0}\right] \qquad \text{(IV-101)}$$

$$m = 1, 2, \ldots, S-1, S$$

$$\left[z+1+2r\left(1+\frac{H}{S}\right)(\gamma z+1-\gamma)\right]\bar{t}_0 - 2r(\gamma z+1-\gamma)\bar{t}_1 =$$

$$\frac{2rH}{S}(\gamma z+1-\gamma)\bar{t}_f \qquad \text{(IV-102)}$$

$$+ z\left[\left\{1+2r\gamma\left(1+\frac{H}{S}\right)\right\}t_{0,0} - 2r\gamma t_{1,0} - 2r\frac{H\gamma z}{S}t_{f,0}\right]$$

$$\bar{t}_{S+1} = \bar{t}_{S-1} \qquad \text{(IV-103)}$$

These three equations are, respectively, a second-order ordinary difference equation, IV-101, and its two boundary condition equations, IV-102 and IV-103. The first equation is the transformed one-dimensional diffusion difference equation for any initial distribution.

Application of Initial Conditions. In order to find the solution of the difference equation IV-101 and its boundary equations IV-102 and IV-103, the initial conditions $t_{m-1,0}$, $t_{m+1,0}$, and $t_{m,0}$ must be expressed as functions of m and in a form that applies for each value of m for which the difference equation applies. In equation IV-101 if γ is not zero an expression for $t_{m,0}$ must apply for m of 0 to (S+1) even though the difference equation applies for m from 0 to S. In most cases, a suitable function of m may be obtained directly from the initial distribution for the continuous problem. However, in some cases, particularly when the surface temperature is specified, the required expression is not obvious. Consider for a moment the case when the surface temperature $t_{0,0}$ is specified as one, but the initial

temperature distribution $t_{m,0}$ for $m = 1, \ldots (S+1)$ is zero. In this case a suitable function that fits the initial distribution is:

$$t_{m,0} = \delta_{m,0} \qquad m = 0, 1, \ldots, S \qquad \text{(IV-104)}$$

where $\delta_{m,p}$ = Kronecker delta function = $\begin{cases} 1 & m = p \\ 0 & m \neq p \end{cases}$

(Note second subscript on δ does not refer to time in this case as does the second subscript on t.) In general, a series of terms using the Kronecker delta function may be used:

$$t_{m,0} = \sum_{p=0}^{S} t_{p,0}\, \delta_{m,p} \qquad m = 0, 1, \ldots, S \qquad \text{(IV-105)}$$

If an analytic expression for the Kronecker delta function is necessary to allow a solution of the difference equation, either a form with 0 roots or the Lagrangian interpolation formula may be satisfactory, but both are complicated. These are, respectively,

$$\delta_{m,p} = \frac{m(m-1) \ldots (m-p+1)(0)^{m-p}}{p!} \qquad m,p = 0, 1, \ldots, S \qquad \text{(IV-106)}$$

$$\delta_{m,p} = \frac{m(m-1) \ldots (m-p+1)(m-p-1) \ldots (m-S)}{p(p-1) \ldots (1)(-1) \ldots (p-S)} \qquad m,p = 0, 1, \ldots, S \qquad \text{(IV-107)}$$

In equation IV-106 to keep the term $[m(m-1) \ldots (m-p+1)(0)^{m-p}]$ finite for m less than p, it is considered as the limit of

$[m(m-1) \ldots (m-p+1)\,(\epsilon)^{m-p}]$ as ϵ goes to zero as used in equation IV-89. For mesh $\Delta\xi/2$ the variable m would be replaced by $(m - \frac{1}{2})$.

Another method that avoids the series above is to first solve the explicit case ($\gamma = 0$). The partial difference equation is then

$$\bar{t}_{m-1} - \left(\frac{z-1+2r}{r}\right)\bar{t}_m + \bar{t}_{m+1} = \frac{-z}{r}\, t_{m,0} \qquad m = 1, 2, \ldots, S \quad \text{(IV-108)}$$

In this case, the expression for the initial distribution does not need to fit the initial boundary temperature $t_{0,0}$ as the equation applies only to $m = 1, 2, \ldots, S$. The boundary equation allows for the initial boundary temperature in equation IV-102 for point 1, which can be rewritten for zero γ:

$$\left[z-1+2r\left(1+\frac{H}{S}\right)\right]\bar{t}_0 - 2r\bar{t}_1 = \frac{2rH}{S}\,\bar{t}_f + z t_{0,0} \qquad \text{(IV-109)}$$

The solution for the explicit case is found; then, by setting n to zero in this solution an expression results which fits the initial condition for all $m = 0, 1, \ldots, (S+1)$: This expression is then substituted in equations IV-101 and IV-102 for the implicit case. Although this expression is a series, because of the way it was found, it simplifies the solution of the original system of equations IV-101, IV-102, and IV-103.

For cases where the Y/A matrix and the particular solution are not functions of the time differencing parameters $\Delta\tau$ and γ, an additional simplification can often be made. Here, in addition to the particular solution, neither the eigenvectors of the Y/A matrix nor the initial vectors are functions of the time differencing parameters. Consequently,

these quantities can be found by using the z-transform to solve the problem for a convenient combination of time increment $\Delta\tau$ and weighting γ; this usually corresponds to an r of $\frac{1}{2}$ and a γ of zero. This solution also gives the q_j's that correspond to the specific time differencing parameters used, and the eigenvalues of the Y/A matrix, λ_j, are then found from equation II-56. Once these values are known, the Q matrix for any combination of $\Delta\tau$ and γ is known from equation II-57 and the analog solution is obtained by replacing the Q^n matrix with the $e^{+\Lambda\tau}$ matrix. This latter fact means that in these cases the Laplace transform could be applied to the analog system of differential equations.

$$\frac{d\,t_m}{d\tau} = \frac{1}{(\Delta\xi)^2}\left[t_{m-1} - 2t_m + t_{m+1}\right] \qquad \text{(IV-110)}$$

together with the analog boundary conditions. After taking the Laplace transform this gives a boundary value ordinary difference problem as found in equations IV-101, IV-102, and IV-103. Solving these equations and inverting the Laplace transform gives the eigenvectors and initial vector for the approximate method and the eigenvalues λ_j from which not only the analog solution is found, but also the Q matrix for the difference solution can be found. These simplifications can be applied to methods G and C as applied to problems where a true steady-state solution exists.

Solution of Transformed Difference Equation. Returning to the example problem, to solve the ordinary difference equation we take the initial temperature distribution as zero at all points,

$$t_{m,0} = 0 \qquad 0 \le m \le S+1 \tag{IV-111}$$

Thus the equations IV-101, IV-102, and IV-103 become,

$$\bar{t}_{m-1} - \left[\frac{z-1+2r(z\gamma+1-\gamma)}{r(z+1-\gamma)}\right]\bar{t}_m + \bar{t}_{m+1} = 0 \tag{IV-112}$$

$$m = 1, 2, \ldots, S, S+1$$

$$\left[z-1+2r\left(\frac{H}{S}+1\right)(\gamma z+1-\gamma)\right]\bar{t}_0 - 2r(\gamma z+1-\gamma)\bar{t}_1 = 2r\frac{H}{S}\left[(\gamma z+1-\gamma)\bar{t}_f - z\gamma t_{f,0}\right] \tag{IV-113}$$

$$\bar{t}_{S-1} = \bar{t}_{S+1} \tag{IV-114}$$

The difference equations IV-101, IV-102, and IV-103 are more easily solved with the following definition (15, 21):

$$\cos\alpha = \frac{z-1+2r(z\gamma+1-\gamma)}{2r(z\gamma+1-\gamma)} \tag{IV-115}$$

or inversely

$$z = \frac{1-2r(1-\gamma)(1-\cos\alpha)}{1+2r\gamma(1-\cos\alpha)} \tag{IV-116}$$

Substitution gives, simply, for the difference equation IV-112

$$\bar{t}_{m+1} - 2\cos\alpha\,\bar{t}_m + \bar{t}_{m-1} = 0 \qquad m = 1, 2, \ldots, S, S+1 \tag{IV-117}$$

The transformed boundary equation is simplified by solving equation IV-116 for (z-1)

$$z-1 = 2r(\gamma z+1-\gamma)(1-\cos\alpha) = \frac{2r(\cos\alpha-1)}{1+2r\gamma(1-\cos\alpha)} \qquad \text{(IV-118)}$$

Substituting this in the coefficient for $\bar{t}_0$ in equation IV-113 gives:

$$\left\{\cos\alpha+\frac{H}{S}\right\}\bar{t}_0 - \bar{t}_1 = \frac{H}{S(\gamma z+1-\gamma)}\left[(\gamma z+1-\gamma)\,\bar{t}_f - z\gamma t_{f,0}\right] \qquad \text{(IV-119)}$$

The solution of the homogeneous equation IV-117 is:

$$\bar{t}_m = P\cos m\alpha + G\sin m\alpha \qquad m = 0,\ 1,\ \ldots,\ S,\ S+1 \qquad \text{(IV-120)}$$

where P and G are constants to be found from the two boundary relationships or their modified forms, equations IV-114 and IV-119, so that the solution also satisfies these boundary equations. If a non-zero initial condition were present, equation IV-117 would not be homogeneous and a particular solution found by the method of undetermined parameters would be added to the above complementary solution.

To evaluate the constants P and G we first find the expressions for $\bar{t}_{S+1}$ and $\bar{t}_{S-1}$ from equation IV-120 by setting m equal to (S+1) and (S-1), respectively.

$$\bar{t}_{S+1} = P\cos(S+1)\alpha + G\sin(S+1)\alpha \qquad \text{(IV-121)}$$

$$\bar{t}_{S-1} = P\cos(S-1)\alpha + G\sin(S-1)\alpha \qquad \text{(IV-122)}$$

and substituting them in the boundary equation IV-114 gives:

$$\bar{t}_{S+1} - \bar{t}_{S-1} = 0 = P\cos(S+1)\alpha + G\sin(S+1)\alpha - P\cos(S-1)\alpha - G\sin(S-1)\alpha \quad \text{(IV-123)}$$

Simplifying with trigonometric identities gives

$$P\tan S\alpha = G \quad \text{(IV-124)}$$

Now G is eliminated from equation IV-120 which, using trigonometric identities, gives

$$\bar{t}_m = \frac{P\cos(S-m)\alpha}{\cos S\alpha} \quad \text{(IV-125)}$$

From equation IV-125 the transforms $\bar{t}_0$ and $\bar{t}_1$ are:

$$\bar{t}_0 = \frac{P\cos S\alpha}{\cos S\alpha} \quad \text{(IV-126)}$$

$$\bar{t}_1 = \frac{P\cos(S-1)\alpha}{\cos S\alpha} \quad \text{(IV-127)}$$

and substituting them in equation IV-119 gives:

$$\frac{P}{\cos S\alpha}\left[\left(\cos\alpha + \frac{H}{S}\right)\cos S\alpha - \cos(S-1)\alpha\right] = \frac{H}{S(\gamma z+1-\gamma)}\left[(\gamma z+1-\gamma)\bar{t}_f - z\gamma\bar{t}_{f,0}\right] \quad \text{(IV-128)}$$

Solving for $P/\cos S\alpha$ and substituting in equation IV-125 together with trigonometric simplification of the bracketed form on the left gives the solution for the transformed difference temperature.

$$\bar{t}_m = \frac{H}{S}\left[\frac{(\gamma z+1-\gamma)\bar{t}_f - z\gamma t_{f,0}}{\gamma z+1-\gamma}\right]\left[\frac{\cos(S-m)\alpha}{\frac{H}{S}\cos S\alpha - \sin S\alpha \sin\alpha}\right] \qquad \text{(IV-129)}$$

At this point, it is usually advantageous to check the solution in the boundary equations, and also to check the initial conditions using the initial value theorem, item 14, Table IV-7.

Inversion of Transform. The next and last step is the inversion of the transform. To do this we specify how the fluid temperature varies with time. Assuming that it is constant at unity for all n equal to and greater than zero we have from item 1, Table IV-8:

$$\bar{t}_f = \frac{z}{z-1} \qquad \text{(IV-130)}$$

$$n = 0, 1, \ldots, \infty$$

$$t_{f,n} = 1$$

Substituting in the transform temperature in equation IV-129,

$$\bar{t}_m = \left[\frac{z}{(z-1)(\gamma z+1-\gamma)}\right]\left[\frac{H\cos(S-m)\alpha}{H\cos S\alpha - S\sin S\alpha \sin\alpha}\right] \qquad \text{(IV-131)}$$

The inversion theorem must be used to find $t_{m,n}$. The quantity $(z^{n-1}\bar{t}_m)$ has only poles for singularities. These poles occur at values of z for which the quantities (z-1) and $(H\cos S\alpha - S\sin\alpha \sin S\alpha)$ go to zero. Although it appears that $(z\gamma+1-\gamma)$ gives another pole, we will show first that $\bar{t}_m$ is actually analytic at a z of $(\gamma-1)/\gamma$; then we will discuss the roots of the trigonometric quantity and evaluate the residues.

To show that $(\gamma-1)/\gamma$ is not a pole of $\bar{t}_m$, from trigonometric identities and the definition of $\cos\alpha$:

$$\sin\alpha = \frac{\sqrt{-(z-1)(z-1+4r\{\gamma z+1-\gamma\})}}{2r(z\gamma+1-\gamma)} \qquad \text{(IV-132)}$$

Now from the formulas for sine and cosine of multiple angles we know that $\cos(S-m)\alpha$ is a polynomial that contains $(\cos\alpha)$ to the $(S-m)$ power; $(\cos S\alpha)$ contains $(\cos\alpha)$ to the S power; $(\sin S\alpha)$ contains $(\sin\alpha)$ to the S power or $(\cos\alpha)$ times $(\sin\alpha)$ to the $(S-1)$ power, depending on S odd or even. Thus, since both $(\cos\alpha)$ and $(\sin\alpha)$ contain the factor $(\gamma z+1-\gamma)$ in the denominator, the term $(H\cos S\alpha - S\sin\alpha\sin S\alpha)$ is a polynomial that contains the factor $(\gamma z+1-\gamma)$ to the $(S+1)$ power in its denominator; and $\cos(S-m)\alpha$ with $(\gamma z+1-\gamma)$ to the $(S-m)$ power in its denominator. Consequently, $\bar{t}_m$ would contain $(\gamma z+1-\gamma)$ to the $[S+1-(S-m)-1]$ or to the mth power in its numerator; and as z goes to $(\gamma-1)/\gamma$, $\bar{t}_m$ has a limit; therefore, $\bar{t}_m$ is analytic for all m from 0 to $(S+1)$.

Next the residue at a z of one is evaluated using equation IV-82. Note that in taking limits as z goes to one we must also evaluate α which, from equation IV-115, goes to 0.

$$\lim_{z\to 1}\cos\alpha = \lim_{z\to 1}\left[\frac{z-1+2r(\gamma z+1-\gamma)}{2r(\gamma z+1-\gamma)}\right] = 1 = \cos(0) \qquad \text{(IV-133)}$$

$$\rho_{1,n} = \lim_{\substack{z\to 1\\ \alpha\to 0}} (z-1)z^{n-1}\,\bar{t}_m = \lim_{\substack{z\to 1\\ \alpha\to 0}} \left\{ \frac{z^n H\cos(S-m)\alpha}{(\gamma z+1-\gamma)(H\cos S\alpha - S\sin\alpha\sin S\alpha} \right\} \qquad \text{(IV-134)}$$

$$= 1$$

This corresponds to the steady-state or particular solution.

The roots of the quantity ($H\cos S\alpha - S\sin\alpha\sin S\alpha$) are the poles for the transient part of the solution and must satisfy

$$S\sin\alpha_j \tan S\alpha_j = H \qquad j = 1, 2, \ldots, S+1 \qquad \text{(IV-135)}$$

Using the relationship between z and α, equation IV-116 gives:

$$\lim_{\substack{\alpha\to\alpha_j\\ z\to q_j}} z = q_j = \frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r\gamma(1-\cos\alpha_j)} \qquad \text{(IV-136)}$$

$$j = 1, 2, \ldots, S+1$$

We can write the $j = 1, 2, \ldots, (S+1)$, because method G has a Y/A matrix of dimension (S+1) by (S+1) and we know from equation IV-135 that (S+1) eigenvalues or q_j's exist. Assuming that the q_j's are distinct and simple, the residue for the jth root is found from equation IV-83.

$$\rho_{j,n} = \lim_{\substack{z\to q_j\\ \alpha\to\alpha_j}} \frac{N(z)}{\frac{d}{dz}D(z)} \qquad \text{(IV-137)}$$

where, for this problem,

$$N(z) = \frac{z^n H[\cos(S-m)\alpha]}{(z-1)(\gamma z+1-\gamma)} = \frac{z^n H[\cos(S-m)\alpha][1+2r\gamma(1-\cos\alpha)]^2}{2r(\cos\alpha-1)} \qquad \text{(IV-138)}$$

$$D(z) = H\cos S\alpha - S\sin S\alpha \sin\alpha \qquad \text{(IV-139)}$$

$$\frac{d}{dz}D(z) = \frac{S[H\sin S\alpha + S\sin\alpha\cos S\alpha + \sin S\alpha\cos\alpha][1+2r\gamma(1-\cos\alpha)]^2}{2r\sin\alpha} \qquad \text{(IV-140)}$$

The second equality for $N(z)$ comes from the relationship of $(z-1)$ and α, equation IV-118. The residue becomes then

$$\rho_{j,n} = \frac{H(\sin\alpha_j)[\cos(S-m)\alpha_j](q_j^n)}{S[1-\cos\alpha_j][(H\cos\alpha_j)(\sin S\alpha_j)+(S)(\sin\alpha_j)(\cos S\alpha_j)]} \qquad \text{(IV-141)}$$

or, after simplification using a trigonometric identity:

$$\rho_{j,n} = \frac{-H[\cos(S-m)\alpha_j][q_j^n]}{S\left[\tan\frac{\alpha_j}{2}\right]\left[(H\cos\alpha_j)(\sin S\alpha_j)+(S)(\sin\alpha_j)(\cos S\alpha_j)\right]} \qquad \text{(IV-142)}$$

Before summing over residues to find the solution, the assumption of distinct q_j's must be investigated. Study of the transcendental characteristic equation IV-135 shows that $(S+1)$ q_j's are real and distinct if the range of real α's is restricted from 0 to π and that one α_j is complex of the form $[\pi + \sqrt{-1}\,\phi(\alpha)]$. If the S real α_j's satisfy the equation, both $-\alpha_j$ and $(\alpha_j \pm p2\pi)$ also satisfy the equation. However, if two different α_j's are used that give the same q_j they have identical eigenvectors. This is seen as the mth element in the jth eigenvector

is given by $[\cos(S-m)\alpha_j]$; but from trigonometry we have

$$\cos(S-m)(\alpha_j) = \cos(S-m)(-\alpha_j) \qquad m = 0, 1, 2, \ldots, S \qquad \text{(IV-143)}$$

$$\cos(S-m)(\alpha_j) = \cos(S-m)(\alpha_j + p2\pi) \qquad \text{(IV-144)}$$

Thus, to have a matrix of eigenvectors which is non-singular, as required from matrix arguments, the q_j's must be distinct and the real α_j's are conveniently selected as those in the range 0 to π. Analogous arguments apply for the complex α. Thus we have (S+1) roots α_j's, S of which are real and fall between 0 and π and one of which is complex of the form $[\pi + \sqrt{-1}\,\phi(\alpha)]$ which result in (S+1) real, distinct q_j's.

Now, summing over all residues, the complete difference solution is

$$t_{m,n} = 1 - \frac{H}{S} \sum_{j=1}^{S+1} \frac{[\cos(S-m)\alpha_j][q_j^{n-}]}{\left[\tan\frac{\alpha_j}{2}\right]\left[(H+\cos\alpha_j)(\sin S\alpha_j)+(S)(\sin\alpha_j)(\cos S\alpha_j)\right]} \qquad \text{(IV-145)}$$

where the q_j's are given by equation IV-136 and the α_j's are the roots of equation IV-135. From equations IV-145 and II-53 and II-56, the analog solution is given by

$$t_{m,n} = 1 - \frac{H}{S} \sum_{j=1}^{S+1} \frac{[\cos(S-m)\alpha_j]\left[e^{-2S^2(1-\cos\alpha_j)\tau}\right]}{\left[\tan\frac{\alpha_j}{2}\right]\left[(H+\cos\alpha_j)(\sin S\alpha_j) + S\sin\alpha_j \cos S\alpha_j\right]} \qquad \text{(IV-146)}$$

Reviewing, the difference and analog solutions above are for using method G to approximate the continuous problem for a slab of width L, initially at zero temperature throughout with the right boundary insulated and the left boundary brought into contact with a fluid at unit temperature and heat transfer-coefficient h. In terms of dimensionless variables

$$\frac{\partial T}{\partial \tau} = \frac{\partial^2 T}{\partial \xi^2} \qquad 0 \le \xi \le 1 \qquad \text{(IV-147)}$$

Initial Condition:

$$T(\xi, 0) = 0 \qquad 0 \le \xi \le 1 \qquad \tau = 0 \qquad \text{(IV-148)}$$

Boundary Conditions:

$$H[T(0,\tau)-1] = \frac{\partial T}{\partial \xi}(0,\tau) \qquad \xi = 0 \qquad \tau \ge 0 \qquad \text{(IV-149)}$$

$$\frac{\partial T}{\partial \xi}(1,\tau) = 0 \qquad \xi = 1 \qquad \tau \ge 0 \qquad \text{(IV-150)}$$

The exact solution of the continuous problem is:

$$T(\xi,\tau) = 1 - 2H \sum_{j=0}^{\infty} \frac{[\cos(1-\xi)\nu_j]\left[e^{-\nu_j^2 \tau}\right]}{(\nu_j)[(H+1)(\sin\nu_j + \nu_j \cos\nu_j)]} \qquad \text{(IV-151)}$$

where the υ_j are defined by the characteristic equation,

$$\upsilon_j \tan \upsilon_j = H = \frac{hL}{k} \tag{IV-152}$$

The form of both the difference approximation and the continuous solution are the same; as a matter of fact, the continuous solution could be derived from the difference solution by properly taking limits as $S \to \infty$, $\Delta\xi \to 0$, and $\Delta\tau \to 0$. For a direct comparison, the only modification necessary is to substitute difference parameters for the continuous ξ, τ, and υ according to:

$$\psi = \upsilon \Delta\xi = \frac{\upsilon}{S} \tag{IV-153}$$

$$\frac{1}{S} = \Delta\xi \tag{IV-154}$$

$$\xi = m\Delta\xi \tag{IV-155}$$

$$\tau = n\Delta\tau \tag{IV-156}$$

$$r = \frac{\Delta\tau}{(\Delta\xi)^2} \tag{IV-157}$$

which then allows us to write all the solutions in difference form.

Difference Solution:

$$t_{m,n} = 1 - \frac{H}{S} \sum_{j=1}^{S+1} \frac{[\cos(S-m)\alpha_j]\left[\frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r\gamma(1-\cos\alpha_j)}\right]^n}{\left[\tan\frac{\alpha_j}{2}\right]\left[H+\cos\alpha_j)(\sin S\alpha_j) + S(\sin\alpha_j)(\cos S\alpha_j)\right]} \tag{IV-158}$$

Analog Solution:

$$t_{m,n} = 1 - \frac{H}{S} \sum_{j=1}^{S+1} \frac{\left[\cos(S-m)\alpha_j\right]\left[e^{-2r(1-\cos\alpha_j)}\right]^n}{\left[\tan\frac{\alpha_j}{2}\right]\left[H+\cos\alpha_j)(\sin S\alpha_j)+ S(\sin\alpha_j)(\cos S\alpha_j)\right]} \tag{IV-159}$$

where α_j is defined by

$$S(\sin\alpha_j)(\tan S\alpha_j) = H \qquad j = 1, 2, \ldots, S+1 \tag{IV-160}$$

Continuous Solution:

$$T(m,n) = 1 - 2H \sum_{j=1}^{\infty} \frac{\left[\cos(S-m)\psi_j\right]\left[e^{-\psi_j^2 r}\right]^n}{(S\psi_j)\left[H+\cos\psi_j)(\sin S\psi_j) + S\psi_j(\cos S\psi_j)\right]} \tag{IV-161}$$

where

$$S\psi_j(\tan S\psi_j) = H \qquad j = 1, 2, \ldots, \infty \tag{IV-162}$$

Thus, the z-transform leads to results which allow a comparison of the analytic solution of the difference formulation with the continuous solution. The above equations are scalar representations of the mth component* of the vector solution derived before. The particular solution

* As this method is based on mesh $\Delta\xi$, m starts at zero instead of 1 as it should in conventional matrix-vector nomenclature. The mth component of the vector is then the temperature at point m_{th} which in conventional vector notation is the (m+1) component; the mth row of the eigenvector matrix consists of the coefficients in the matrix row that are

vector for all solutions has dimension (S+1) with a one for each element. The eigenvector matrix for the difference and analog solution has an element in the mth row* and j^{th} column defined by $[\cos (S-m)\alpha_j]$ and the corresponding element in the continuous solution is defined by $[\cos (S-m)\psi_j]$. The jth element of the initial vector, or vector of Fourier coefficients, is, for the approximate solutions, $H/S(\tan \frac{\alpha_j}{2})[(H+\cos\alpha_j)(\sin S\alpha_j)+ S(\sin\alpha_j)(\cos S\alpha_j)]$, and for the continuous solution, $2H/(S\psi_j)[(H+\cos\psi_j)(\sin S\psi_j)+S\psi_j(\cos S\psi_j)]$. The characteristic equations that define the α_j and ψ_j are given by equations IV-160 and IV-162.

Another method of comparison of the solutions that would apparently avoid the lengthy inversion procedure would be to derive the z-transforms of the analog and continuous solutions and compare them to the z-transform of the difference solution. However, to convert the Laplace transform to a z-transform also requires a complex integration. Most chemical engineers (including the author) would have difficulty comparing the meaning of the transforms.

---

* (continued)
in the equation for the mth point, which is the (m+1) row in conventional matrix notation. A similarcomment applies if method C is being discussed as m starts at $\frac{1}{2}$ and the mth component would be the $(m+\frac{1}{2})$ component in conventional matrix notation.

2. Problem I, Averaged Method G--Ramp-Step Equivalence.

For problems with certain types of initial temperature distributions, a backward-averaged solution is identical to that obtained by replacing a constant fluid temperature with a ramp-step fluid temperature function, as shown previously.

The solution for method G applied to problem I using the ramp-step function will now be found. The ramp-step function for this problem is given by equation IV-54, and is

$$t_{f,n} = \begin{cases} \frac{1}{2} & n=0 \\ 1 & n \geq 1 \end{cases} \qquad \text{(IV-163)}$$

To generalize this for implicit methods, we use the fluid temperature from equation IV-163 above directly and do not use the weights of γ and $(1-\gamma)$ as shown in equation IV-94. The z-transform for equation IV-163 is easily derived from the series definition, equation IV-67.

$$\mathcal{Z}\{t_{f,n}\} = \tfrac{1}{2}\, z^{-0} + z^{-1} + z^{-2} + \ldots + z^{-n} + \ldots$$

$$= \tfrac{1}{2} + \sum_{n=1}^{\infty} z^{-n} = \tfrac{1}{2} - 1 + \sum_{n=0}^{\infty} z^{-n} \qquad \text{(IV-164)}$$

$$= \frac{z}{z-1} - \tfrac{1}{2} = \frac{z+1}{2(z-1)}$$

To derive the solution using this "ramp-step" function instead of a true step, the transformed partial difference equation IV-112 and the

adiabatic boundary condition equation IV-114 apply if a zero initial condition is assumed throughout. The boundary condition represented by equation IV-113 is modified by the substitution of equation IV-164 for $\bar{t}_f$ and ignoring the γ weighting for $\bar{t}_f$ (γ set to zero when it is a coefficient of $\bar{t}_{f,0}$):

$$\left[z-1+2r\left(1+\frac{H}{S}\right)(\gamma z+1-\gamma)\right]\bar{t}_0 - 2r(\gamma z+1-\gamma)\bar{t}_1 = \frac{2rH}{S}\left[\frac{z+1}{2(z-1)}\right] \qquad \text{(IV-165)}$$

Since the transformed partial difference equation and adiabatic boundary condition are unchanged, the solution for these two equations is written directly,

$$\bar{t}_m = \frac{P\cos(S-m)\alpha}{\cos S\alpha} \qquad \text{(IV-166)}$$

Using the previous procedure of substituting $\bar{t}_0$ and $\bar{t}_1$ as found from equation IV-166 into equation IV-165 and solving for $(P/\cos S\alpha)$, we find that

$$\bar{t}_m = \left[\frac{z+1}{2(z-1)}\right]\left[\frac{H\cos(S-m)\alpha}{(\gamma z+1-\gamma)\left\{H\cos S\alpha - S\sin\alpha\sin S\alpha\right\}}\right] \qquad \text{(IV-167)}$$

(This equation could be found directly by substituting equation IV-146 for $\bar{t}_f$ into equation IV-129, ignoring γ.)

This transform is the average of the transform for the true step function and that transform divided by z. Further, as $\bar{t}_m$ goes to a constant value as z goes to zero, $\bar{t}_m$ contains neither a pole nor zero at zero.

According to the previous discussion, a simple pole at zero is introduced when multiplying by z^{n-1}; and in order that the solution fit the initial conditions, it must be included. Further, by comparison of equation IV-167 with the previous transform, equation IV-131, the solution is found to be the average of the result for n and for (n-1) in addition to the residue of the zero singularity.

$$t_{m,n} = \tfrac{1}{2}\left\{2 - \frac{H}{S}\sum_{j=1}^{S+1} \frac{\left[1+\frac{1}{q_j}\right]\left[\cos(S-m)\alpha_j\right]\left[q_j\right]^n}{\left[\tan\frac{\alpha_j}{2}\right]\left[(H+\cos\alpha_j)(\sin S\alpha_j)+S\sin\alpha_j\cos S\alpha_j\right]}\right\} \qquad \text{(IV-168)}$$

$$-\tfrac{1}{2}\left\{\frac{[H\cos(S-m)\beta][0]^n}{[H\cos S\beta - S(\sin\beta)(\sin S\beta)][1-\gamma]}\right\}$$

$$m = 0, 1, \ldots, S$$

$$n \geq 0$$

where the q_j's and α_j are as defined previously by equations IV-136 and IV-160, respectively, and

$$\cos\beta = \frac{2r(1-\gamma)-1}{2r(1-\gamma)} \qquad \text{(IV-169)}$$

The solution equation IV-168 applies for all values of γ including a γ of 1, even though $(\cos\beta)$ becomes infinite for a γ of 1. The argument used previously to show that the term $(\gamma z+1-\gamma)$ in the transform in equation IV-131 does not give a pole can be used here to show that the β term has a finite limit for all values of γ including a value of

one. The necessity for adding the zero root is obvious, as a comparison of this solution, equation IV-168, with the previous one, equation IV-158, shows that if the first meets the initial conditions the second cannot unless another term is added. The above result has been checked numerically to make sure it does fit the initial condition. Since, as mentioned before, the zero root only gives a contribution for n of zero, we may neglect it in comparing the results. The effect of using the "ramp-step" is then essentially an averaging process; that is, the same results could be obtained by stepping out the solution with a true step and averaging the results at n and $(n-1)$ and using this average at n. This is an alternate proof of the relationship between averaged methods and the ramp-step replacement for a step change in fluid temperature. The factor $(1+1/q_j)$ actually modifies the Fourier coefficients in the initial vector and further studies would be required to determine any improvement in accuracy.

3. Problem II, Infinite H, Method G

For the case of infinite heat-transfer coefficient with a zero initial condition and a boundary at unity,

$$t_{m,0} = \delta_{m,0} \tag{IV-170}$$

The transformed partial difference equation IV-117 is no longer homogeneous; substituting for $t_{m+1,0}$, $t_{m,0}$, and $t_{m-1,0}$ from

equation IV-170 into equation IV-101 together with equation IV-115, and noting that $t_{m+1,0}$ and $t_{m,0}$ are zero for the range of m's used, this equation is:

$$\bar{t}_{m-1} - 2\cos\alpha\,\bar{t}_m + \bar{t}_{m+1} = \left[\frac{z\gamma}{(\gamma z+1-\gamma)}\right]\left[\delta_{m-1,0}\right] \qquad \text{(IV-171)}$$

$$m = 1, 2, \ldots, S$$

and the left boundary condition is

$$t_{0,n} = t_{f,n} \qquad \text{(IV-172)}$$

and the transformed conditions are:

$$\bar{t}_0 = \frac{z}{z-1} \qquad \text{(IV-173)}$$

$$\bar{t}_{S+1} = \bar{t}_{S-1} \qquad \text{(IV-174)}$$

Assuming the particular solution is of the form,

$$\bar{t}_{m,P} = F\,\delta_{m,0} \qquad \text{(IV-175)}$$

substituting into equation IV-171 according to the method of undetermined parameters:

$$F\,\delta_{m-1,0} - 2F\,\delta_{m,0}\cos\alpha + F\,\delta_{m+1,0} = \left(\frac{z\gamma}{\gamma z+1-\gamma}\right)\delta_{m-1,0} \qquad \text{(IV-176)}$$

$$m = 1, 2, \ldots, S$$

or, as $\delta_{m,0}$ and $\delta_{m+1,0}$ are zero for the m's used,

$$F = \left[\frac{z\gamma}{z\gamma+1-\gamma}\right] \tag{IV-177}$$

and

$$\bar{t}_m = P \cos m\alpha + G \sin m\alpha + \frac{z\gamma\delta_{m,0}}{[z\gamma+1-\gamma]} \tag{IV-178}$$

$$m = 0, 1, \ldots, S, S+1$$

Solving for P and G according to the previous procedure gives the complete transformed difference temperature as,

$$\bar{t}_m = \frac{z}{z-1}\left[\frac{\cos(S-m)\alpha}{(z\gamma+1-\gamma)(\cos S\alpha)}\right] + \frac{z\gamma\,\delta_{m,0}}{(z\gamma+1-\gamma)} \tag{IV-179}$$

$$m = 0, 1, \ldots, S+1$$

This is not the transform that would be obtained by taking limits of equation IV-131 as H or h goes to infinity; because of the change in initial conditions, the term involving the Kronecker delta function is added. However, the "main" transform could be found by taking limits.

Inversion may be carried out as before. The first term has simple poles at

$$z = 1 \tag{IV-180}$$

and at

$$z = q_j = \frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r(1-\gamma)(1-\cos\alpha_j)} \qquad \text{(IV-181)}$$

$$j = 0, \ldots, S-1$$

where α_j are the roots of

$$\cos S\alpha_j = 0 \qquad \text{(IV-182)}$$

or

$$\alpha_j = \frac{(2j+1)\pi}{2S} \qquad j = 0, \ldots, S-1 \qquad \text{(IV-183)}$$

Note that because $t_{0,n}$ is fixed there is one less degree of freedom, thus one less q_j and α_j. (The complex α_j fortunately is eliminated.) Now considering the pole in the first term at $(\gamma-1)/\gamma$, and using arguments like those for the case of finite h, we conclude that a zero exists at $z = (\gamma-1)/\gamma$ of the order of $(m-1)$. Consequently, only for an m of zero is there a simple pole at $(\gamma-1)/\gamma$, or, equivalently, only for an m of zero does this pole have a non-zero residue. However, its residue is the negative of the residue of the second term of the transform. Thus, this pole gives no contribution to the solution. Evaluating and summing the residues for the other poles we obtain (after trigonometric simplification)

$$t_{m,n} = 1 - \frac{1}{S} \sum_{j=0}^{S-1} \left[\frac{1}{\tan\frac{\alpha_j}{2}} \right] \left[\sin m\alpha_j \right] \left[\frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r\gamma(1-\cos\alpha_j)} \right]^n \qquad \text{(IV-184)}$$

where

$$\alpha_j = \frac{(2j+1)\pi}{2S} \tag{IV-185}$$

This result could have been obtained by taking the limit of equations IV-158 and IV-160, the characteristic and the solution equations, as H goes to infinity even though the transform might not be found correctly by taking limits. The solution is identical to that reported in O'Brien et al. (2) for the same problem solved by von Neumann's Method and a finite Fourier harmonic analysis. The eigenvectors and eigenvalues reported agree with the results from many workers (15, 10, 21, 17, 29, 22, 2, 3, 23, 24, 4, 30) as derived by separation of variables technique or matrix analysis.

4. Problem V, Semi-Infinite Solid, Method G--Branch Cut Singularity

The last type of example solution carried out here is that of applying method G to a semi-infinite solid, whose left boundary temperature is specified as unity and which extends to infinity in the ξ direction. The initial condition is taken as zero. Since an implicit method would require the solution of an infinite number of equations only the explicit case is considered. The transformed partial difference equation is then

$$\bar{t}_{m-1} - 2\cos\alpha\bar{t}_m + \bar{t}_{m+1} = 0 \tag{IV-186}$$

where

$$\cos\alpha = \frac{z-1+2r}{2r} \qquad m = 1, 2, \ldots, \infty \qquad \text{(IV-187)}$$

The transformed boundary conditions are:

$$\bar{t}_0 = \frac{z}{z-1} \qquad \text{(IV-188)}$$

$$\bar{t}_\infty = 0 \qquad \text{(IV-189)}$$

Assuming that the solution to the partial difference equation is equation IV-120 P and G may be determined from the above boundary equation giving

$$\bar{t}_m = \frac{z}{z-1}\left[e^{-m\alpha\sqrt{-1}}\right] \qquad \text{(IV-190)}$$

From complex variable theory the formula for the inverse cosine may be written:

$$\alpha = \cos^{-1}\left(\frac{z-1+2r}{2r}\right) = -\sqrt{-1}\ln\left[\frac{z-1+2r}{2r} + \sqrt{\left(\frac{z-1+2r}{2r}\right)^2 - 1}\right] \qquad \text{(IV-191)}$$

Substituting this into equation IV-190 and simplifying, we have,

$$\bar{t}_m = \frac{z}{z-1}\left[\frac{2r}{z-1+2r+\sqrt{(z-1)(z-1+4r)}}\right]^m \qquad \text{(IV-192)}$$

This function has, in addition to the pole at +1, a branch cut from (1-4r) to +1 on the real axis because of the square root of the two

factors (28). To invert the transform we integrate $z^{n-1}\,\bar{t}_m$ along a dumbbell-shaped contour in the z-plane. This contour is shown in Figure IV-5(b) for $r = \frac{1}{2}$; and the direction is counter-clockwise about the closed dumbbell indicated by the dashed line. (Actually the inversion integral is along ABC; but as $z^{n-1}\,\bar{t}$ is analytic within ABCDA, the inversion integral may be shown to be equal to that found by going counter-clockwise around the dumbbell.) The integrals are evaluated allowing the radius of the circles δ and distance from the real axis ϵ to go to zero.

The path around the pole and branch point +1 is given by

$$\Gamma_1 \qquad z = 1 + \delta e^{\sqrt{-1}\,\varphi_1} \qquad -\pi \le \varphi_1 \le \pi \tag{IV-193}$$

where φ_1 is allowed to vary from $-\pi$ to π. Substituting this into the transform and taking the limit and integrating we obtain

$$J_1 = \lim_{\delta \to 0} \int_{\Gamma_1} z^{n-1}\,\bar{t}_m\,dz = 2\pi\sqrt{-1} \tag{IV-194}$$

The path around the branch point at (1-4r) is

$$\Gamma_3 \qquad z = 1 + \delta e^{\sqrt{-1}\,\varphi_2} \qquad 0 \le \varphi_2 \le 2\pi \tag{IV-195}$$

Substituting and carrying out the integration gives

$$J_3 = \lim_{\delta \to 0} \int_{\Gamma_3} z^{n-1}\,\bar{t}_m\,dz = 0 \tag{IV-196}$$

The path for the line above the real axis is

$$\Gamma_2 \qquad z = x + \sqrt{-1}\,\epsilon \tag{IV-197}$$

and below is

$$\Gamma_4 \qquad z = x - \sqrt{-1}\,\epsilon \tag{IV-198}$$

After making the above substitutions and taking the required limits, the integrals for Γ_2 and Γ_4 may be combined. The result is, after simplification and trigonometric substitution:

$$J_2 + J_4 = \lim_{\epsilon \to 0} \int_{\Gamma_2+\Gamma_4} z^{n-1} \bar{t}_m \, dz = 2\sqrt{-1} \int_0^{\pi} \frac{(\sin m)[1-2r(1-\cos\alpha)]^n \, d\alpha}{\tan\frac{\alpha}{2}} \tag{IV-199}$$

The actual integrations are not shown as they involve a considerable amount of algebra, and are straightforward after the limits are taken. The inverse transform is $(1/2\pi\sqrt{-1})$ times the sum of the integrals:

$$t_{m,n} = 1 - \frac{1}{\pi} \int_0^{\pi} \frac{(\sin m)[1-2r(1-\cos\alpha)]^n \, d\alpha}{\tan\frac{\alpha}{2}} \tag{IV-200}$$

This result could be obtained from equations IV-184 and IV-185 by letting S go to infinity without $\Delta\xi$ going to zero. Briefly let

$$\alpha = \alpha_0 + j\Delta\alpha$$

$$\text{where } \alpha_0 = \pi/2S \tag{IV-201}$$

$$\text{and } \Delta\alpha = \pi/S$$

Now, noting from the characteristic equation IV-182 that as S goes to infinity any value of α between 0 and π satisfies the equation, or, equivalently, α becomes continuous, and taking limits on the above equations, we have

$$\lim_{S \to 0} \alpha_0 = 0 \qquad \text{(IV-202)}$$

$$\lim_{S \to \infty} \frac{1}{S} = \lim_{S \to \infty} \frac{\Delta\alpha}{\pi} = \frac{d\alpha}{\pi} \qquad \text{(IV-203)}$$

Substituting for the limit of 1/S and changing the sum to an integral we obtain the above result.

5. Conclusions

Most of the results presented and compared in following subsections were derived using the principles, techniques, and manipulations shown in deriving the previous results. The only major difference for methods using mesh $\Delta\xi/2$ is that the m's are 1/2, 3/2, ... S-1/2 for convenience in derivation and in comparisons; no other complications appear.

The z-transform method for solving partial difference equations was demonstrated above by finding the analytic solution for the difference approximation of the one-dimensional diffusion equation for several types of boundary and initial conditions. These analytic difference solutions are of the same form as the solution of the differential equation; thus, comparisons between the continuous solution and difference

approximations are readily possible. The z-transform method suffers the same restrictions as the Laplace transform method, and thus the difference solutions can be derived only for those cases where the continuous solution can be found. This is not a disadvantage in comparing several approximate methods to the exact continuous solution.

The z-transform technique has several advantages over the other methods for solving partial difference equations. It is a methodical procedure mostly consisting of algebraic manipulations that allow one to find the complete solution to the difference problem. The only "trick" in determining the solution is the substitution of α (equation IV-115); but all the other methods require this or an equivalent substitution also. This substitution is suggested by a form of the solution of the ordinary difference equation, if it is not made. The procedure has the advantage that it gives the particular or quasi-steady-state solution and the Fourier coefficients in the initial vector, in addition to the eigenvectors and eigenvalues. As demonstrated in the next subsection, the Fourier coefficient in the initial vector is one of the most important quantities in obtaining accurate approximations. The direct matrix method requires significant additional work just to determine the analytic expressions for the eigenvectors after the eigenvalues are found. The separation of variables technique determines the eigenvectors and eigenvalues in a way very much like that for the z-transform method. Even after these two quantities are found, the particular or quasi-

steady-state solution must be found. And possibly the most difficult step with these more conventional techniques is to determine the coefficients that fit the eigenvectors to the initial condition. This requires determining the orthogonality relationships and the analytic evaluation of a series of finite sums.

No rigorous proof has been presented that the z-transform solution of a partial difference equation is the same as the eigenvector-eigenvalue problem. Indeed, no proof has been presented that z-transforms are applicable to partial difference equations. However, the basis for the z-transforms has been presented and no mathematical inconsistency seems to have been introduced in applying them to the partial difference equations. The difference solutions derived have been compared with published results derived by other methods and in each case the transform results are identically the same for quantities that other researchers have obtained. Usually this is restricted to a comparison of the eigenvalues and sometimes eigenvectors for the case of infinite heat-transfer coefficient. The only comparison for a complete solution including the initial vector and particular solution was mentioned previously. Further, almost all of the results have been checked numerically. That is, the difference solution has been computed for fixed values of S, r, and γ, by stepping out the solution and from the z-transform solution. In all cases (after algebraic and numerical errors were removed) the two solutions checked except

for discrepancies which could be accounted for by round-off error (usually four-digit accuracy). Thus, the z-transform method is concluded to be a valid, useful procedure for studying the one-dimensional partial difference equation of diffusion.

In addition to the type of problems studied and compared here, there are many other applications where this method should be useful. Already it has been successfully applied to Richardson's (2) unstable explicit three-level approximation to the diffusion equation, where the temperature-time derivative is replaced with $(t_{m,n+1} - t_{m,n-1})/2\Delta\tau$ instead of the two-level formula used above. The other main application to the diffusion equation is to solve its partial difference equation for the radial direction in either cylindrical or spherical coordinates. Preliminary attempts to derive this solution for the radial coordinate in cylindrical coordinates have not been successful, but, in this author's opinion, this was caused by an inability to find the substitution equivalent to the relationship between z and α that gives simplification to the transformed difference equation. This in turn is caused by not having a sufficient familiarity with Bessel functions and their identities.

In addition to the study of the approximate methods for the diffusion equation, z-transforms should be of use in studying approximate methods for other equations which lead to a linear partial difference equation of an initial-value type with linear boundary conditions for which the exact solution may be found by the Laplace transform.

Specifically, two partial differential equations for which approximate solution methods might be studied are the wave equation and the equation governing transient bending of beams. The form of these are, respectively,

$$\frac{\partial^2 y}{\partial \theta^2} = \frac{\partial^2 y}{\partial x^2} \tag{IV-204}$$

and

$$\frac{EI}{m} \frac{\partial^4 y}{\partial x^4} + \frac{\partial^2 y}{\partial \theta^2} = 0 \tag{IV-205}$$

where $\frac{EI}{m}$ = flexural rigidity.

(E) COMPARISON OF METHODS--z-TRANSFORM SOLUTIONS

In order to compare the approximate methods just derived, and to demonstrate the utility of z-transforms in studying and understanding approximate methods, complete analytic solutions have been found for the approximate methods applied to several commonly encountered problems. As has been shown above, these approximate solutions have the same form as the continuous solution. A comparison of corresponding terms of the approximate and continuing solutions will now be made. Also, for one problem approximated by graphical methods, it is possible to allow for the interrelationship between the several terms in making a comparison. These results not only show us which are the more accurate methods, but also how to select the differencing parameters so that the approximate solution is sufficiently accurate. The problems considered are summarized in Table IV-9. The comparison of terms of the solution is in Tables IV-10 through IV-29.

The precise mathematical statements of the five problems for which analytic solutions for the approximate methods have been found are in Table IV-9. In problem I the solid has a zero initial condition throughout. The left boundary is in contact with a fluid which has a heat-transfer coefficient, h, and which undergoes a step increase in temperature from 0 to 1 at time zero and remains at 1 for all

succeeding times. The right boundary is adiabatic. Problem II is the limiting case of problem I as the heat-transfer coefficient h is increased to infinity so that the left surface undergoes the unit step increase in temperature and then remains constant at 1. In problem III the initial distribution increases linearly from 0 at the left boundary to 1 at the right. Both boundaries are adiabatic. The solutions for approximate methods G, A, and C are compared to the corresponding continuous solution for each of these three problems; the solution for graphical method F for problem II is also shown. The main conclusions concerning these methods are based on how accurately they approximate the continuous solution to these problems.

Problem IV has an initial temperature distribution of zero; the left surface temperature is maintained at zero and the right surface at one. Problem V has the same initial and left boundary conditions as II, but its right boundary is assumed to be located at a ξ of infinity. For these two problems only the solutions for method G are derived and compared to the continuous solution. Method G applied to problem IV was the first problem solved by z-transforms. It was selected because the eigenvalues and eigenvectors of the Y/A matrix for this case are published (29), and the z-transform results were subjected to an immediate (successful) check. Method G applied to problem V was used to show that the z-transform procedure is applicable to the case of a semi-infinite solid where the transform has a branch cut as a singularity.

The approximate and continuous solutions have the same form as shown in general by the vector equations II-17, II-43, and II-52, and specifically by the components of the solution vector for method G for problem I, equations IV-145, IV-146, and IV-161. These solutions are in difference form for problems I through IV:

Continuous:

$$T(m\Delta\xi, n\Delta\tau) = T_P(m\Delta\xi, n\Delta\tau) - \sum_{j=1}^{\infty} a_j b_j(m\Delta\xi)\left[e^{-\psi_j^2 r}\right]^n \qquad \text{(IV-206)}$$

$$n = 0, 1, \ldots, \infty$$

where ψ_j's are the roots of a characteristic equation, $F_C(\psi_j) = 0$.

Approximate:

$$t_{m,n} = t_{P\,m,n} - \sum_{j=1}^{S} g_j c_{mj} q_j^n \qquad n = 0, 1, \ldots, \infty \qquad \text{(IV-207)}$$

where

$$q_j = \frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r\gamma(1-\cos\alpha_j)} \qquad j = 1\ 2, \ldots, S \qquad \text{(IV-208)}$$

and where $g_j c_{mj}$, as well as q_j, depend upon the roots of the characteristic equation $F_D(\alpha_j) = 0$. For the analog solution the term q_j^n is replaced by $e^{-2S^2(1-\cos\alpha_j)\tau}$. Table IV-10 summarizes the solution equations, the corresponding quantities, and other notes.

Although in the continuous solution m and n (and n in the analog) can be treated as continuous variables in the above equations

for T_n and $t_{m,n}$ when comparing with an approximate method, both the continuous and approximate solutions are written only for the spatial points and values of n where the difference method gives results. Thus, n is only allowed to be non-negative integers, and when comparing methods G and A based on mesh $\Delta\xi$, the above equations are written for $m = 0, 1, \ldots, S$; and for methods C and F based on mesh $\Delta\xi/2$, $m = 1/2, 3/2, \ldots, S - 1/2$. The summation for the difference solutions is a finite sum over the number of degrees of freedom or number of points for which the temperature is calculated. The upper limit of the summation changes with the mesh and specific problem and the use of S above is to represent a finite sum.

The error in the approximate solution for the mth point at time n is found by taking the difference between the above two equations, IV-207 and IV-208, and is

$$v_{m,n} \equiv T(m\Delta\xi, n\Delta\tau) - t_{m,n} = T_P(m\Delta\xi, n\Delta\tau) - t_{P\,m,n}$$

$$+ \sum_{j=1}^{S} g_j c_{mj} q_j^n - \sum_{j=1}^{\infty} a_j b_j (m\Delta\xi) e^{-r\psi_j^2 n} \qquad \text{(IV-209)}$$

This error is made up of the error in the particular solution, $T_P(m\Delta\xi, n\Delta\tau) - t_{P\,m,n}$, and the error in the transient solution which is represented by the difference between the finite and infinite sums.

Because the fluid temperatures are constant with time and/or the boundaries are adiabatic for the five problems considered, the particular solutions are also constant and are discussed first. Actually no error occurs for the particular solution in these cases and the study is limited to the important and complicated transient solution. This solution is discussed in general followed by a detailed comparison of the characteristic equations, damping factors, eigenvectors, and initial vectors. Also in the section E-6 on initial vectors, a study of the inter-relationships of the oscillatory behavior for graphical methods is shown and averaged solutions are discussed. A numerical comparison of the analytic solution of graphical methods A and F applied to problem II is included under this topic and a special section on the analytic solution of problem V is presented.

The detailed summaries of the continuous and approximate solutions are in Tables IV-10 through IV-29; these are arranged by terms in the solutions and the complete analytic solution* is found by recombining the terms according to equations IV-206 and IV-207.

1. Particular Solutions

The particular or steady-state solutions for both the continuous and approximate methods derived are in Table IV-11. As the fluid

---

* Zero roots (q_j's = 0) are not included in these summaries unless they require no change in form.

temperatures in each of the five problems are constant with time, the particular solution, which is of the same form as the fluid temperature function, is not a function of time and therefore is a true steady state. The particular solution for every approximate solution found is identically equal to the continuous particular solution.

$$T_P - t_P = 0 \qquad \text{(IV-210)}$$

Further, for one-dimensional problems arguments based either on matrix solution of the system of difference equations or on the z-transform of the solution and the final value theorem, item 15, Table IV-7, can be used to show that methods G, A, and C have the same steady-state solution as the continuous solution. Thus, for problems when a true steady state occurs, i.e., where the fluid temperatures are constant and/or where the boundaries are adiabatic, the steady states for both the approximate and continuous solutions are equal.

Of more interest for possible future study would be a comparison of particular solutions for problems where the fluid temperature is a function of time. This would give a particular solution which is a quasi-steady-state solution, and one of the approximate methods might show an advantage in accuracy for these problems. Two types of fluid temperature functions which would be most interesting to study are: one which changes linearly with time, a ramp function, and one which changes sinusoidally with time. The problem for a sinusoid function

could be studied using a frequency response technique (13) on the z-transform without inversion; this frequency response technique is exactly analogous to that for Laplace transforms and ordinary differential equations. The final value theorem cannot be used to study these solutions where the quasi-steady-state solution changes with time, as it applies only when a final value or true steady state occurs.

Although a complete solution or frequency response study is necessary to determine the exact influence of the differencing parameters on the accuracy of particular solutions that change with time, some of the comments and conclusions made for the eigenvectors and eigenvalues of the transient solution are probably true in a general way for the particular solution. This can be shown by writing the vector solution for t_n when the boundary temperature vector is a function of time $t_{B,n}$ as

$$t_n = C\,Q^n\,C^{-1}\,t_0 + \sum_{p=1}^{n} C\,Q^{n-p}\,C^{-1}\,\frac{Y_B}{A}\,t_{B\,p-1} \qquad \text{(IV-211)}$$

instead of as

$$t_n = t_{P,n} + C\,Q^n\,C^{-1}\,(t_0 - t_{P,0}) \qquad \text{(IV-212)}$$

Thus, the summation in equation IV-211 which involves the eigenvectors and eigenvalues of the Y/A matrix contains the particular solution and part of the transient solution.

$$\sum_{p=1}^{n} C\, Q^{n-p}\, C^{-1}\, \frac{Y_B}{A}\, t_{B\,p-1} = t_P - C\, Q^{n}\, C^{-1}\, t_{P,0} \qquad \text{(IV-213)}$$

Since the continuous solution of the partial differential equation can be written in a form similar to equation IV-211, using Duhamel's Theorem where an integral replaces the sum, the selection of the differencing parameters so that the eigenvector matrix C and eigenvalue matrix Q are close to the corresponding quantities in the continuous transient solution also would be expected to result in an accurate particular solution.

Conclusions and Summary--Particular Solutions. For problems where the particular solution is a true steady-state solution, the approximate particular solutions for all the methods are equal to the continuous particular solutions. Additional studies are proposed for problems where the fluid temperature forcing functions give a quasi-steady-state solution, although arguments are presented that methods which give the most accurate transient solutions also probably give the most accurate particular solutions for this case.

2. Transient Solution

The transient solutions for the continuous and the approximate solutions are the infinite and finite summations, respectively, in equation IV-209. As mentioned before, the transient solution gives the exponential rate of decay to the final steady state, and because of

the nature of the problems studied, the major conclusions and comparisons are for the transient solution. In this section the transient solution is discussed in general in order to introduce detailed discussions and to show the specific goals of those discussions.

Each term in the transient solution is a product of an initial vector component a_j or g_j times the mth component of an eigenvector $b_j(m\Delta\xi)$ or c_{mj} times an exponential $e^{-\psi_j^2 rn}$ or exponential form q_j^n. The initial vector component a_j or g_j is the weighting for the jth eigenvector-exponential product. The mth component of the eigenvector $b_j(m\Delta\xi)$ or c_{mj} gives the relative weighting for the jth exponential at the mth point. The jth exponential function $e^{-\psi_j^2 rn}$ or the exponential form q_j^n determines the rate at which the jth term goes to zero with time or n. The continuous quantities a_j, $b_j(m\Delta\xi)$, and $e^{-\psi_j^2 rn}$ depend upon the jth root ψ_j of a characteristic equation $F_C(\psi_j) = 0$ for the continuous problem; the approximate quantities g_j, c_{mj}, and q_j depend upon the jth root α_j of the characteristic equation $F_D(\alpha_j) = 0$ for the approximate problem. Also, the summations are ordered so that, as j increases, both ψ_j and α_j increase. This means that the lower-subscripted terms are the slowest decaying terms for the continuous solution; $e^{-\psi_1^2 r}$ corresponds to q_1 or q_{MAX}. The higher-subscripted terms for the continuous solution decay rapidly. The higher-subscripted terms for the approximate solution are the ones that contain the negative q_j's, if such are allowed by the selection of the differencing

paramcters; the minimum q_j is q_S or q_{S+1}, depending upon the method and the problem, and is usually designated q_{min}. Each of the first S terms of the continuous transient solution is considered to correspond to the term in the approximate solution with the same subscript. For convenience, the quantities q_j and $e^{-\psi_j^2 r}$ which are raised to the nth power in the solutions are called damping factors in this study, as the term eigenvalue applies only to q_j, and exponential function only to $e^{-\psi_j^2 r}$.

The error in the transient solution for the mth point at the nth time increment is the difference between the infinite and finite summations in equation IV-209, and because no error occurs in the particular solution for the five problems studied, this is the total error:

$$v_{m,n} = \sum_{j=1}^{S} g_j c_{mj} q_j^n - \sum_{j=1}^{\infty} a_j b_j(m\Delta\xi)\, e^{-\psi_j^2 rn} \qquad \text{(IV-214)}$$

By combining corresponding terms with the same subscripts from both summations and assuming that terms containing products like $(g_j - a_j)[c_{mj} - b_j(m\Delta\xi)]$ are negligible, this error is approximately:

$$v_{m,n} \simeq \sum_{j=1}^{S} \left[(g_j - a_j)c_{mj} + g_j\left\{c_{mj} - b_j(m\Delta\xi)\right\} + g_j c_{mj}\left\{1 - \left(\frac{e^{-\psi_j^2 r}}{q_j}\right)^n\right\}\right] q_j^n$$

$$- \sum_{j=S+1}^{\infty} a_j b_j(m\Delta\xi) e^{-\psi_j^2 rn} \qquad \text{(IV-215)}$$

The second summation, which is from (S+1) to infinity, occurs because the difference system has a finite number of roots for a finite number of degrees of freedom, but the continuous solution has an infinite number of roots for an infinite number of degrees of freedom. This summation contains the most rapidly decaying exponentials, and for any values of S a time sufficiently large exists when this summation is negligible; however, for short times, some of the terms in this summation are significant and cause significant error. The finite summation in equation IV-215 represents the error caused by the fact that the initial vector components, eigenvector components, and damping factors for the approximate solution are not equal to the corresponding factors in the continuous solution. The accuracy of the transient solution for each of the methods is studied by comparing these corresponding factors; that is, a separate section is devoted to each of the quantities, the damping factors ratio $e^{-\psi_j^2 r} / q_j$ in E-4, the difference in eigenvector components $(c_{mj} - b_j(m\Delta\xi))$ in E-5, and the difference in initial vector components $(g_j - a_j)$ in E-6. The form of equation IV-215 used here is convenient because expressions for the ratios and differences shown can be easily derived from the complete analytic continuous and approximate solutions. Since each of these quantities depends upon the roots ψ_j and α_j of the characteristic equations, these roots are discussed in the next section. However, before starting these detailed comparisons, some additional comments on the behavior of the transient

solution and on the requirements for accurate solutions are to be made.

As an example of how terms in the transient solutions behave, refer to Figure IV-6, where the first two terms (j=1 and j=2) in both the continuous solution and an approximate solution for a temperature point at a ξ of 0.2 for problem II are plotted versus time on semi-logarithmic paper. The approximate solution terms shown are for graphical method A with five points (S=5). Two vertical scales are shown, one on the left, a logarithmic scale, representing the term in the transient solution; the other on the right, an arithmetic scale, representing the natural logarithm of the term. Two horizontal scales are also shown, one for the time τ and one for the time increment n. Each of the four terms decreases linearly from an intercept of $a_j b_j(m\Delta\xi)$ or $g_j c_{mj}$ on the logarithmic scale at time zero. The slopes of these lines, based on the scale for the natural logarithms of the term and the time (τ) scale, are $-\nu_j^2$ for the continuous solution and the natural logarithm of $q_j^{1/\Delta\tau}$ for the approximate solution; if the slopes are based on the n-scale, the slopes are the logarithms of the damping factors or $(-\nu_j^2\Delta\tau)$ for the continuous solution terms and $(\ln q_j)$ for the approximate solution terms. The differences in intercepts of the terms are caused by the errors in the initial vector and eigen-vector components which are, from equation IV-215,

$$g_j c_{mj} - a_j b_j(m\Delta\xi) \simeq (g_j - a_j)c_{mj} + g_j\left\{c_{mj} - b_j(m\Delta\xi)\right\} \qquad \text{(IV-216)}$$

Since the $g_j c_{mj}$ product for an unaveraged method is affected only by the differencing parameters that are involved in the Y/A matrix, the error in the intercept for an unaveraged solution depends on the method used and the number of points for methods G, C, and graphical A. For averaged methods and generalized method A the intercept is also a function of the time increment and weighting γ; and for the averaged methods it is increased or decreased, depending upon the time at which the average is applied. The slope of the line for the approximate solution depends on the method, number of points S, and the time differencing parameters $\Delta\tau$ and γ, and it is unaffected by averaging. Also, this graph indicates that the second term of either solution decays rapidly, so that for this problem at an n of 7 or dimensionless time of 0.14, it is less than 4 per cent of its original value and about 5 per cent of the first term. At a dimensionless time of 0.2, it is insignificant compared to the first term; therefore, at long times only the first term (j=1) in the summation for the error is significant, and, if a true steady state exists, a plot of the solution on semi-logarithmic paper is linear with time for these long times. This fact should be used to extrapolate a stepped-out solution to very long times, and q_{MAX} can be calculated from the slope of this portion of the graph, if desired.

However, for short times more terms in the continuous transient solution are significant, and if accuracy is required for these times

the corresponding terms in the approximate solution must be good estimates of the terms in the continuous solution. That is, for example, if J terms are significant at time of interest τ_0, the first J terms of the approximate solution must be good approximations to the corresponding terms in the continuous solution. (In some cases very accurate approximate solutions are obtained at several low values of n because the approximate solution crosses the continuous solution near these values of n; however, this type of crossing is not predictable.) Further, Figure IV-6 shows that neither the intercept nor the slope for the second term (j=2) of the approximation is as close to those of the continuous second term as for the first term (j=1).

These facts, although shown only for one approximation, are typical of all approximate transient solutions and can be used to sketch a procedure to select the differencing parameters intelligently. If the maximum error that can be tolerated in the approximate calculation is V, at time τ_0 or increment n_0, and for all later times, then we must have:

$$V \geq |v_{m,n}| \cong \left| \sum_{j=1}^{J} \left[(g_j - a_j) c_{mj} + g_j \left\{ c_{mj} - b_j (m\Delta\xi) \right\} + g_j c_{mj} \left\{ 1 - \left(\frac{e^{-\psi_j^2 r}}{q_j^{1/\Delta\tau}} \right)^{\tau_0} \right\} q_j^n \right] \right| \qquad \text{(IV-217)}$$

where only the first J terms are significant in the continuous solution. Although in principle each of the J terms in the expression for the

error $v_{m,n}$ could be estimated and used to select the approximating parameters, this is not practical if J is more than one. Instead, an upper bound for the sum in equation IV-217, or equivalently $v_{m,n}$, is found by determining an upper bound for the absolute value of each term in the sum. The largest q_j^n in the summation occurs for a j of 1, and from the above comparison of terms subscripted 1 and 2 the largest coefficient of q_j^n occurs for the largest j or J. Therefore, each term in the summation must be less than a product of these quantities and the summation of J terms is less than J times the product. The inequality in equation IV-217 can then be conservatively written as

$$V = J\left|(g_J - a_J)c_{mJ} + g_J\left\{c_{mJ} - b_J(m\Delta\xi)\right\} + g_J c_{mJ}\left\{1 - \left(\frac{e^{-v_J^2}}{q_J^{1/\Delta\tau}}\right)^{\tau_0}\right\}\right| q_1^n \geq |v_{m,n}| \tag{IV-218}$$

If the contribution to the error from the difference in intercepts is taken equal to that from the slope at time τ_0, we have from equations IV-216 and IV-218,

$$\frac{V}{2Jq_1^{n_0}} \geq \left|(g_J - a_J)c_{mJ} + g_J\left\{c_{mJ} - b_J(m\Delta\xi)\right\}\right| \simeq \left|g_J c_{mJ} - a_J b_J(m\Delta\xi)\right| \tag{IV-219}$$

and

$$\frac{V}{2Jq_1^{n_0}} \geq \left|g_J c_{mJ}\left\{1 - \left(\frac{e^{-v_J^2}}{q_J^{1/\Delta\tau}}\right)^{\tau_0}\right\}\right| \tag{IV-220}$$

By selecting the error in the intercept equal to that in the slope, if the terms have opposite signs, the actual error at time τ_0 is much less than the bound V, but if they are of the same sign, the error is within the requirements. Now since the intercept error is a function primarily of S for all methods, the first restriction, equation IV-219, can be used to fix the number of points; then after the points are selected, the time differencing parameters $\Delta\tau$ and γ are fixed so that the second restriction is satisfied. Although the above equations should apply for any values of J, they are particularly useful when J is 1 or probably not more than 5.

However, the time differencing parameters must also be selected so that the oscillatory effects caused by any negative approximate damping factors do not ruin the accuracy of the solution. This requires that the largest oscillatory term be insignificant compared to the smallest significant term in the continuous transient solution, or equivalently:

$$\min_{1 \le j \le J} |g_j c_{mj} q_j^{n_0}| \gg \max_{J < j \le S} |g_j c_{mj} q_j^{n_0}| \qquad \text{(IV-221)}$$

The quantity on the left of the inequality sign is the term in the approximate solution that has the minimum absolute value of the first J terms in the sum; the quantity on the right of the inequality is the term that has the maximum absolute value of the remaining terms subscripted (J+1) to S. A more useful criterion is to make the contribution from

the maximum oscillatory term much smaller than the error bound.

$$V >> g_S \, c_{mS} \, Y^{n_0} \tag{IV-222}$$

where Y is an upper bound for $|q_{min}|$. This oscillatory contribution should probably be no larger than 1/10 of V and possibly should be smaller than 1/100 of V. Two ways of minimizing the oscillatory effects are by making Y^{n_0} small by the selection of the time differencing parameters and/or by making the amplitude of the oscillations small, either by using a method which has a small $g_S \, c_{mS}$ or by using a technique such as averaging to reduce this amplitude. The selection of $\Delta\tau$ and γ to obtain a given Y was discussed thoroughly in Chapter III, using matrix techniques, although a brief section is included here as part of the discussion of damping factors as found by z-transforms. However, one of the conclusions in that section is that the matrix technique is not only easier to use, but also gives an Y that is closer to q_{min}; therefore, the methods of Chapter III are better for studying stability. The amplitude of q_{min} is also considered, and the averaged methods are studied not only from the standpoint of oscillatory behavior, but also for their effect on the accuracy of the intercept for the larger positive damping factors. Because the parameters used for graphical solutions are so close to the limiting conditions for stability, the restriction in equation IV-222 does not apply directly to graphical solutions; however, a similar restriction concerning

oscillatory behavior is required, and this is discussed in the section on damping factors. Not all the methods are suitable for graphical solutions because they do not meet the requirement, but graphical method A is shown to be capable of giving accurate solutions. If the assumption is made that a graphical method does meet the oscillatory requirement, then as both γ and r are fixed, S must be increased until both the accuracy requirements, equations IV-219 and IV-220 are satisfied. (For fixed r increasing S reduces $\Delta\tau$.)

In selecting the time differencing parameters consideration must also be given to minimizing the number of calculations to the maximum time τ_1. If a true steady state exists this time τ_1 need never be much greater than the time when only one term is significant in the transient solution which is, for problem II, about 0.2. To find the solution for longer times a semi-logarithmic extrapolation should be used. For other problems, this time can be estimated from q_1 and q_2, from a trial calculation with a coarse differencing grid, or from the calculation itself. The number of non-zero multiplications is proportional to the number of points and the time τ_1, and inversely proportional to the time increment for an explicit calculation or an implicit calculation where the tridiagonal solution described in Chapter III is used. The number of multiplications is equivalently inversely proportional to r, the cube of the number of points, S^3, and τ_1.

$$N = \frac{P S \tau_1}{\Delta \tau} = \frac{P S^3 \tau_1}{r} \qquad \text{(IV-223)}$$

where N = number of non-zero multiplications and P = proportionality constant which depends upon r and γ and how the vector-matrix product $[I + (1-\gamma)\Delta\tau(Y/A)]\, t_n$ is found.

If this vector-matrix product is found in the most efficient way by making use of the fact that the off-diagonal elements in a row are equal, P is 1 for a graphical solution, 2 for all other explicit calculations ($\gamma = 0$), and 7 for any implicit calculation ($\gamma \neq 0$). If the calculation of the vector-matrix product is carried out by a vector-matrix subroutine each value of P is increased by 1 giving 2 for graphical, 3 for any other explicit, and 8 for implicit calculations. In practice, differencing parameters which give the minimum number of multiplications, and which meet the restriction for accuracy, are never found but some rules are presented that allow this minimum to be approached. Assuming the most efficient calculation, these equations do indicate that, for an implicit method to give the smallest N, it must allow an $S/\Delta\tau$ or S^3/r which is less than 1/7 times that for a graphical solution or less than 2/7 times that for an explicit method. This means that no other explicit calculation except the graphical solution need be considered if the graphical solution meets the restrictions with the same number of grid points as does any other explicit calculation.

In the detailed discussion and comparison of the trigonometric roots, the damping factors, eigenvectors, and initial vectors of approximate methods, the following points are emphasized as leading to a thorough understanding of the approximate solutions and to the selection of differencing parameters that give the required accuracy with a small amount of calculation.

(1) A comparison of the methods is made to determine which are the most accurate for any selection of approximating parameters and which are the most appropriate for graphical solutions. Included in this is the effect of oscillatory roots and a study of averaging.

(2) The second major consideration is to present ways of simply estimating the quantities in the inequalities for obtaining accuracy, equations IV-219, IV-220 and IV-222, and showing a methodical procedure for selecting the differencing parameters that give this accuracy with small although probably not minimum effort. This requires presenting an expression for the difference in the intercepts, $[g_J c_{mJ} - a_J b_J(m\Delta\xi)]$, and the slopes $(-\nu_J^2 - \ln q_J^{1/\Delta\tau})$ of the Jth terms, as functions of the differencing parameters which will apply in all cases. It also requires obtaining estimates of the damping factors or, equivalently, their trigonometric roots for either an approximate or continuous solution to determine the number of significant terms J at the earliest time of interest τ_0.

3. Characteristic Equations and Roots

The characteristic equations are summarized for problem I in Table IV-12, and for problems II, III, and IV in Table IV-13. (Problem V does not have a characteristic equation as such.) For both continuous and approximate methods, the characteristic equations for problems II and III can be found from the characteristic equation for problem I by taking the limits as H approaches ∞ or 0, respectively. That this is true for the difference method is seen by noting that the eigenvalues of the matrix are continuous functions of its elements. Thus the solutions of problems II and III are developed and discussed here as special cases of I. Problem IV, with both H's infinite, is a different problem and is discussed separately.

The characteristic equations in Tables IV-12 and IV-13 can be thought of as equation II-47 with a trigonometric substitution, equation IV-208, for q. For the difference methods G and C and graphical method A the Y/A matrix is not a function of γ or $\Delta\tau$; thus these characteristic equations apply for all values of γ and $\Delta\tau$ including $\Delta\tau = 0$ for the analog solution, and are independent of the initial condition or how the fluid temperature changes with time. However, when the Y/A matrix is a function of $\Delta\tau$ or r for a fixed S its characteristic equation is a function of r; this is true for generalized method A. The approximate characteristic equations apply also to any problem where the Y/A matrix is the same as for problem I; this includes problems where, in addition

to the heat flux from the heat-transfer coefficient, a heat flux specified as a function of time but independent of the temperature is present at the boundary.

The difference between the characteristic equations for the approximate methods shown in Table IV-12 for problem I and the characteristic equation for the continuous solution is that the factor $(S\psi)$ in the continuous case is replaced with factors such as $(S \sin\alpha)$, $(S \tan\alpha)$, or $[2S \tan(\alpha/2)]$, which differ from $(S\alpha)$ only slightly for small angles. As the characteristic equations themselves are transcendental and the roots can be obtained in analytic form only for the special cases when H is either zero or infinity, the best way of gaining a quantitative understanding of the effect of the replacement is by defining for each characteristic equation shown in Table IV-12 a function $H(\psi)$ or $H(\alpha)$. A plot of $H(\psi)$ or $H(\alpha)$ versus the angle parameter is constructed for an assumed S. From these curves the required roots can be found for any H and a comparison of the behavior of the equations is possible. Although the results from this study apply only to a given value of S, the conclusions are generalized qualitatively to other values of S by a series expansion of the replacement function. First the graphs for $H(\psi)$ and $H(\alpha)$ are constructed. This is followed by the series expansion to explain and generalize the results.

The graphs for the characteristic equations of the continuous solution and methods G, C, and graphical A are in Figures IV-7

through IV-12, for an S of 5 and angles of 0 to 2π radians. Since generalized method A would require a graph for several r's it is not shown, but it is discussed briefly later. Before discussing each graph in detail several points in common can be shown. For an H of zero (problem III) each of the characteristic equations becomes

$$\sin S\alpha_j = \sin S\psi_j = 0 \tag{IV-224}$$

which is satisfied by

$$\psi_j = \alpha_j = \frac{(j-1)\pi}{S} \qquad j = 1, \ldots, \infty \tag{IV-225}$$

For an infinite H (problem II) by taking limits the equations become

$$\cos S\psi_j = \cos S\alpha_j = 0 \tag{IV-266}$$

which is satisfied by

$$\psi_j = \alpha_j = (2j-1)\pi/2S \qquad j = 1, \ldots, \infty \tag{IV-227}$$

Thus each function is zero at an even multiple of $\pi/2S$, including zero, and infinite at an odd multiple of $\pi/2S$.

The continuous function in Figure IV-7 shows that $H(\psi)$ rises with increasing slope from 0 at a ψ of 0 to ∞ at a ψ of $\pi/10$. Between ψ's of $\pi/10$ and $\pi/5$ the function $H(\psi)$ is negative and is not shown. At a ψ of $\pi/5$, this function $H(\psi)$ becomes zero and again rises to infinity at a ψ of $3\pi/10$. This behavior is repeated so that an infinite sequence of such curves is obtained, each curve falling within an interval of π/S, or S roots falling in each interval of π radians. Thus an infinite number

of positive ψ_j's satisfy the characteristic equation for the continuous solution. The shape of each curve in the sequence is different. If ψ_j satisfies this continuous characteristic equation, $-\psi_j$ does also; however, the form of solution is written so that only the positive ψ_j's are used. The most significant roots are for the low ψ's and therefore ψ_1 is taken as the root for the curve between 0 and $\pi/10$, ψ_2 as the root between $\pi/5$ and $3\pi/10$, etc.

Figures IV-8 and IV-9 for the $H(\alpha)$ for methods G and C show a sequence of curves directly analogous to that for the continuous $H(\psi)$ for α's between 0 and π. That is, each curve rises from an H of zero at an even multiple of $\pi/10$ to infinity at an odd multiple of $\pi/10$ with a shape like that for the continuous equation. However, these graphs are symmetric about π radians, i.e., between π and 2π the curves are mirror images of those between 0 and π. This is shown on the graphs by a scale going from π to 2π as one moves from right to left on the top of the graph. Thus, an infinite number of roots also satisfy the difference equation; however, because of the symmetry of location of these roots and the requirement from matrix theory for independent eigenvectors only those α's in an interval of π radians are used. This interval is conveniently selected as zero to π. Thus, these graphs locate the required five roots for method C and five of the six roots for method G. If H is infinite, then method G has only five degrees of freedom and all five roots are shown.

The sixth root for method G when H is not infinite is found by allowing it to be complex and of the form

$$\alpha_6 = \pi + \sqrt{-1}\, \mathcal{J}(\alpha_6) \qquad \text{(IV-228)}$$

Substituting this into the characteristic equation for method G from Table IV-12 and using complex variable theory, the equation is modified to

$$5[\sinh \mathcal{J}(\alpha_6)]\,[\tanh 5\,\mathcal{J}(\alpha_6)] = H \qquad \text{(IV-229)}$$

A plot of this as a function of α is in Figure IV-10. The function $\{5[\sinh \mathcal{J}(\alpha_6)]\,[\tanh 5\,\mathcal{J}(\alpha_6)]\}$ rises from 0 at an $[\mathcal{J}(\alpha_6)]$ of zero to infinity at an $[\mathcal{J}(\alpha_6)]$ of infinity. Thus, for both H's of zero the sixth root is π and contains no complex part. Even though α_6 is complex for all other H's between zero and infinity, its corresponding eigenvector, damping factor, and initial vector component are real, as indeed they are known to be from matrix arguments. Since this is the (S+1) root, it is the one that affects stability and oscillatory behavior, and it is the reason why method G has a more severe stability restriction. Thus, for a good approximation, the effect on the solution of this root should be minimized.

The graph for graphical method A is in Figure IV-11. Here the curves for the first three roots up to $\pi/2$ show behavior like those for methods G and C and the continuous method, and the zero points and asymptotes are as before. However, the curves between $\pi/2$ and π

are mirror images of those from 0 to $\pi/2$. That is, as α goes from $\pi/2$ to $3\pi/5$, the function $H(\alpha)$ drops from infinity to zero and, although not shown, is negative from $3\pi/5$ to $7\pi/10$; the other methods and the continuous solution rise from an H of zero to infinity as the angle goes from $3\pi/5$ to $7\pi/10$. Thus, for graphical method A the ψ and α do not agree at all accurately for the higher subscripted roots. However, the accuracy obtained with graphical method A is actually caused by this symmetry or reversal at $\pi/2$, as explained later. This reversal is caused by the sign change in $(\tan\alpha)$ at $\pi/2$ which does not occur at $\pi/2$ for the corresponding factors $(\sin\alpha)$, $[2\tan(\alpha/2)]$, or ψ. It is seen that all six roots for graphical method A can be found from this graph and are real. The graph shown is for S odd; if S is even, an α of $\pi/2$ is a root for all values of H and the symmetry about $\pi/2$ is maintained.

From these graphs and comments it is seen that, in general, for real ψ's and α's, $H(\psi)$ and $H(\alpha)$ are made up of S sections of curves between 0 and π. Further, as S becomes large the asymptote for ψ_1 and α_1 approaches zero; therefore, ψ_1 and α_1 become small. Also the asymptote for ψ_S and α_S approaches π and this root also approaches π.

In later comparisons for the damping factors and for components of the initial vector, a difference parameter related to α, as the continuous parameter ψ is related to υ (equation IV-153), is more

convenient to use than α. It is defined by:

$$\mu = \alpha S \qquad \text{(IV-230)}$$

The graphs just discussed for the characteristic function of α could easily be modified to be of μ by multiplying the α scale by S, which is 5 for the graphs shown. Thus, H would rise from zero for a μ that is a multiple of π (including 0) to approach infinity as μ approaches a multiple of $\pi/2$. The graphs for the difference methods would then be symmetric about $S\pi$. The advantage of defining and using μ is that as the roots ν_j are not a function of S and as the μ_j correspond to the ν_j, the order of magnitude of the μ_j is not a function of S. Note, however, that the exact value of each μ_j is a function of S unless the problem and approximate method is one of the special cases where the S roots for the approximate solution are equal to the first S roots for the continuous solution.

In order to quantitatively show the deviation in these roots, residual graphs of $(\psi-\alpha)$ for S of 5 are shown in Figures IV-12 and IV-13 for the several roots. Note that the scale for $(\psi-\alpha)$ is different on the different graphs. For methods G, graphical A, and C, the quantity $(\psi_1-\alpha_1)$ is zero for zero H and its absolute value increases to a maximum at an H of about 2 to 3; then it approaches zero asymptotically as H goes to infinity. Graphical method A shows the largest deviations in absolute value and these are positive. Method C's deviations are also positive, but they are only about $\frac{1}{4}$ those of graphical

method A. The first root for method G shows negative deviations which are about twice as much in absolute value as those for C. The graphs for $(\psi_2-\alpha_2)$, $(\psi_3-\alpha_3)$, ..., $(\psi_5-\alpha_5)$ show the same behavior; however, the order of magnitude of the deviations increases by a factor of 20 for root 2 and again by a factor of 8 for root 3, with even larger deviations for roots 4 and 5.

These results can be explained by the following manipulations. First let D be the factor in a characteristic equation for an approximate method that replaces $S\psi$ in the characteristic equation for the continuous solution. Then, combining the continuous characteristic equation and the approximate characteristic equation written in terms of D and υ and μ instead of ψ and α, and using straightforward algebraic manipulations, there is obtained:

$$(\upsilon_j-\mu_j)(\upsilon_j+\mu_j) = (\upsilon_j^2-\mu_j^2) = \mu_j(D_j-\mu_j)\left[\frac{\mu_j^2-\upsilon_j^2}{\mu_j(\mu_j-\upsilon_j \tan\upsilon_j \operatorname{ctn}\mu_j)}\right] \qquad \text{(IV-231)}$$

$$j = 1, 2, \ldots, S$$

or, expanding the second bracketed term using trigonometric series, which shows that $(\mu_j^2-\upsilon_j^2)/\mu_j$ is a factor of $(\mu_j-\upsilon_j \tan\upsilon_j \operatorname{ctn}\mu_j)$ we obtain:

$$(\upsilon_j-\mu_j)(\upsilon_j+\mu_j) =$$

$$\mu_j(D_j-\mu_j)\left[1-\frac{\upsilon_j^2}{3} - \frac{\upsilon_j^2}{45}(\mu_j^2+\upsilon_j^2) - \frac{\upsilon_j^2}{945}(2\mu_j^4-5\upsilon_j^2\mu_j^2+2\upsilon_j^4\cdots\right] \qquad \text{(IV-232)}$$

The expanded term depends only slightly on S because the order of magnitude of μ_j does not change with S, and v_j, which is only a function of H, could be substituted for μ_j with little error. Therefore, this term is like a proportionality constant and consequently the difference between the roots, either $(v_j^2-\mu_j^2)$ or $(v_j-\mu_j)$ is approximately proportional to the quantity $(D_j-\mu_j)$. These expansions for these quantities for the methods are below:

Method

G $$D_j-\mu_j = S \sin\frac{\mu_j}{S} - \mu_j = -\frac{\mu_j^3}{6S^2} + \frac{\mu_j^5}{120S^4}\cdots \qquad \text{(IV-233)}$$

$$\mu_j < \alpha$$

A $$D_j-\mu_j = \frac{S \sin\frac{\mu_j}{S}}{1-2r(1-\cos\frac{\mu_j}{S})} - \mu_j \qquad \text{(IV-234)}$$

$$= \frac{\mu_j^3}{6S^2}(6r-1) + \frac{\mu_j^5}{120S^4}(1-30r\{1-4r\})\cdots$$

Graphical A $r = 1/2$

$$D_j-\mu_j = S \tan\frac{\mu_j}{S} - \mu_j = \frac{\mu_j^3}{3S^2} + \frac{2\mu_j^5}{15S^4}\cdots \qquad \text{(IV-235)}$$

$$|\mu_j| < \frac{S\pi}{2}$$

$$C \qquad D_j - \mu_j = 2S \tan \frac{\mu_j}{2S} - \mu_j = \frac{\mu_j^3}{12S^2} + \frac{\mu_j^4}{120S^4} \cdots \qquad \text{(IV-236)}$$

$$|\mu_j| < S\pi$$

The coefficient of the first term, μ_j^3, in these expansions explains the relative size and direction of the deviations of the first several roots (low values of μ_j or α_j). That is, the roots for method G show a negative deviation in absolute value about twice that for method C and its coefficient is $-1/6S^2$ which is negative and twice the coefficient for method C which is $1/12S^2$. Likewise, graphical method A, which showed positive deviations four times those of method C, has a positive coefficient which is four times that of method C. Further, these series allow the important conclusion that the difference between the roots ν_j and μ_j or between ν_j^2 and μ_j^2 decreases with the square of the number of points.

The probable reason that method C gives the most accurate trigonometric roots is related to the fact that it uses mesh $\Delta\xi/2$ with the points located away from the surface. As pointed out in section B-4 of this chapter, this means that the approximate boundary equations represent a direct discretization of the continuous boundary equation as shown by equation IV-32; no heat capacity is associated with the surface as in the continuous boundary equation; no heat capacity of the solid is neglected, and the Y/A matrix is symmetric. Thus, method C

has the most accurate trigonometric roots probably because it is based on mesh $\Delta\xi/2$ which leads to an approximate formulation that is close mathematically and physically to the continuous problem.

Although a series expansion of $(\upsilon_j^2-\mu_j^2)$ can be found by multiplication of the series for $(D_j-\mu_j)$ with the series in equation IV-232, the resulting series does not give a useful representation of that difference. This is true because the series in equation IV-232 is only very slowly convergent for most values of μ_j and υ_j as it involves the series for the cotangent and tangent of these quantities. Moreover, the series does not relate the difference in roots directly to the dimensionless heat-transfer coefficient H.

Since the quantities $(\upsilon_j^2-\mu_j^2)$ and $(\upsilon_j-\mu_j)$ appear in the comparisons of the damping factors and initial vectors, and since, for the above reasons, a useful analytic expression for these quantities probably cannot be developed, a numerical determination of the first several of the roots for both method C and graphical method A would be useful. The result would be a graph showing the deviation for the smallest two or three roots as a function of H for each of several S's.

A possible reason why the α's and ψ's are not equal when H is between zero and infinity is that the continuous boundary condition is actually in a difference-derivative form, and only the continuous derivative part is approximated. For the limiting case of H's of infinity or zero the boundary condition is not of this mixed form and the

trigonometric parameters, α_j and ψ_j or υ_j and μ_j, are identical for the number of difference roots.

Because the initial vector component, the eigenvector, and the damping factor are all functions of the trigonometric roots, a method where the first roots were equal might be desirable

$$\mu_1 = \upsilon_1 \tag{IV-237}$$

or

$$\psi_1 = \alpha_1 \tag{IV-238}$$

A comparison of the results from such a method with one of the above methods would also give an indication as to the importance of this deviation in roots. Two methods for which the roots are equal are possible. The first would be to use the value of H that would give the equality of the first roots rather than the H given for the problem. This procedure could be applied to any of the above methods, but it probably would change the components of the initial vector so much that the accuracy would be decreased. The second method utilizes the fact that the Y/A matrix for generalized method A is a function of r; consequently its replacement factor can be set equal to ψ_1:

$$\frac{\sin\psi_1}{1-2r(1-\cos\psi_1)} = \psi_1 \tag{IV-239}$$

Solving this equation for r gives:

$$r = \frac{\psi_1 - \sin\psi_1}{\psi_1(1-\cos\psi_1)} = \frac{1}{6} + \frac{\psi_1^2}{180} + \frac{\psi_1^4}{5040} + \cdots \qquad \text{(IV-240)}$$

A numerical study would be necessary to construct a graph for the r that solves the equation versus H. From the series expansion in equation IV-240, one sees that by making r = 1/6 the angles can be made almost equal for any value of H. Since using the value of r of 1/6 for an explicit method also gives very accurate damping factors, this method could give accurate approximations and would be simple to use. Additional numerical work must be done in order to determine any advantage for these methods that make the roots more accurate.

Table IV-13 shows that problems II and III are the two special cases of problem I where the continuous ψ and the difference α's, as shown above, agree identically for the required number of α's, as shown previously. Note that for graphical method F applied to problem II the characteristic equation is

$$[\tan(S-\tfrac{1}{2})\alpha_j][\tan\alpha_j] = 2 \qquad j = 1, \ldots, S \qquad \text{(IV-241)}$$

which would give α_j not equal to ψ_j for the infinite H problem; but these α_j are less than the corresponding ψ_j's. A numerical comparison of these roots is in Table IV-22. Despite the fact that they are not equal to the continuous roots, graphical method F does give the best results of the graphical methods for problem II. It is because of this fact that any firm conclusions about improvement in accuracy

by making α_1 equal to ψ_1 as described above cannot be made. Graphical method F is compared in more detail later.

The characteristic equations for method G and the continuous solution for problem IV are identical:

$$\sin S\psi_j = 0 \qquad \text{(IV-242)}$$

$$\psi_j = j\pi/S \qquad j = 1, 2, \ldots, \infty \qquad \text{(IV-243)}$$

$$\alpha_j = \psi_j \qquad j = 1, 2, \ldots, S-1 \qquad \text{(IV-244)}$$

For problem V the characteristic equations are the same as for problem II, if S is allowed to become infinite. Thus, any value of α or ψ satisfies the equation and α and ψ become continuous variables, and the summation is replaced with an integration.

Before leaving the discussion of the roots, the size of the smaller damping factors can be estimated because of the periodic location of the roots. The estimating relationship is:

$$\upsilon_{j+1} \simeq \mu_{j+1} \simeq \mu_j + \pi \simeq \upsilon_j + \pi \qquad \text{(IV-245)}$$

The first root μ_1 or υ_1 is necessary to start the calculation and if either is not known, μ_1 can be determined by finding q_{MAX} from a trial calculation using a coarse grid, or it can be taken as an angle between 0 and $\pi/2$ radians for problems with boundary equations like those for problems I and II, or as π for problems with boundary equations like those for problems III and IV. The relative size of

several of the terms for the smaller damping factor can be estimated as $e^{-\nu_j^2 \tau_0}$ and the number of terms J that are significant at time τ_0 can be estimated when the (J+1) term is a smaller fraction of the error bound, V. The general application of equation IV-245 should be further studied for problems where both H's are between 0 and infinity; however, surprisingly, that equation appears to apply even to the characteristic roots for the continuous problem with a finite H in cylindrical coordinates.

Conclusions and Summary--Trigonometric Roots. The S or (S+1) trigonometric roots for the approximate methods G, A, and C, as applied to problems II and III and for method G applied to problem IV are equal to the corresponding roots of the continuous solution. For methods G, A, and C applied to problem I and for graphical method F applied to problem II these roots are not equal.

The important points for problem I where H is not zero or infinite are:

(1) The difference between the squares of the roots $(\nu_j^2 - \mu_j^2)$ goes to zero with $1/S^2$.

(2) The roots for method C are more accurate than method G which in turn is more accurate than graphical method A.

(3) The probable reason that method C has the most accurate roots is that it is based on mesh $\Delta\xi/2$ with no points on the boundary. This then leads to an approximate formulation that is closer in both

mathematical and physical senses to the continuous problem (see section B-4, this chapter).

(4) The deviations for the first α_j's or μ_j's that determine the larger positive damping factors are relatively small. Deviations for the α's beyond $\pi/2$ are large, but these roots correspond to damping factors that can be negative; consequently, the differencing parameters are usually selected so that these damping factors are close to zero and therefore their effect is insignificant.

(5) The roots for graphical method A do not even approximately follow those for the continuous solution for α's greater than $\pi/2$, but are symmetric about $\pi/2$.

(6) Method G for problem I, when H is neither zero nor infinity, has a complex trigonometric root, α_{S+1}, which has no counterpart in the continuous solution.

A method that allows the estimation of the number of significant terms in the transient solution is also shown. This method is based on the approximate periodic distribution of the roots.

4. Damping Factors

The damping factors q_j are the quantities in the approximate transient solution that are raised to the nth power, and determine how the transient solution changes with time. In the solution to the partial differential equation the factors to the nth power are actually representations of the exponential $e^{-\nu_j^2 \tau}$ which is represented in difference

form by either $(e^{-\nu_j^2 \Delta\tau})^n$ or $(e^{-\psi_j^2 r})^n$. The discussion on the effect of damping factors is in two sections. First, the stability and oscillatory criteria as derived by the z-transforms are shown and related to criteria developed from matrix theory arguments in Chapter III. A subsection on graphical solutions discusses quantitatively the effect of oscillations on accuracy. The second part presents a series expansion which allows an estimation of $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ as a function of the differencing parameters if estimates of ν_j or μ_j are available. This ratio appears in the inequality for accuracy, equation IV-220.

The approximate damping factors are, for all methods applied to all problems,

$$q_j = \frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r(1-\cos\alpha_j)} = \frac{1-2S^2\Delta\tau(1-\gamma)(1-\cos\frac{\mu_j}{S})}{1+2S^2\Delta\tau\gamma(1-\cos\frac{\mu_j}{S})} \qquad \text{(IV-246)}$$

The expression in terms of α_j and r is used for the stability and oscillatory study; the expression with μ_j, $\Delta\tau$, and S is used in the analysis of accuracy. The eigenvalues of the Y/A matrix are given by

$$-\lambda_j = 2S^2(1-\cos\alpha_j) = 2S^2(1-\cos\frac{\mu_j}{S}) \qquad \text{(IV-247)}$$

In the analog solution, which is the special case for a $\Delta\tau$ of 0, $e^{-\lambda_j\tau}$ replaces q_j^n. The damping factors depend upon all the differencing parameters, the method used, number of points, time increment, and weighting γ. However, because their form is the same for all the

methods and problems, the conclusions made can be generalized to other problems and methods by using the different characteristic roots μ_j for the new case. Consequently, many of the conclusions made here are quite general.

Stability and Non-Oscillatory Behavior. The classical methods for studying stability of approximations to the diffusion equation derive equation IV-246 by a separation of variables technique for specific boundary equations. Then if the α_j that corresponds to q_{min} is known, the necessary and sufficient conditions for stable or non-oscillatory behavior are obtained. For the cases usually considered, the α_j's are all real, and

$$(1 - \cos\alpha_j) \leq 2 \qquad j = 1, \ldots, S \qquad \text{(IV-248)}$$

Using this inequality, the same sufficient conditions are derived that are obtained from the matrix criteria discussed in Chapter III, when the minimum Y/A matrix norm is $4S^2$. This condition is that for

$$\frac{1 - 4r(1-\gamma)}{1 + 2r\gamma} = -Y < q_{min} \qquad \text{(IV-249)}$$

where

$$Y > 0 \qquad \text{(IV-250)}$$

r and γ must satisfy the following equivalent inequalities:

$$r[1 - \gamma(Y + 1)] \leq \frac{Y + 1}{4} \qquad \text{(IV-251)}$$

or

$$\gamma \geq \frac{4r - (1+Y)}{4r\,(Y+1)} \qquad \text{(IV-252)}$$

These sufficient criteria apply to all approximations where either the damping factor is given by equation IV-246 and the α_j's are all real, or a norm of the Y/A matrix is equal to or less than $4S^2$. Consequently, they apply to all approximations to all problems discussed here except the approximations for either method G, when applied to problem I where for either case a complex α_{S+1} can occur, or for generalized method A when r is less than $\frac{1}{4}$. From the location of the α_j's for the cases when all of them are real, it is seen that as S is increased the α-root that determines q_{min} approaches π, and the inequality approaches equality. Therefore, q_{min} is close to its bound of -Y for large S. This is the basis for a prior statement in Chapter III that $|\lambda_{min}|$ for a Y/A matrix with equal column and row sums approaches the minimum matrix norm.

In Figure IV-14, the sufficient conditions based on the inequality IV-252 are shown. This is a graph of γ and r found from IV-252 using the equality sign for each of several Y's between 0 and 1. Thus, for example, selecting a combination of r and γ that falls on or above and to the left of a line for an Y of 0.2 assures one that the minimum q_j is greater than -0.2. For cases where the α_j's are real or where the minimum norm of the Y/A matrix is $4S^2$ or less, this graph can be used to find Y for a selection of γ and r to see if the inequality, equation IV-222, holds.

The curves are a family of hyperbolas that have a common asymptote of the γ axis (r=0) with the second asymptote a constant value of γ. For an Y of zero, the sufficient condition for non-oscillatory solutions, this asymptote is one,(γ=1); for an Y of one, the sufficient condition for stability, the asymptote is $\frac{1}{2}$,($\gamma = \frac{1}{2}$). From this chart, the following conclusions can be made about the effect of S, $\Delta\tau$, and γ on q_{min}, if it is assumed that Y is very close to q_{min}. First, increasing the weighting γ at a constant difference modulus r reduces Y rapidly for values of Y close to one: as Y approaches zero, a reduction in γ does not reduce Y as much as before. But reducing r at constant weighting factor lowers Y less when Y is close to 1 than as it approaches zero. This point becomes important as increasing r by increasing the time increment decreases $\tau_0/\Delta\tau$, and the allowable bound on $|q_{min}|$, Y, must also be decreased to keep $Y^{\tau_0/\Delta\tau}$ constant. Also, if the number of points S is large, either a very small time increment $\Delta\tau$ must be used or the weighting factor γ might have to approach or equal one to obtain an approximate solution that satisfies the inequality IV-222 for satisfactory oscillatory behavior. Although the curves for the sufficient conditions are not extended for negative weighting factors, negative weightings can be used as long as stable solutions are obtained.

However, the sufficient conditions, equation IV-252, and the graph, Figure IV-14, do not apply when a complex α_j root occurs.

In this case, unless the actual value of the complex α is determined, a study of its equation does not give a simple useful criterion. For example, the q_{min} corresponding to the complex root for method G applied to problem I is

$$q_{min} = q_{S+1} = \frac{1-2r(1-\gamma)[1+\cosh \mathcal{J}(\alpha_{S+1})]}{1+2r\gamma[1+\cosh \mathcal{J}(\alpha_{S+1})]} \qquad \text{(IV-253)}$$

where $\mathcal{J}(\alpha_{S+1})$ satisfies equation IV-229. However, a good simple upper bound for $[\cosh \mathcal{J}(\alpha_{S+1})]$ cannot be easily found from the characteristic equation, and the only criterion found from equation IV-253 without actually solving the transcendental equation IV-229 is that the solutions are stable for γ's of $\frac{1}{2}$ or larger and are non-oscillatory for a γ of 1. These are the same results which are obtained by assuming the maximum eigenvalue $|\lambda_{min}|$ or minimum norm of the Y/A matrix is infinity. However, the norms of the Y/A matrix are not infinity and do give good, simple, and sufficient criteria which are shown in Table IV-1.

Thus, the important conclusion can be made that the time-consuming z-transform solution does not always provide us with bounds for $|q_{min}|$ which are convenient or useful. Although the necessary and sufficient conditions for stability can always be found from a z-transform solution for a specific problem by solving the characteristic equation either analytically or numerically, this must be done for each value of S and for each value of H for an approximate method. However, for the

methods and problems where the roots are real, the z-transform solution gives a simple sufficient condition that applies for all S, but this sufficient condition is identical to that found much more easily using matrix norms. Thus, unless the roots of the characteristic equation are to be found for the specific value of S to be used, even for regular meshes, the sufficient conditions for stable or non-oscillatory solutions found by the simple additions to calculate the Y/A matrix norms give a better or equivalent bound for $|q_{min}|$ much more simply than the extended z-transform solution. Since the z-transform solution gives results for the characteristic equation and damping factors that are identical to those found by von Neumann's method or the separation of variables technique, the above conclusion shows the superiority of matrix theory for stability over any other known technique.

If the sufficient conditions for method G applied to problem I, as shown in Table IV-1, were to be superimposed on Figure IV-14, a separate condition would be required for each value of H. These curves would fall above and to the left of the curve shown as the stability restriction is more severe. However, for zero or infinite H they would coincide with those drawn. Thus, as graphical method G is unstable for intermediate values of H, it is defined only for zero or infinite H.

Graphical Solution. A graphical solution is defined as an explicit calculation ($\gamma=0$) with an r of $\frac{1}{2}$. With this selection of γ and r the damping factors become

$$q_j = \cos\alpha_j = \cos\frac{\mu_j}{S} \qquad j = 1, 2, \ldots, S \qquad \text{(IV-254)}$$

For all the methods as applied to the problems here, the roots α_j are located between 0 and π, and the largest α_j closest to π gives a damping factor q_{min} that has an absolute value of the same order of magnitude as q_{MAX}. For approximate methods applied to problems where the α_j roots are symmetrically located about $\pi/2$ for each positive q_j a negative q_j with the identical absolute value occurs. This is true because in these cases if α_j is a root $(\pi-\alpha_j)$ is also a root and

$$q_j = \cos\alpha_j = -\cos(\pi-\alpha_j) \qquad \text{(IV-255)}$$

These damping factors that are equal in absolute value, but opposite in sign, occur in approximate solutions of methods G, graphical A, and C applied to problems II, III, and IV, and graphical A applied to problem I. For methods G and C applied to problem I and graphical method F applied to problem II, the location is only approximately symmetric and a negative damping factor slightly smaller in absolute value than a corresponding positive damping factor occurs. However, for each of these cases, each negative damping factor can be paired to the corresponding damping factor with the same or larger absolute value. The requirement for graphical methods which replaces the oscillatory requirement of equation IV-222 is that the negative damping factor must be sufficiently smaller in absolute value than its paired positive

damping factor, or the weighting given the negative damping factor must be sufficiently smaller than the weighting for its paired positive damping factor.

Graphical method F as applied to problem II satisfies the first requirement as shown by its damping factors, summarized for an S of 5 in Table IV-22. This is one of the reasons graphical method F gives accurate approximations. Also, when the negative damping factor is sufficiently small, a somewhat larger value for its weighting can be tolerated.

For the methods where the α_j roots are symmetric, the roots that have equal magnitude, but opposite signs, each decay at the same rate; the positive one with no oscillations and the negative one with oscillations; hence an oscillatory effect is always significant. However, the rate of decay is shown to be a good estimate of the exponential rate of decay of the continuous solution, and by using the temperatures only at alternate intervals, the rate of decay is observed. This can be proved by computing the sum of the two terms in the transient solution that correspond to the pair of equal and opposite damping factors. If $[g_j\, c_{mj}(\cos\alpha_j)^n]$ is the term for the positive factor and $[g_{S+1-j}\, c_{m\,S+1-j}\,(\cos\alpha_{S+1-j})^n]$ is the term for the negative factor where α_{S+1-j} is equal to $(\pi-\alpha_j)$, their sum can be written

$$g_j\, c_{mj}(\cos\alpha_j)^n + g_{S+1-j}\, c_{m\,S+1-j}(-\cos\alpha_j)^n =$$

(IV-256)

$$\left[g_j + (-1)^n\, g_{S+1-j}\left(\frac{c_{m\,S+1-j}}{c_{mj}}\right)\right] c_{mj}(\cos\alpha_j)^n$$

$$0 \le m \le S$$

$$n \ge 0$$

The quantity $[g_j + (-1)^n\, g_{S+1-j}\,(c_{m\,S+1-j}/c_{mj})]$ is called the effective initial vector g_{jE} and the second requirement mentioned above is that the coefficient of $(-1)^n$ be relatively small. Consequently, each of the positive damping factors can be considered to be associated with two intercepts, for even values of n, $(g_j\, c_{mj} + g_{S+1-j}\, c_{m\,S+1-j})$, and for odd values of n, $(g_j\, c_{mj} - g_{S+1-j}\, c_{m\,S+1-j})$. This is shown in Figure IV-15, which is a semi-logarithmic plot of the single term $(g_1\, c_{m1} q_1^{\,n})$ and the summation as given from equation IV-256 versus time, for graphical method A as applied to problem II, using five points (S=5). The summation of the two terms is represented by the two parallel dot-dash lines, one for n odd, the other for n even, respectively, above and below the dashed line for $(g_1\, c_{m1}\, q_1^{\,n})$. The accuracy of the graphical solution then is determined primarily by the accuracy of the effective initial vector components, and this is discussed in the section on initial vectors. However, by comparing the effective term for n-odd for graphical method A for the point at m of 1 with the first term in the

continuous solution, one can conclude that accurate approximations are expected from graphical method A.

One further point should be mentioned about unaveraged graphical methods G and A applied to problem III, where both H's are zero. In this case an α_S of π occurs, giving a q_S of -1.* This is the α_j that is symmetric to the α of zero that gives a q of 1 for the steady-state solution.** Graphical solutions for problem III do not show the equal oscillations of this root because, for the specific linear initial distribution of the problem and for S even, the component of the initial vector for this root is zero. However, the graphical solution for S odd does show equal oscillations, and graphical solutions for other initial distributions should, in general, give equal oscillations for these methods. If the graphical solution is averaged, the oscillations of the -1 root are eliminated as this involves equivalently multiplying the weighting of q_{min} of -1 by $(q_{min} + 1)$. For the adiabatic problem the -1 damping factor cannot be excited by the boundary forcing functions, i.e., a term of the form $n(-1)^n$ cannot be in its solution. But since the Y/A matrix and its derived matrices for a problem with a specific flux at both boundaries are the same as would be used for both H's zero, an

---

* The angle π also satisfies the characteristic equation for method C applied to problem III, but it would be associated with an eigenvector which has zero for all its components; consequently, it can never occur in a solution for method C applied to problem III.

** The calculation matrix, when both H's are zero, must have an eigenvalue or q_j of 1 to give a non-zero steady-state solution.

oscillatory flux with a frequency of $1/2\Delta\tau$ would excite the q of -1, and the solution would have oscillations with a linearly increasing amplitude going to infinity with time. (The change in the matrix equation would be that the specified boundary flux would replace a component of the Y_B t_B vector directly.) Consequently, an unaveraged graphical method G or A for boundary conditions of specified flux is technically an unstable approximation, according to the definition; however, the conditions under which this can occur are rather unusual. Further, if the solution is averaged, the oscillations with a linearly increasing amplitude for the unaveraged solution become equal oscillations in the averaged solution, which is then the same form of solution as the continuous solution.

Accuracy of Damping Factors. The above discussion is concerned with selecting the differencing parameters so that the inequality in equation IV-222 is satisfied or, equivalently, so that a negative q_{min} does not ruin the accuracy of the approximate solution. The accuracy of an approximate solution is determined in part by how close an agreement there is between the slope, on a semi-logarithmic plot, of the terms in the approximate transient solutions containing the larger damping factors, and the slope of the corresponding term in the continuous solution. As pointed out previously, these slopes are, respectively, $-\upsilon_j^2$, and $[(1/\Delta\tau) \ln q_j]$. The difference between these quantities is the natural logarithm of the ratio $e^{-\upsilon_j^2}/q_j^{1/\Delta\tau}$ or

$e^{-\psi_j^2 S^2}/q_j^{1/\Delta\tau}$ which appears in the one of the inequalities IV-220 that must be satisfied to achieve a certain accuracy.

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = -\nu_j^2 - \frac{1}{\Delta\tau} \ln q_j \qquad \text{(IV-257)}$$

In taking the logarithm of q_j, q_j is assumed to be positive; therefore, the following series, which are derived using the logarithm of q_j, apply only to the positive damping factors. The oscillatory effect of any negative q_j's has already been discussed.

A series expansion for the logarithm of the approximate damping factor can be developed from a series for the damping factor. The series for q_j is found by substituting the series for $[\cos(\mu_j/S)]$ into equation IV-246 for q_j and carrying out the indicated long division. This series is shown on Table IV-14, together with the exponential series for the continuous damping factor $e^{-\nu_j^2 \Delta\tau}$. A comparison of the damping factors in the form shown on this table is not entirely satisfactory for two reasons. First, neither infinite series converges rapidly except for very small time increments. Second, both damping factors are strong functions of the time increment, and decreasing the time increment changes both factors rapidly, making them appear to approach one at zero $\Delta\tau$. Consequently, it is difficult to determine quantitatively the effect on accuracy as the time increment is reduced. The series for the natural logarithm of q_j can then be found by

substituting the series for q_j into the series for the natural logarithm given in Dwight (27), equation 601.5. The resulting expression is at the top of Table IV-15.

Before starting the detailed comparison of this series, one further manipulation should be mentioned, namely that a series for q_j^n can be developed from the series for the logarithm of q_j. The logarithm series is multiplied by n and then substituted into the exponential series and an expansion for q_j^n results. This series is shown in Table IV-14, together with a series for $e^{-\nu_j^2 n\Delta\tau}$. Again, a comparison of these two series suffers from the same disadvantages as a comparison of the damping factors; however, they are useful in showing that the analog solution is the limiting case of an approximate solution as the time increment goes to zero and n goes to infinity so that the product $n\Delta\tau$ is always equal to time.

$$\lim_{\substack{\Delta\tau \to 0 \\ n \to \infty \\ n\Delta\tau \to \tau}} q_j^n \longrightarrow e^{-2S^2\left(1-\cos\frac{\mu_j}{S}\right)\tau} \qquad \text{(IV-258)}$$

Moreover, taking the limit of the q_j^n as the time increment goes to zero and the number of points goes to infinity, together with a similar limiting process on the characteristic equation, shows that the continuous exponential results.

$$\lim_{\substack{\Delta\tau \to 0 \\ n \to \infty \\ S \to \infty}} \quad q_j^{\,n} \quad \underset{n\Delta\tau \to \tau}{\longrightarrow} \quad e^{-\nu_j^2 \tau} \tag{IV-259}$$

The first several terms in the series expansion for the logarithm of the ratio $e^{-\nu_j^2}/q_j^{\,1/\Delta\tau}$, or the difference in the slopes, are found by subtracting the infinite series for $[(1/\Delta\tau)\ln q_j]$ from $-\nu_j^2$. The result is:

$$\ln \frac{e^{-\nu_j^2}}{q_j^{\,1/\Delta\tau}} = (\mu_j - \nu_j)(\mu_j + \nu_j) + \mu_j^4 \left[(\tfrac{1}{2} - \gamma)\Delta\tau - \frac{1}{12S^2} \right] + \tag{IV-260}$$

$$+ \mu_j^6 \left[(\tfrac{1}{3} - \gamma + \gamma^2)(\Delta\tau)^2 - \frac{(\frac{1}{2} - \gamma)\Delta\tau}{12S^2} + \frac{1}{360S^4} \right] \cdots$$

The advantage of using this expression for comparing damping factors is, first, a series for only one of the damping factors is involved. Second, since $e^{-\nu_j^2}/q_j^{\,1/\Delta\tau}$ should be one and its logarithm zero, each non-zero term in the series is an error term and is a direct result of the space and/or time discretization. Third, the term $(e^{-\nu_j^2}/q_j^{\,1/\Delta\tau})^{\tau}0$ can be readily found and used directly in equation IV-220. The series is shown in Table IV-15 for the common special cases of $\gamma = 0$, $(\frac{1}{2} - 1/12S^2\Delta\tau)$, $\frac{1}{2}$, and 1, for graphical solutions where $\gamma = 0$ and $S^2\Delta\tau$ or r is $\frac{1}{2}$, and for the analog case of zero $\Delta\tau$.

Equation IV-260 for $\ln(e^{-\nu_j^2}/q_j^{\,1/\Delta\tau})$ is valid for all positive

q_j where the series converges. However, in studying the size of this logarithm the series must converge rapidly enough so that the three terms shown give a good estimate of the logarithms. Because of this, and because of the difficulty in obtaining a general term for this series, the convergence was not studied directly, but a numerical study was made to see if and when the three terms give a good estimate of the logarithm. This study was based on q_{MAX} for problem II where μ_1 and ν_1 are equal to $\pi/2$ or 1.57, and ν_1^2 is 2.46740. The range of the logarithm of the ratio tested was from 0.00000 to 0.1327.

The main conclusion from this study is that the terms shown give an accurate estimate of the logarithm for as large a time increment as one would expect to encounter, and it can be used to study quantitatively the effect of differencing parameters on the accuracy of the slope, or to estimate the ratio $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ for use in equation IV-221. Specifically, a good estimate was found for the logarithm for γ's of zero and one as long as the μ^6 term was less than 5 per cent of the μ^4 term and for γ's of $\frac{1}{2}$ or $(\frac{1}{2} - 1/12S^2\Delta\tau)$ up to where the μ^6 term was twice that of the μ^4 term. This means that accurate estimates of the logarithm can be made for time increments up to 0.10 for γ's of zero or one, and up to time increments of 0.2 for γ's of $\frac{1}{2}$. This compares to either series for damping factors which, using the terms shown, do not give accurate estimates of the damping factor for time increments above about 0.04 to 0.08.

The series for $\ln(e^{-\upsilon_j^2}/q_j^{1/\Delta\tau})$ is arranged in increasing powers of μ_j^2. By selecting the weighting factor γ and time increment $\Delta\tau$ correctly, it is possible to make the coefficients of some of the terms very small or even zero. However, because the exact values of the μ_j are different not only for each problem but within the same problem, it is not practical to try to select $\Delta\tau$ and γ so that compensation between terms is obtained, although, in some methods, this type of compensation does occur. Therefore, the following discussion treats the three terms separately.

The first term in the series is $(\mu_j-\upsilon_j)(\mu_j+\upsilon_j)$, which was thoroughly discussed in the section on the characteristic roots. It is zero for all approximations to problems II, III, and IV which have been considered except graphical method F applied to problem II. For problem I its magnitude depends upon the method used, and it goes to zero with $1/S^2$. Because of this dependence on $1/S^2$, this term in general should be of about the same order of magnitude as the μ^4 term. In practice, it is usually smaller than the μ^4 term, except when special conditions are used to make the μ^4 term zero. Its contribution to the logarithm is greatest for graphical method A and smallest for method C. However, one of the reasons graphical method A does give accurate results is that the size and the sign of this term cancel out up to one-third of the $\mu^4/6S^2$ term for q_{MAX} depending upon the value of H. For method F applied to problem II the difference between μ and υ for an

S of 5 is almost negligible, although some compensation occurs there also.

Although in this analysis this $(\mu_j^2-\nu_j^2)$ quantity is treated as a special term, by carrying out the indicated multiplication of series in equation IV-232 a series expansion can be developed for this difference. If this expansion is substituted into equation IV-260, a quantity would be added to each of the other terms in the series. The first term in the series for $(\mu_j^2-\nu_j^2)$ would be proportional to μ_j^4/S^2 and the proportionality constant would be a function of the dimensionless heat-transfer coefficient, H. When added to the μ^4 term in equation IV-260, it would change the coefficient of $1/S^2$.

The second term of the expansion of the logarithm of the ratio $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ is the dominant term in the series for all problems, and is the first non-zero term for problems II, III, and IV. This term is $\mu_j^4\left[(\frac{1}{2}-\gamma)\Delta\tau - \frac{1}{12S^2}\right]$ and for most useful combinations of the differencing parameters, it exerts the most influence of all the terms in the series on the logarithm of the ratio, and consequently, from its value alone, one can tell much about the size and direction of the error in the slope of the approximate transient terms on a semi-logarithmic plot. For example, this single term agrees with the ratio $\ln(e^{-\nu^2}/q^{1/\Delta\tau})$, to within one per cent, for a graphical solution applied to problem II, using an S of 5. And, in general, for most combinations of differencing parameters that would be used, conclusions about the damping factors

based only on this term are usually correct, so the effect of the differencing parameters on this term are studied in some detail.

In Figure IV-16 the coefficient of μ^4 which is $[(\frac{1}{2}-\gamma)\Delta\tau - 1/12S^2]$ is related to the increment for each of several combinations of γ of 0, $\frac{1}{2}$, and 1, and S of 5, 10, and infinity. The lines can be classified by either of two groupings; one for constant γ for each S, and one for constant S for each of the γ's. For a constant weighting factor γ and the different S's the lines are parallel with slope of $(\frac{1}{2} - \gamma)$ with different intercepts at $-1/2S^2$. For a constant S the intercept at a time increment of zero, the analog solution, is $-1/12S^2$ and the lines leave that point with slope $(\frac{1}{2} - \gamma)$.

Examination of the graph indicates that the time increment and weighting γ affect this term much more than does the number of points S, providing S is selected larger than about 5. Changing S from 5 to ∞ changes this coefficient only 0.0033, which is equivalent to reducing the time increment 0.00167 for a γ of one. Further, by doubling the number of points, the contribution to this term caused by space discretization is reduced by $\frac{3}{4}$. Only for time increments below 0.01 does the space discretization error term $1/12S^2$ for an S of 5 have the same size as that caused by the time discretization $(\Delta\tau)(\frac{1}{2}-\gamma)$. Also, raising the weighting γ from 0 to 1 at constant time increment, and where r is greater than 1/6, changes the coefficient from a positive to a negative quantity with larger absolute value,

resulting in a larger change in the term than increasing S from 5 to infinity. Consequently, increasing S does not improve the accuracy of the damping factors as rapidly as changing the time differencing parameters, $\Delta\tau$ and γ, does, and the contribution of $1/S^2$ or $(\Delta\xi)^2$ to this error term is small.

The most important conclusions about this coefficient are based upon the fact that the intercept is negative $(-1/12S^2)$ but that its slope is either positive or negative and is given by $(\frac{1}{2}-\gamma)$. Thus, the backward difference method with a weighting of one always has a slope of $-\frac{1}{2}$ and this term is then always negative. Therefore, when a weighting of 1 is used, the damping factors for the approximate solution are too large and the approximate transient solution does not decay as rapidly as does the continuous transient solution. For the central difference calculation, when the weighting γ is one-half, the coefficient is equal to the intercept and is independent of the time increment. Thus, the approximate damping factors are accurate and the approximate solution is close to the continuous solution and also to the exact analog solution. For the explicit forward difference scheme with a γ of zero, the slope of the coefficient is $+\frac{1}{2}$ and the coefficient is zero when r is 1/6. The coefficient for the graphical solutions is the last point on the solid line. The logarithm of the ratio is positive here and the terms in the transient solution for a graphical construction decay more rapidly than those in the continuous solution. Although the damping factors for

the graphical solutions are not as accurate as those for the central difference calculation, $\gamma = \frac{1}{2}$, they should be sufficiently accurate for many calculations. Two other interesting points about the forward difference calculation are, first, that even when the solution is unstable as shown by the dashed lines, the larger damping factors are slightly more accurate than those for the backward difference calculation and, second, that increasing the number of points at some larger constant time increments actually reduces the accuracy of the damping factors.

The negative intercept and a positive slope can compensate to reduce the error in the damping factor; this requires that the weighting γ is selected with the range of 0 to $\frac{1}{2}$.

$$1 \leq \gamma < \tfrac{1}{2} \tag{IV-261}$$

Further if the weighting γ is selected according to

$$\gamma_o = \tfrac{1}{2} - \frac{1}{12S^2\Delta\tau} = \frac{6r-1}{12r} \tag{IV-262}$$

the μ_j^4 term is zero. This equation was first derived by putting the third term of the damping factor series in the same form as the third term of the exponential series (15); it also makes the third term of the q^n series of the same form as the third term of the series for $e^{-\nu^2 n\Delta\tau}$. Richtmyer (3) has found the relationship directly from the discretization error in deriving the partial difference equation by Taylor series

expansions. For problems II, III, and IV, where ν and μ are equal, very accurate solutions can be obtained by selecting γ and $\Delta\tau$ for a given number of points based on equation IV-262 as this gives a very accurate largest damping factor q_{MAX}, and also significantly improves the accuracy of the other larger damping factors. However, for problem I, the weighting factor γ should be selected slightly larger than γ_o given by equation IV-262 for method G and slightly smaller than that γ_o given by equation IV-262 for method C to compensate for the difference $(\mu^2-\nu^2)$. The amount larger or smaller would depend upon H, how the quantity $(\mu_j^2-\nu_j^2)$ changes with S for a specific method, or upon the proportionality constant of the term μ_j^4/S^2 in an expansion for $(\mu_j^2-\nu_j^2)$. Consequently, the selection of γ_o by using equation IV-262 is not a general relationship that gives the best approximation for all problems, as it only applies directly for methods applied to problems where the approximate trigonometric roots are equal to the first several roots of the continuous solution. A numerical study would probably be required to determine the modified relationship for an optimum weighting γ_o for problem I.

When the weighting is selected by equation IV-262 the oscillatory components must be considered. In Figure IV-14 the relationship between γ and r is superimposed upon the sufficient conditions for bounding the oscillatory damping factors. The relationship is a hyperbola with one asymptote at the γ axis (r=0) and the other at γ of $\frac{1}{2}$. As it always

is above or to the left of the sufficient condition for a q_{min} of -1, the approximation given by equation IV-262 is always stable. But, as can be seen from the graph, a large negative q_{min} is expected for large r's. In these cases, unless the q_{min} becomes negligible compared to the larger q's at the time of interest, the oscillatory effect masks the accuracy of the largest damping factor. Also, for large r's, it is possible that $\Delta\tau$ may be so large that the logarithmic series does not converge rapidly; and even selecting γ by equation IV-262 does not give an accurate q_{MAX}. Therefore, for problems where a large time increment is to be used, or where r is large, the weighting γ_o cannot be used.

The expression in equation IV-262 can also be used to estimate if the approximate solution decays too rapidly, or too slowly. If the weighting γ used for an approximation is greater than that γ_o for the time increment and number of points used, then the logarithm of $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ is less than zero, the ratio $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ is less than one, and the approximate solution decays too slowly. Conversely, if the weighting γ used is smaller than γ_o, then the logarithm of the ratio is greater than zero, and the approximate solution decays more rapidly than the continuous transient solution. The above comments are for weightings that are different from γ_o by a significant amount and they then hold even for approximations to problem I. Further, the above statements hold for all damping factors in the approximate

solution compared to the continuous solution; only when the weighting γ is very close to γ_o are some damping factors greater and some less than their corresponding continuous damping factors.

The μ^6 term in the series for the logarithm is only significant for larger time increments or for the ratios for larger μ_j roots. The coefficient of μ_j^6 is plotted versus time for the several combinations of S and γ studied previously in Figure IV-17. These plots are a series of parabolas, and the size of the coefficient for the range of differencing parameters shown is always less than about 1/25 of that for μ^4. This coefficient is small for γ_o as given by equation IV-262 and for a γ of $\frac{1}{2}$. Also, from the previous numerical study of the series for the logarithm, it could be concluded that for either of these γ's the μ^8 and higher order terms are very small. Moreover, although not apparent from the graph when γ_o is used, this coefficient can be made zero when $r = S^2\Delta\tau = 0.223607$. The coefficient when the weighting γ is zero or one is somewhat larger and positive. This tends to compensate the μ^4 term when γ is one, but this term never is large enough to dominate the negative effect of the μ^4 term. For the explicit case ($\gamma = 0$) this term does not compensate the μ^4 term but increases the logarithm more when r is greater than 1/6.

Conclusions and Summary--Damping Factors. A series expansion was developed for the difference between the slopes on a semi-logarithmic graph for corresponding terms in the transient solution for

the continuous and approximate methods. This is equation IV-260 which relates the $\ln (e^{-\upsilon_j^2} / q_j^{1/\Delta\tau})$ to the jth trigonometric roots and the differencing parameters. This expansion was shown to be suitable for estimating $e^{-\upsilon_j^2} / q_j^{1/\Delta\tau}$ for use in the inequality concerning damping factors, equation IV-220, which must be satisfied to have an approximate solution of a certain accuracy. From this expression the following observations may be made:

(1) The difference in slopes is approximately proportional to $1/S^2$ and to $\Delta\tau$; the proportionality constant for $\Delta\tau$ is a function of the weighting γ.

(2) For all methods and problems, a term proportional to μ_j^4 exerts the primary influence on the difference in slope. Three important points based on this coefficient are: (a) the coefficient of μ_j^4 is primarily a function of the time differencing parameters, and, if S is greater than about five, further increase in the number of points does not reduce this coefficient nearly as much as do appropriate changes in the time increment and/or γ; therefore, for problems where the trigonometric roots are equal, damping factor accuracy is not greatly improved by increasing the number of points. (b) The coefficient of μ_j^4 can be made zero by selecting the weighting γ as γ_o according to equation IV-262. This selection of differencing parameters gives the most accurate damping factors for methods and problems where the approximate trigonometric roots are equal to the

corresponding continuous trigonometric roots. (c) If γ is significantly smaller than γ_o for any method applied to any problem, then the transient solution decays too slowly; the reverse is true when γ is significantly larger than γ_o. Since $0 \leq \gamma_o < \frac{1}{2}$, the least accurate larger positive damping factors are found with $\gamma = 1$.

(3) For problems where the roots are not equal, the difference in slopes is equal to $(\mu_j^2 - \upsilon_j^2)$ when $\gamma = \gamma_o$. Some improvement in the accuracy of damping factors could be obtained by use of a slightly different value of γ to allow compensation with the μ_j^4 term. For instance, for a given S and $\Delta\tau$ the optimum γ for method C on problem I is slightly less than γ_o.

The oscillatory behavior was also studied based on the z-transform solutions. The conclusions are:

(1) Although the exact value of $|q_{min}|$ can be found in theory from the complete z-transform solution, only a good bound for $|q_{min}|$ is necessary in practice. Even for regular meshes the easily calculated matrix norms give a better or equivalent bound for $|q_{min}|$ than can be easily found from the z-transform solution. Thus, the matrix norms should always be used in studying stability and oscillatory criteria. Even when the trigonometric root for the characteristic equation can be found analytically, because it appears in a cosine function within the equation for $|q_{min}|$, it is inconvenient to use.

(2) The sufficient coonditions that $|q_{min}|$ be less than Y were derived as equation IV-252 and shown in Figure IV-14. This figure is for use with all methods applied to problems where the trigonometric roots α_j for the approximation are real or where the minimum norm of the Y/A matrix is less than $4S^2$. This graph is suitable for finding Y for use in equation IV-222 which assures satisfactory oscillatory behavior of method C for problems where the heat-transfer coefficients at the left (H_0) and the right (H_S) boundaries can take on any values:

$$0 \leq H_0 \ , \ H_S \leq \infty \qquad \text{(IV-263)}$$

It also applies for the following cases: (a) method G where H's are zero or infinity, or one H is zero and the other H infinity; (b) generalized method A to all problems when r is greater than or equal to $\frac{1}{4}$; when r is less than $\frac{1}{4}$ only for cases where the graph can be applied to method G.

(3) The requirement for a graphical solution to give satisfactory behavior, even though it contains negative damping factors of the same order of magnitude as the positive damping factors, is that either the negative damping factor must be somewhat smaller than a paired corresponding positive damping factor or, if the paired damping factors are equal in magnitude, the weighting for the negative damping factor must be low. For the latter situation, the effect of the oscillations is to change the intercept on the semi-logarithmic graph of the transient term for alternate time intervals, but the slope is the same, if the values for only even n or odd n are considered.

5. Eigenfunctions and Eigenvectors

Associated with each damping factor is an eigenfunction for the continuous solution or an eigenvector for an approximate solution. The jth eigenfunction $b_j(\xi)$ gives the relative weighting of the jth continuous damping factor $e^{-\psi_j^2 r}$ for all points ξ within the solid; the jth eigenvector c_j shows the relative weighting of q_j^n at each point in the difference network. These eigenvectors c_j are the eigenvectors of the Y/A matrix. Thus, they are determined by the number and location of the points and the boundary equations, and apply to all problems where the same Y/A matrix is used in the calculation. A "continuous eigenvector" is defined as an S-dimensional vector with components found by evaluating the eigenfunction at each point, $b_j(m\Delta\xi)$. The eigenvector matrices are then the matrices which have as columns the S-dimensional eigenvectors, and are the B matrix for the continuous solution and the C matrix for the approximate solution. Since the intercept on the semi-logarithmic graph for the corresponding terms in the transient solutions for the mth point is the product of the jth initial vector component times the mth component of the jth eigenvector, $g_j c_{mj}$ for the approximate solution and $a_j b_j(m\Delta\xi)$, a difference between corresponding elements of the continuous and approximate eigenvector matrices contributes to the error in this intercept which affects the total error as indicated in equation IV-229.

For all the approximations considered, and for the continuous solution to problems I through IV, the weightings of the jth damping

factor in the solution for the mth temperature point are of the same form and are:

$$\text{Continuous:} \quad \cos(S-m)\,\psi_j \qquad j = 1, 2, \ldots, \infty \qquad \text{(IV-264)}$$

$$\text{Approximate:} \quad \cos(S-m)\,\alpha_j \qquad j = 1, 2, \ldots, S \text{ or } S+1 \qquad \text{(IV-265)}$$

where the values of m depend upon the mesh. Consequently, for the approximate method applied to a problem that gives the same trigonometric roots α_j as the first S-roots for the continuous solution, the S-eigenvectors of the approximate method are equal to the first S-"continuous eigenvectors."

$$b_j(m\Delta\xi) - c_{mj} = 0 \qquad j = 1, \ldots, S \qquad \text{(IV-266)}$$

This equation IV-266 is valid for methods G, A, and C, applied to problems II and III, and to method G applied to problem IV. Further, it is valid for those methods applied to problems like those above where the boundary equations are the same, but the forcing functions and initial distribution are different. Thus, the error in the intercept for the transient term in these cases is caused only by the error in the initial vector component.

For methods G, A, and C used to approximate the continuous solution of problem I where H is neither infinite nor zero and for method F applied to problem II the α_j's are not equal to the corresponding ψ_j's. The difference eigenvectors thus are not identical to the continuous eigenvectors. The difference between corresponding elements of the continuous and approximate eigenvector matrices can be found

using the trigonometric identity for the difference between cosines. This gives, after a series expansion for a sine of a small angle, and replacing $(\mu_j+\upsilon_j)$ with $2\mu_j$:

$$c_{mj}-b_j(m\Delta\xi) \simeq$$

$$(\upsilon_j-\mu_j)(1-\frac{m}{S})\left[1-\frac{(\mu_j-\upsilon_j)^2(1-\frac{m}{S})^2}{6}\right]\sin(1-\frac{m}{S})\mu_j \qquad \text{(IV-267)}$$

(This equation does not apply to the complex α_{S+1} for method G and problem I.)

This is the contribution to the error in the intercept of the jth transient solution term in the equation for the mth point, caused by the approximate eigenvector. It appears in the inequality equation IV-219 that is used to determine the number of points required. Since $(\mu_j-\upsilon_j)$ goes to zero with $1/S^2$, the error in the eigenvector matrix elements also goes to zero with $1/S^2$, but the proportionality constant would be smaller for points close to the adiabatic boundary ($m \rightarrow S$) and larger at the heat-transfer surface. Also, the error in the eigenvectors for the smaller damping factors would be larger than those for the larger damping factors. The error in the elements of the eigenvector for the largest damping factor always has the same sign as does $(\mu_j-\upsilon_j)$; consequently, for method G these elements are smaller than the continuous quantities, and for method C and graphical A they are larger than for the continuous solution.

As the eigenvectors shown in equations IV-264 and IV-265 have not been normalized to unit length, they do not satisfy the orthogonality relationships, equations II-18 and II-49. However, if the eigenvectors,

as shown in the above equations, are multiplied by a proportionality constant which has a j subscript, they would satisfy the orthogonality relationship. The eigenvectors as shown in equations IV-264 and IV-265 are in a convenient form for comparison of accuracy. However, a simpler form for the eigenvectors for several of the methods and problems can be found by substituting the trigonometric roots into these equations and then cancelling a factor that is in the initial vector. The simplest form for the eigenvectors is Table IV-16 , and the initial vectors shown in the following tables, and to be discussed in the next section, are consistent with these eigenvectors. That is, in reconstructing the complete solutions the eigenvectors shown in Table IV-16 should be used with the initial vectors tabulated, not the eigenvectors given in equations IV-264 and IV-265.

These expressions are all sines and cosines of real angles except for α_{S+1} for method G applied to problem I. Consequently, the absolute value of the elements of the eigenvector matrix vary from zero to one, and they are usually about the same size. Therefore, the size of the $g_S\ c_{mS}$ product that is the amplitude for the damping factor that can be negative, q_{min}, is determined largely by the size of its initial vector component, and the eigenvector elements usually are not the reason for a large amplitude. However, part of the difference in the behavior of the graphical solutions for methods C and G in Figure IV-1 is explained by the eigenvector components. In this

case, with S=5, q_{min} is -0.9511 for both methods, and the expression for the eigenvector corresponding to q_{min} for both methods is $[\sin m(2S-1)\pi/2S]$. However, for method C m is $\frac{1}{2}$ and for method G it is 1, and this component for method C is about $2S/\pi$ times the component for G, partially accounting for the larger amplitude of oscillations.

The eigenvector for method G that corresponds to q_{min} and the complex root α_{S+1} is

$$\cos(S-m)\alpha_{S+1} = (-1)^{S-m} \cosh(S-m) \mathcal{J}(\alpha_{S+1}) \qquad \text{(IV-268)}$$

Since the hyperbolic cosine is large for large $[\mathcal{J}(\alpha_{S+1})]$ this can have large components; however, the amplitude of any oscillation is affected by the size of its initial vector component which is small for this case.

Conclusions and Summary--Eigenvectors. The eigenvectors as shown by equations IV-264 and IV-265 or in Table IV-16 only contribute to the error in the intercept of the semi-logarithmic graphs of a transient term when the approximate trigonometric roots are not equal to the corresponding continuous roots. This error thus does not occur for methods G, A, and C applied to problems with one heat-transfer coefficient zero and one infinity, problem II, and with both heat-transfer coefficients zero, problem II, and for method G applied to problem IV. The eigenvector matrix is in error for all methods applied to problem I

and for graphical method F applied to problem II. The error in each element is proportional to $(\mu_j - \nu_j)$ and consequently goes to zero with $1/S^2$.

The elements in the eigenvector matrix are of the same order of magnitude, and thus the eigenvectors are not usually responsible for a large amplitude, $g_S\, c_{mS}$, for the minimum and possibly negative damping factor. An exception is that for the point located at $\frac{1}{2}$ for method C applied to problem I, the large size of its component of the Sth eigenvector partially explains the oscillatory behavior of this method.

6. Initial Vectors

The transient solution of either the partial differential equation, or the partial difference equation, is a vector which is the sum of weighted eigenvector-damping factor products. That is,

$$t_n = t_{P\,n} - \left[g_1 c_1 q_1^{\,n} + g_2 c_2 q_2^{\,n} + \ldots + g_S c_S q_S^{\,n} \right] \qquad \text{(IV-269)}$$

where the bracketed quantity represents the vector $C\,Q^n g$ for the transient solution; the c_j are the eigenvectors; $q_j^{\,n}$ are the corresponding damping factors to the nth power; and the g_j are the components of the initial vector. Each g_j is the weighting given to the eigenvector-damping factor product, $c_j q_j^{\,n}$. The product of the initial vector component and the mth component of its eigenvector $g_j c_{mj}$ determines the intercept on the semi-logarithmic graph of the jth term in the transient solution for the mth point. For the large damping factors, this quantity

should be close to the continuous intercept $a_j b_j(m\Delta\xi)$, but for the higher-subscripted damping factors which can be negative, one desires this quantity to be small.

In addition to being a function of the initial temperature distribution, the initial vector is a function of the Y/A matrix as shown by the orthogonality relationship, equation II-49. Thus, for methods G, graphical A, and C, applied to a specific problem, the initial vectors are a function only of the number of points; for generalized method A or an averaged method they are a function also of the time differencing parameters. Because the initial vectors are specific functions for a given initial condition and do not follow a given form as do both the damping factors, conclusions which are as precisely stated and generally applicable cannot be made. Probably because of this and also because of the difficulty of obtaining a function for the initial vector from the orthogonality relationship, the initial vectors apparently have not been studied previously. However, the initial vectors can and often do have a decisive effect on the accuracy of an approximate solution both from the standpoint of accuracy of lower-subscripted components, and the absolute value of the higher-subscripted components.

For these reasons, the specific initial vectors for each approximate method with its variations to each problem are not discussed individually. (These initial vectors and related quantities are summarized in Tables IV-17 through IV-28.) Instead, first, the accuracy

of the components for the slowest-decaying terms and the size of the components for the possibly negative damping factors are discussed. Second, the effect of the amplitude of the oscillatory components for graphical methods is discussed using the effective initial vector components based on equation IV-26 . This is really a consideration of the sum of two paired terms in the approximate solution. Third, the effect of averaging on accuracy is discussed for both general approximations and graphical solutions.

Unmodified Initial Vector Components. The unmodified initial vector components are summarized in Tables IV-17, IV-18, IV-19, and IV-20, for the several methods and problems. These unmodified initial vector components are the components for the unaveraged methods as shown in equation IV-269 and apply directly to all S damping factors. That is, g_1 is the weighting factor for $c_1 q_1^n$, the slowest decaying term, and g_S or g_{S+1} is the weighting for $c_S q_S^n$ or $(c_{S+1}\, q_{S+1}^n)$ which is the term that can be negative and give oscillations. Thus, although the components are shown for graphical method A, these components do not include the effects of oscillations; they are included in the effective initial vector components discussed in the next subsection, and the accuracy of any graphical solution cannot be inferred directly from the table or the following discussion.

A study of these unmodified components for problems I, II,

and III shows that the g_j components for method C are closest to the corresponding a_j of the continuous solution; method G has the next most accurate components, and the difference between g_j and a_j is greatest for graphical method A. This means that if a combination of time increment and weighting factor is used so that all q_j's are positive, method C would be expected to give a more accurate approximate solution than method G. Since for graphical method A the time differencing parameters $\Delta\tau$ and γ are fixed, this cannot be done, and it is not included in the comparison. Further, from a comparison of the accuracy of the g_j one also cannot be sure which of the graphical solutions gives the best approximation, because of the effect of oscillations.

The above conclusion about the difference between g_j and a_j is based upon the series expansions which are developed directly from series expansions for the trigonometric functions in g_j for the components for problems II and III, and from comparison of each function in the expression for the components, and a brief numerical study for g_1 for problem I. The initial vector for problem II is a special limiting case of problem I as H goes to infinity; however, problem III is not the limiting case of problem I with an H of zero, as its initial distribution is not constant. A minor exception to the above conclusion is that method G can be more accurate than method C for problem I for values of H so that H/2S is about the size of 2S, and for values of H near where $(g_j - a_j)$ goes to zero for method G. For problems II and III

the first term in the expansion for $(g_j - a_j)$ is proportional to $1/S^2$ and the proportionality constant for method C is always $\frac{1}{2}$ that for method G. If method C always gives an intercept that has this relationship, then to obtain a given accuracy $(g_j - a_j)$ in intercepts, 41 per cent fewer points would be required than for method G.

The probable reason why method C gives more accurate initial vectors is that the Y/A matrix for method C is the only matrix approximation to the Laplacian which is symmetric, and consequently, its eigenvectors are orthogonal with respect to the same form of weighting factor as is the continuous solution. The integration of the orthogonal functions to find the Fourier coefficients a_j for Cartesian coordinates uses a dimensionless area weighting $\sigma(\xi)$ of 1 for the volume weighting:

$$\sigma(\xi)d\xi = d\xi = dV \tag{IV-270}$$

where dV is a dimensionless differential volume.

The integration is:

$$a_j = \int_0^1 b_j(\xi) \left[T(\xi, 0) - T_P(\xi, 0) \right] d\xi \tag{IV-271}$$

The summation of the eigenvectors to find the initial vector coefficients g_j is

$$g_j = \sum_{m=\frac{1}{2}}^{S-\frac{1}{2}} c_{mj} \left[t_{m,0} - t_{P\ m,0} \right] A_m \tag{IV-272}$$

For method C the A_m are $\Delta\xi$ for each value of m; this represents an integration of the area from $-\Delta\xi/2$ to $+\Delta\xi/2$.

$$A_m = \int_{-\Delta\xi/2}^{\Delta\xi/2} \sigma(\xi)d\xi = \Delta\xi \qquad \tfrac{1}{2} \le m \le S - \tfrac{1}{2} \qquad \text{(IV-273)}$$

However, for method G the weightings A_m for the summation are $\Delta\xi$ only for the interior points and are $\Delta\xi/2$ for the points adjacent to the fluid temperature.

$$A_m = \frac{\Delta\xi}{2} \qquad m = 0,\ S \qquad \text{(IV-274)}$$

$$A_m = \Delta\xi \qquad m = 1, 2, \ldots,\ S-1 \qquad \text{(IV-275)}$$

Thus, the weightings for method G are not directly analogous to the weightings in the continuous integration.

For a network of nodes to have an area equal to the integral of the volume weighting, the points must be located away from the surface. For example, for radial coordinates we would have:

$$A_m = \int_{\rho_m - \Delta\xi/2}^{\rho_m + \Delta\xi/2} 2\pi\rho\, d\rho = 2\pi\rho_m \Delta\rho \qquad \text{(IV-276)}$$

where $m\Delta\rho = \rho_m$ and $1/\Delta\rho = S$.

In order to use this expression to calculate all the A_m's, m must be $1/2, 3/2, \ldots, S-1/2$ or the points must be located a distance $\Delta\rho/2$

away from the surface (not at m=0 or m=S). The location of the points for more complicated geometries is not as clear; however, it might be possible using numerical techniques of orthogonalizing (31) to find such a system for a given geometry. However, based on the above discussion, locating the points away from the boundary appears to be advantageous. A different argument was used in Chapter III to justify locating the points away from boundaries about which the temperature distribution is symmetric, or an adiabatic boundary.

The importance of the accuracy of the initial vector components can be shown by a study of the application of generalized method A to problem II. This method is studied in detail only for problem II, as it appears to be suitable mainly for graphical methods. As the Y/A matrix is a function of the time increment for this method, the initial vector is also a function of the time increment as is shown by the series expansion for the initial vector in Table IV-18. The first term in an expansion for $(g_j - a_j)$ is proportional to (12r-1) and by making r=1/12 this term is made zero. A numerical calculation, based on this method for 5 points, gives a first initial vector component that agrees with the continuous component to within 0.01 per cent for generalized method A compared to an error for method G of about 1 per cent. The approximate temperatures found using generalized method A show an error that is from 1/50 to 1/100 of that of the approximate temperature found using method G for the same number of points and same time

increment, which shows that the accuracy of the initial vectors does greatly influence the accuracy of the solution.

However, in order to obtain an accurate solution even when the initial vectors are accurate, the oscillatory effects must be small. Consequently, if the time differencing parameters are selected so that q_{min} is negative, the weighting associated with this damping factor should be small, and the closer q_{min} is to q_{MAX} in absolute value, the smaller this weighting or amplitude should be. Since the eigenvector matrix elements all have about the same absolute value, which is less than or equal to one, except for method G applied to problem I, the size of the initial vector g_S or g_{S+1} determines the size of the amplitude. A study of the initial vector components together with a knowledge of the trigonometric roots that correspond to q_{min} indicates that the component of the initial vector is excessively large only for method C applied to problem I where the heat-transfer coefficient is large, $H > 2S$, and to problem II where H is infinity. For the other methods and problems, this component is sufficiently small and does not give amplitude to the oscillation. This means that, if the accuracy in method C is to be obtained, the differencing parameters must be selected so that q_{min} is either positive or so that $(q_{min})^{n_0}$ is negligible at the time of interest.

Several comments should be made about the possibly oscillatory components of the other methods. First, for stable method G applied

to problem I, the product $(g_{S+1}\, c_{m,S+2})$, which is the amplitude associated with the damping factor q_{min} and the complex α_{S+1}, is of the order of $1/S^2$, and should not cause any trouble as the amplitude for q_{MAX} is about 1. Even though the amplitude of q_{min} is small for graphical solutions using methods G and A, for problem III q_{min} is -1, and unless averaging is used the solutions are technically unstable.

Graphical Solutions--Effective Initial Vectors. The effective initial vectors for graphical solutions are defined by equation IV-256 and are found by taking the sum of the terms in the approximate solution containing damping factors of equal magnitude but opposite sign. Their components are

$$g_{j\,E} = \left[g_j + (-1)^n\, g_{S+1-j}\, \frac{c_{m\,S+1-j}}{c_{mj}} \right];\ j < \frac{S-1}{2} \qquad \text{(IV-277)}$$

The product of the jth effective initial vector component with a component of the jth eigenvector gives the intercept of the semi-logarithmic plot of a transient term for either n even $[g_j + g_{S+1-j}\,(c_{m\,S+1-j}/c_{m\,S+1})]$ or n odd $[g_j - g_{S+1-j}\,(c_{m\,S+1-j}/c_{m\,S+1})]$. This is shown in Figure IV-15 for graphical method A and problem I. Consequently, for a graphical solution where equal and opposite damping factors occur, in determining the accuracy the important difference is $(g_{j\,E} - a_j)$ and not $(g_j - a_j)$.

Although the effective initial vector components exist and determine the accuracy for graphical methods G, A, and C applied to

problems II and III and for graphical method A applied to problem I, the effective initial vectors have been derived analytically only for problem II, and a brief numerical study has been made of the effective initial vector component for q_{MAX} for graphical method A applied to problem I. The series expansions for these components are in Table IV-21 for problem II and from a study of them the following conclusions are made:

(1) Of the graphical methods shown, graphical methods A and G are the most accurate, and the highest accuracy occurs when the quantity (m+n) is odd. Even under these circumstances, the first error term of an expansion of $(g_{j\,E} - a_j)$ for an (n+m) odd which is $+\mu_j/3S^2$ is twice the error term for the expansion of $(g_j - a_j)$ for generalized method G with no oscillations or 4 times the first term of the difference for generalized method C with no oscillations.

(2) Graphical method C never gives accurate results for problem II as its first error term in $(g_{j\,E} - a_j)$ which is $[(-1)^{n+m-\frac{1}{2}}\, a_j/2m]$ can be considered a zero order term because it does not go to zero as S goes to infinity. This first error term of the effective initial vector also completely explains the behavior shown in Figure IV-1 for an m of $\frac{1}{2}$, as $(g_{1\,E} - a_1)$ would be expected to be either $-a_j$ or $+a_j$ making $g_{1\,E}$ oscillate between zero and more than twice its correct value.

The numerical study for method A and problem I shows that the component for the slowest decaying damping factor g_1 is always

greater than the corresponding a_1, and that the compensation obtained when the term $[(-1)^n g_{S+1} (c_{m\,S+1}/c_{m1})]$ is negative accounts for the more accurate effective initial vector. This occurs for problem I when $(n+m)$ is odd and this is the time at which the graphical result should be applied.

Thus, in order for graphical methods with equal and opposite damping factors to be accurate, the size of the oscillatory part of the effective initial vector, $[g_{S+1-j} (c_{m\,S+1-j}/c_{mj})]$, should be of the same order of magnitude as $(g_j - a_j)$ and further, one should know at which odd or even n to use the graphical solution calculated at a point. For graphical method A and the problems here, the correct order of magnitude does occur on the oscillatory part of the initial vector and the more accurate solutions occur for (n+m) odd; however, both the amplitude of the oscillations and the error in $(g_j - a_j)$ for this method should be further studied. Possibly, means should be devised by which the sign of $(g_j - a_j)$ could be predicted, and when the oscillatory part compensates.

It should be added that a procedure similar to the effective initial vector method could be used to analyze graphical methods where the damping factors are not equal and opposite as, for example, graphical method F applied to problem II. In this case, the two damping factors that are closest in absolute value but that have opposite signs could be paired and assuming that j and (S+1-j) are paired and that

q_{S+1-j} is negative, the sum of the corresponding two terms in the transient solution is:

$$g_j c_{mj} q_j^n + g_{S+1-j} c_{m\,S+1-j} q_{S+1-j}^n =$$

$$\left[g_j + \left(\frac{q_{S+1-j}}{q_j} \right)^n \frac{c_{m\,S+1-j}}{c_{mj}} \right] c_{mj} q_j^n \qquad \text{(IV-278)}$$

$$1 \le j \le \frac{S}{2}$$

The effective initial vector would be like that before but with the ratio $[(q_{S+1-j}/q_j)^n]$ instead of $(-1)^n$ for the oscillatory term. In this case, the effective initial vector is not constant at alternate time intervals but oscillates about g_j with decreasing amplitude. Because of this decreasing amplitude, even a numerical comparison of these effective initial vectors is not useful.

Because graphical method F is the most accurate of the graphical methods for this problem, its solution is compared numerically to the continuous solution for 5 points in Table IV-22; graphical method A is compared similarly in Table IV-23. The success of graphical method F is due to a complicated system of cancelling errors; however, two characteristics which are apparent for method F and problem I are:

(1) The fact that $|q_{min}|$ is significantly less than q_{MAX} allows a larger initial vector component for q_{min}, and this oscillatory component is rapidly damped.

(2) Both the intercepts for j of 1 and 2 are more accurate than are the corresponding intercepts for method A.

Averaged Methods. The advantage of averaged methods is that they greatly reduce the amplitude of the oscillations of any negative damping factors, as discussed in Chapter III and Chapter IV, section B-7. It was mentioned that an averaged solution has as its initial vector components $[(q_j+1)/2q_j^{\,k}]$ times the unaveraged solution initial vector.

$$g_{j\,Ave} = \frac{q_j+1}{2q_j^{\,k}} \qquad \text{(IV-279)}$$

The quantitative effect of this on the accuracy of the initial vector components is studied by expanding $[(q_j+1)/2q_j^{\,k}]$ into a series. Multiplication of this series times an expansion for the unaveraged initial vector components gives a series for the averaged initial vector components, and subtraction of a_j gives the required expression. In Table IV-24, the series for $[(q_j+1)/2q_j^{\,k}]$ is shown together with the series for the special cases of k: k=0 the forward average, k=$\frac{1}{2}$ the central average, and k=1 the backward average. Carrying out of the multiplication of the series for $[(q_j+1)/2q_j^{\,k}]$ times the series for the unaveraged initial vector gives the series for the averaged initial vector. This series shows that the major effect of averaging is that the term $[a_j(k-1/2)r\mu_j^{\,2}/S^2]$ is added to the first term for the expansion for

the error in the unaveraged initial vector component $(g_j - a_j)$ to obtain the error in the averaged initial vector component $(g_{j\,Ave} - a_j)$. For the three averages above this term is: forward, $(-a_j r \mu_j^2/2S^2)$; central, 0; backward, $(a_j r \mu_j^2/2S^2)$. Consequently, if an averaged method is desirable for reducing oscillations, and an unaveraged method is known to have very accurate initial vector components, the central average might be the best. On the other hand, if the error in the unaveraged initial vector components is positive, $(g_j - a_j) > 0$, then the forward average should be used; if the reverse is true the backward average is best.

In Tables IV-25, IV-26, and IV-20, the expressions for the initial vector components for the averaged solutions and for the series expansions for $(g_{j\,Ave} - a_j)$ are shown. It should be remembered that the forward average is used with methods A and F and the backward average with methods G and C. In these tables, the coefficients in the series expansions that involve r and γ are the terms added because of averaging. In some cases, the first error term in the expansion for $(g_{j\,Ave} - a_j)$ shows that averaging can compensate; for example, this is true for backward averaged method G applied to problem II. In this problem, this error term is zero if r is 1/6; however, it cannot be generalized to other methods as seen from the other expansions. In particular, when averaging method C for problem II, the error would actually be increased by using the backward average, but, for method

C and problem III, the backward average compensates and the first error term is zero for an r of 1/12. For method C, because of its accurate initial vectors, the central average is probably best. Indeed, from the series for $[(q_j+1)/2q_j^k]$ and the above discussion on forward and backward averages, the important conclusion is made that the central average does not introduce significant error in the initial vector component and, in general, when averaging is necessary to reduce oscillations, the average of the solutions at times t_n and t_{n+1} usually should be applied at time $(n+1/2)\Delta\tau$.

The series in Table IV-24 cannot be used directly on the effective initial vector components to generate the expansion for the effective initial vector components for an averaged graphical method. These expansions can be found by considering the series for the effective initial vector components as the sum of two series; one called the oscillatory series contains all the terms with $(-1)^n$ as a factor; the second, called the non-oscillatory series, contains the remaining terms. The non-oscillatory series for the expansion of the averaged effective initial vector is then the product of the series for $[(q_j+1)/2q_j^k]$ and the non-oscillatory series for the unaveraged solution which is the same as the series for the averaged solution. The comments just made for this series then apply directly to this part of the effective initial vector for the averaged graphical solution. The oscillatory series for

the average effective initial vector components is found by multiplying the oscillatory series for the unaveraged solution by the series for $[(1-q_j)/2q_j^k]$. The series for $[(1-q_j)/2q_j^k]$ for graphical solutions are shown in Table IV-27 in terms of k and for k's of 0, $\frac{1}{2}$, and 1. Each of these series has the same first term, $\mu_j^2/4S^2$. Thus, the major effect of multiplying the oscillatory series by the series for $[(1-q_j)/2q_j^k]$ is that the largest oscillatory term is multiplied by $\mu_j^2/4S^2$.

The effective initial vector components for averaged graphical solutions G, A, and C applied to problem II are shown in Table IV-28, together with the series expansions for the error, $(g_{j\,E\,Ave} - a_j)$. (Note that the forward average is associated with method A and backward averages with methods G and C.) These results show that for averaged methods G and A the oscillatory term now appears only in the second- and higher-order terms and is of the order of $1/S^4$ rather than $1/S^2$ as in the effective components for the unaveraged graphical solution. However, the first error term for the effective initial components for averaged graphical method A is twice that for generalized method G's components, as pointed out previously. The semi-logarithmic plot of the first transient term with time for averaged graphical method A for odd n is coincident with the dot-dash line with the small circles in Figure IV-15; the line for even n has an intercept that is less than 0.2 per cent above that shown for odd n, and is not shown in Figure

IV-15, as it lies within the line thickness for even n. For method C, an oscillatory part still appears in the first error term, but this term is now proportional to $1/S^2$ compared to being independent of S for the unaveraged method. Averaged method C actually gives a ramp-plateau temperature-time plot as does unaveraged G shown in Figure IV-1 and a second averaging would be required to reduce its oscillations to the $1/S^4$ term as for graphical method A.

Thus, for some graphical methods, averaging essentially reduces the oscillatory effect by an order of $1/S^2$; however, for a graphical method which has a tendency to weight the oscillatory component heavily, such as happens for method C, the oscillatory effect is still significant. Further, when using either the forward or backward average, the accuracy of the initial vector is still dependent upon unpredictable error in (g_j-a_j), but, for the problems considered here, graphical method A gives compensation when the forward average is used. This graphical method should be further investigated to see if the average always compensates; a possible reason for this compensation could be that a backward difference was used in deriving the surface equations.

Conclusions and Summary--Initial Vectors. Two reasons have been shown why the initial vector components are important to the accuracy of approximate solutions. First, the difference between the initial vector components, (g_j-a_j), for the larger damping factor terms can make a large difference in the error in the approximation, and

second, if the initial vector component for the minimum damping factor, q_{min}, is large in absolute value, it can ruin the accuracy of an approximation when $|q_{min}|$ is equal to or just smaller than q_{MAX}. Since the initial vector components depend upon the specific initial temperature distribution, and do not have the same form for each method and problem, the general conclusions which can be made about the methods are not as quantitatively precise as are the generalizations about the damping factors.

Method C was shown to have the most accurate unmodified initial vector components of any of the methods for the problems discussed. The probable reason for method C having the most accurate initial vector is that its orthogonality relationship has the same form of volume weighting factor as does the continuous solution. Consequently, method C not only has the most accurate initial vectors for the problems with the specific initial conditions studied here, but probably has the most accurate initial vector for other initial conditions. However, the initial vector component g_S for the minimum damping factor q_{min} is very large for problem I with a large heat-transfer coefficient, and for problem II this method is not suitable for use unless the time differencing parameters $\Delta\tau$ and γ are selected so that q_{min} has a small absolute value or is positive; therefore, it is not suitable for use as a graphical method for this type of problem.

The fact that method C has the most accurate initial vectors, and that its orthogonality relationship has the same form of volume weighting as the continuous solution, was shown to be due to the location of the adjacent points at a distance $\Delta\xi/2$ from the surface. It was also shown that to obtain the directly analogous form for the volume weighting for a regular mesh spacing for radial coordinates the adjacent nodes must be located at distances $\Delta\rho/2$ from the center and from the cylindrical surface. This leads to the important conclusion that for problems where a regular mesh in a coordinate system appropriate to the geometry is used, more accurate initial vectors are obtained if the adjacent points are located a half increment away from the boundaries. This conclusion applies to two- and three-dimensional problems also. Even for irregular networks, location of the adjacent points away from the surface might be advantageous.

Graphical method A was shown to give an accurate approximate solution to the problems considered here because the oscillatory part of its effective initial vector component was about the same size as the error in the corresponding initial vector component, and part of the error is cancelled for alternate values of n when $(n+m)$ is odd. The oscillatory part of the effective initial vector for this graphical method and problem is small and thus, for these problems, accurate solutions are obtained. Thus, graphical method A does have a tendency toward having only a small oscillatory part of the effective initial

vector. However, in order to generalize these results for graphical methods, its effective initial vector components $g_{j\,E}$ should be further studied by deriving solutions for other problems by studying the orthogonality relationships, and by trying to relate the space discretization error. These studies should give information about the size and direction of the error in the initial vector components $(g_j - a_j)$ for the slower decaying positive damping factors, about the size of the oscillatory part of the effective initial vector components, $[(-1)^n g_{S+1-j} (c_{m\,S+1-j}/c_{mj})]$, and about at which values of n, odd or even, the most accurate effective initial vector components occur. However, for graphical method A applied to problem II, the error in the effective initial vector components for (n+m) odd, when the oscillatory part compensates, is still four times the error in the initial vector components of generalized method C.

Averaging a solution, and applying the average at the center of the time interval, reduces the oscillations and changes the initial vector components only insignificantly. Applying the average at the beginning or end of the time interval (forward or backward averages) can change the unaveraged initial vector. However, for graphical method A, the initial vector appears always to have positive error in its unaveraged initial vector components, $(g_j - a_j) > 0$, and the forward average compensates this error and improves the accuracy of graphical method A. The above mentioned study for the initial vector of graphical

method A should also show if the error $(g_j - a_j)$ is always positive and, consequently, if the forward average always should be associated with graphical method A.

Although from the expressions for components of the eigenvectors and the initial vector it is not possible to derive a general expression for the error in the intercept on a semi-logarithmic plot, one concludes that this error is proportional to $1/S^2$. Consequently, in trying to select the differencing parameters or to bound the error, the following rough estimate of this error is suggested:

$$g_j c_{mj} - a_j b_j (m\Delta\xi) < \Phi \frac{\mu_j}{S^2} \qquad j = 1, 2, \ldots, S \qquad \text{(IV-280)}$$

where Φ is a scaling factor which depends upon the range of temperatures present initially in the solid and the surrounding fluid. It can usually be taken as:

$$\Phi = \underset{m_1, m_2}{\text{MAX}} \left| t_{m_1,0} - t_{m_2,0} \right| \qquad \text{(IV-281)}$$

where m_1 and m_2 are taken on all values for the mesh used and f_0 and f_S of the fluid temperatures. This expression is conservative giving an intercept error 12 times that observed for method C for problems II and III and 6 times that for method G. Because of this conservative estimate, it can also be used for graphical methods and averaged methods. Further studies would be required to find a less conservative relationship.

7. Problem V--Complete Problem

Problem V is the limiting case of problem II for the solid that extends to infinity in the ξ direction. The continuous solution and the solution for method G are shown in Table IV-29 for comparison. Both trigonometric expressions for the solution are based on replacing the infinite or finite sum in the transient solution with an integral. Since S is now infinite, $\Delta\xi$ is used in the solution. Based upon the previous discussion of each of the terms in the integral for the approximate solution, the conclusion is that the errors are the same as for the finite solid. Further, because of the similarity of form, the approximate solutions for each of the other methods can be found by taking limits, and then the conclusion is that all comments made concerning problem II apply to problem V also.

A further study and comparison of solutions to this problem might lead to better understanding of the accuracy of the approximations for finite solids at short times. This probably could be done because the continuous solution for a finite solid can also be expressed as an infinite sum of error function terms, and only one or two of these terms are significant at very low times. The continuous solution for the infinite solid as shown in the table is actually the first term of this summation. Further, since the binomial distribution is a finite difference approximation to the normal distribution, a "binomial error function" might be related both to the trigonometric integrals for the

approximate solution and to the normal error function. A comparison between the approximate solution in terms of a "binomial error function" and the continuous solution in terms of its error functions could lead to useful conclusions about the accuracy of approximate solutions at very short times. This has not been attempted.

8. Selection of Differencing Parameters.

As a review and application of the equations derived, a procedure is developed that allows the selection of differencing parameters, S, $\Delta\tau$, and γ, in such a way that the error in the approximate transient solution is never larger than V at any time larger than τ_0. These parameters are also selected by this procedure so that the number of non-zero multiplications required to carry the calculation to a time τ_1 is close to the actual minimum necessary for a given accuracy. This procedure is most useful when a very accurate solution is required at either short or intermediate times, although it can also be used for long times. In this procedure the assumption is made that graphical method A is to be used for any graphical solution, and that method C is to be used for any other calculation. The procedure is based on equation IV-218 which is, together with equation IV-216,

$$V = J\left|g_J\, c_{mJ} - a_J\, b_J(m\Delta\xi) + g_J\, c_{mJ}\left\{1 - \left(\frac{e^{-v_J^2}}{q_J}\right)^{\tau_0}\right\}\right| q_1^{n_0} \geq |v_{m,n}| \qquad \text{(IV-282)}$$

This equation, combined with some of the following remarks, can be used to estimate the error bound V for a given selection of differencing parameters; this is a straightforward calculation and need not be discussed. As shown in section E-2 of this chapter, if the error bound V is equally divided between the intercept and the slope of the semi-logarithmic graph of a transient term with time, the equations which assure that the first J significant terms are sufficiently accurate are given in equations IV-219 and IV-220, and are:

Accuracy of Intercept

$$\frac{V}{2Jq_1^{n_0}} \geq \left| g_J c_{mJ} - a_J b_J (m\Delta\xi) \right| \qquad \text{(IV-283)}$$

Accuracy of Slope

$$\frac{V}{2Jq_1^{n_0}} \geq \left| g_J c_{mJ} \left\{ 1 - \left(\frac{e^{-\nu_J^2}}{q_J^{1/\Delta\tau}} \right)^{\tau_0} \right\} \right| \qquad \text{(IV-284)}$$

The size of the oscillatory component must also be controlled and this restriction is equation IV-222 for solutions other than graphical

$$V >> g_S c_{mS} Y^{n_0} \qquad \text{(IV-285)}$$

where $g_S c_{mS} Y^{n_0}$ can be taken as no larger than 10 per cent of V and probably smaller than 1 per cent. The intercept error is given by

equation IV-280, which is

$$\varepsilon \frac{\mu_J}{S^2} > \left| g_J c_{mJ} - a_J b_J (m\Delta\varepsilon) \right| \qquad \text{(IV-286)}$$

The ratio $e^{-\nu_J^2}/q_J^{1/\Delta\tau}$ can be found from equation IV-260,

$$\ln \frac{e^{-\nu_J^2}}{q_J^{1/\Delta\tau}} = \mu_J^2 - \nu_J^2 + \mu_J^4 \left[(\tfrac{1}{2} - \gamma)\Delta\tau - \frac{1}{12S^2} \right] + \qquad \text{(IV-287)}$$

$$+ \mu_J^6 \left[(\tfrac{1}{3} - \gamma - \gamma^2)(\Delta\tau)^2 - \frac{(\frac{1}{2} - \gamma)\Delta\tau}{12S^2} + \frac{1}{360S^4} \right] \cdots$$

This relation requires a knowledge of $(\mu_j^2 - \nu_J^2)$ or a way to estimate this quantity; this relationship is known for problems with boundary conditions of the types of problems II, III, and IV, but is not yet known for problem I and method C and graphical method A. The bound for $|q_{min}|$, Y, is found for method C from Figure IV-14 or equation IV-252, which is, when solved for Y:

$$Y = \frac{4r(1-\gamma)-1}{4r\gamma + 1} \geq |q_{min}| \qquad \text{(IV-288)}$$

In order to use these equations the following quantities, which are functions of the physical system of the problem, must be estimated:

1. The slowest-decaying damping factor to the nth power, $q_1^{n_0}$.

This can be taken equivalently as:

$$q_1^{n_0} \simeq e^{-\mu_1^2 \tau_0} \simeq e^{-\nu_1^2 \tau_0} \qquad \text{(IV-289)}$$

2. The number of significant terms in the transient solution, J.

3. The Jth trigonometric root, μ_J or ν_J.

4. The size of the initial vector component-eigenvector component products, $g_J c_{mJ}$ for the smallest significant term and $g_S c_{mS}$ for the term that can have the negative damping factor.

Since only estimates of these quantities are required, either of the corresponding continuous or approximate quantities, as indicated above, is satisfactory. Usually, in the following, the approximate quantity is used because once the Y/A matrix and $\Delta\tau$ and γ are fixed, each of the other quantities can be found numerically by appropriate matrix manipulations.

In most cases, the easiest way to estimate the first quantity, $q_1^{n_0}$, is to find an estimate for either μ_1 or ν_1, as an estimate of this trigonometric root is necessary for later use. This can be done in several ways. First, if only a rough estimate is needed, μ_1 can be estimated on the basis of previous experience. Second, if the exact continuous solution can be found for a problem with the same boundary equations but possibly different fluid temperature forcing function and/or different initial conditions, then ν_1 can be found exactly. Third, a trial approximate calculation can be stepped out for a coarse difference mesh for the solution to a problem where the fluid

temperature function gives a steady-state solution. This calculation must be carried to the time when a semi-logarithmic graph for an approximate temperature is a straight line. A graphical construction for about four or five points should give a straight line after about 10 to 15 time increments; its slope with n is $[\cos(\mu_1/S)]$ where n and S are based on the coarse grid. A backward difference implicit calculation with γ of 1 could also be carried out for four or five points with a very large time increment of possibly 0.1 for two or three increments. The angle parameter μ_1 can be estimated from the semi-logarithmic slope and

$$\cos\frac{\mu_1}{S} = \frac{q_1 - 1 + 2r(\gamma q_1 + 1 - \gamma)}{2r(\gamma q_1 + 1 - \gamma)} \qquad \text{(IV-290)}$$

where q_1 is the slope with n and n, r, S, and γ are the parameters for the coarse grid used. After the estimate of μ_1, the quantity $q_1^{n_0}$ can be found from equation IV-289.

The second quantity, the number of significant terms, J, in the solution, can then be estimated by first finding the second and higher order μ_j's from equation IV-245, which is

$$\mu_{j+1} = \mu_j + \pi \qquad \text{(IV-291)}$$

Then, by calculations of $e^{-\mu_j^2\tau_0}$ for the second- and higher-order terms, one finds the Jth term such that the (J+1) term is only a small

fraction of the desired error bound, probably less than 10 per cent of the error bound V. Also, the Jth angle parameter μ_J is estimated and is the third quantity above.

The fourth and last quantity to be estimated is the product $g_J\, c_{mJ}$ for the last significant term and $g_S\, c_{mS}$ for the damping factor which can be negative. These can be taken as the same quantity independent of the j subscript,

$$g_j\, c_{mj} = \Phi \tag{IV-292}$$

For most initial distributions, this gives too large a product for the possibly negative damping factor term; thus, it is a conservative assumption. For an averaged method, the $g_S\, c_{mS}$ product should be taken as Φ / S^2.

Thus, from these estimates $q_1^{n_0}$, J, μ_J, $g_J\, c_{mJ}$, and $g_S\, c_{mS}$ are known, and all the quantities in the restriction equations can now be found as a function of the differencing parameters. The inequality for the intercept error, equation IV-283, can be simplified by substituting for this error from equation IV-286 and solving for S. This gives

$$S \geq \sqrt{\frac{2J q_1^{n_0} \Phi \mu_J}{V}} \tag{IV-293}$$

The inequality for the accuracy of the damping factors, equation IV-284, can also be simplified to

$$\frac{V}{2Jq_1^{n_0}\Phi} \geq \left| 1 - \left(\frac{e^{-\upsilon_J^2}}{q_J^{1/\Delta\tau}} \right)^{\tau_0} \right| \qquad \text{(IV-294)}$$

where the ratio $(e^{-\upsilon_J^2}/q_J^{1/\Delta\tau})$ is given by equation IV-287. Also the oscillatory behavior restriction, equation IV-288, can be found for method C by substitution in that equation for $g_S c_{mS}$ from equation IV-292 and for Y from equation IV-288 obtaining:

$$\frac{V}{\Phi} >> \left[\frac{4S^2\Delta\tau(1-\gamma)-1}{4S^2\Delta\tau\gamma + 1} \right]^{\tau_0/\Delta\tau} = (Y)^{\tau_0/\Delta\tau} \qquad \text{(IV-295)}$$

The double inequality sign can be taken to mean that the quantity on the right is a certain very small fraction of V, probably 0.01 or less, when finding parameters that satisfy this inequality. The problem then is to find the selection of differencing parameters S, $\Delta\tau$, and γ which satisfies the above three inequalities or the graphical solution parameters with $\nu = 0$ and $r = \frac{1}{2}$ which satisfy the two inequalities for accuracy, equations IV-293 and IV-294, and that, at the same time, minimize the number of calculations. This can be done by first analyzing the graphical solution set of parameters, then other explicit methods, and finally the general implicit method. This sequence is used because each analysis of a type of solution can reduce the range of variables considered for the next type of solution.

The number of points to be used for any of the three solutions is found directly from equation IV-293 using the equality sign. This is a very conservative estimate based on the intercept accuracy and it applies to graphical and averaged methods, in addition to other generalized methods. However, as pointed out previously, graphical methods, even allowing for compensating errors of the oscillatory component, do require more points than solutions where Y^{n_0} is made negligible. Anticipating that future studies might indicate that a larger proportionality constant would be associated with the error in graphical solution $g_j c_{mj}$ products than for generalized methods, a larger number of points, S_G, is associated with graphical solutions, compared to S points for implicit or other explicit solutions:

$$S_G \geq S \qquad \text{(IV-296)}$$

(The proportionality constant used in equations IV-286 and IV-293 is one.) The above number of points is based upon accuracy of the intercept and thus is a minimum number of points. In some cases, more points (larger S) than that given by equation IV-293 might have to be used to obtain a mesh point at a desired location, where interpolation was not deemed satisfactory; or to adequately describe a very cyclic initial temperature distribution; or a temperature distribution that is expected to show many cycles with the space variables.

Because the graphical solution with r_G or $s_G^2 \Delta\tau_G$ of $\frac{1}{2}$ and $\gamma = 0$ requires the least number of calculations per time increment, or possibly can be constructed graphically, this solution is considered first. After an S_G has been determined from equation IV-293 or a version of this equation is modified with a different proportionality constant, the accuracy of the damping factors for the graphical solution should be checked by calculating the ratio $e^{-\nu_J^2}/q_J^{1/\Delta\tau}$ from equation IV-287 to see if this ratio satisfies the inequality IV-294. If this inequality is not satisfied, then the number of points for a graphical solution must be further increased until the inequality is satisfied.

Although, in theory, the solution for the selection of parameters of $r = \frac{1}{2}$ and $\gamma = 0$ can be constructed graphically for any value of S_G, for S_G greater than about eight, one requires too large a sheet of graph paper to obtain an accurate construction. The graphical equivalent of numerical round-off error then obscures the solution. Consequently, for an S_G larger than eight, the calculation must be stepped numerically using a desk calculator or digital computer which carries a sufficient number of significant figures. The total number of non-zero multiplications required is then

$$N_G = 2S_G^3\tau_1 \qquad \text{(IV-297)}$$

Since any other explicit calculation requires twice the number of multiplications per time increment. and any implicit calculation requires seven times the multiplications, the only other calculations

that give fewer multiplications are an explicit calculation with

$$r_{Ex} > \left(\frac{S}{S_G}\right)^3 \qquad \text{(IV-298)}$$

and an implicit calculation with

$$r_{Im} \geq 3.5\left(\frac{S}{S_G}\right)^3 \qquad \text{(IV-299)}$$

Because of the third power of the ratio S/S_G in these inequalities, the ratio S/S_G must be close to one if the graphical solution is to be competitive. The following calculations make this clear. First, as the r_{Ex} must be less than $\frac{1}{2}$ when Y^{n_0} is to be small, no explicit calculation except the graphical solution need be considered if

$$\frac{S}{S_G} > 0.79 \qquad \text{(IV-300)}$$

However, if, as indicated by comparing the errors for the effective initial vector components for either averaged or unaveraged graphical method A with those for method C when applied to problem II, the graphical solution requires twice the number of points as does a solution where Y^{n_0} is negligible, the r's that should be considered are:

$$r_{Ex} > 0.125 \qquad \text{(IV-301)}$$

$$r_{Im} > 0.44 \qquad \text{(IV-302)}$$

And since, in any case, an r_{Ex} of 1/6 meets the non-oscillatory requirement and very probably the damping factor accuracy requirement, as these conditions give the most accurate damping factors, the graphical solutions with r of $\frac{1}{2}$ and γ of zero should not be used. However, this statement applies for cases where the very accurate solutions are required and/or when the calculations are to be carried out on a desk calculator or a digital computer, and it does not mean that a graphical construction cannot give an approximation of sufficient accuracy and be available to the engineer much more rapidly (and possibly more cheaply) than if the problem were done on a computer. The above discussion shows the importance of finding a more precise way of estimating the error in the intercept product $(g_j c_{mj})$ than equation IV-286 for both graphical method A and method C. Also it indicates that unless the number of points S_G for the graphical solution is close to S the graphical solution is not competitive.

Next, depending upon the size of S/S_G the maximum r or $\Delta\tau$ for the fixed S should be found that meets the oscillatory and accuracy restrictions for the explicit calculation. This r_{Ex} is probably limited by the oscillatory restriction in most cases, which in this case is, from equation IV-295:

$$\frac{V}{\Phi} >> \gamma^{\tau_0/\Delta\tau_{Ex}} = (4S^2\Delta\tau_{Ex} - 1)^{\tau_0/\Delta\tau_{Ex}} \qquad \text{(IV-303)}$$

and from a trial and error solution of this to find the maximum satisfactory $\Delta\tau_{Ex}$ and Y allowable for a fixed S, the maximum time increment is estimated. In this trial and error solution, the double greater-than sign means that Y^n is equal to a small fraction of V/Φ. The accuracy of the Jth damping factor for the resulting $\Delta\tau_{Ex}$ is then checked to be sure inequality IV-294 is satisfied, using equation IV-287 to find $e^{-\mu_J^2}/q_J^{1/\Delta\tau}$. If this is not satisfied, accuracy rather than oscillatory behavior limits r_{Ex}, and r_{Ex} or $\Delta\tau_{Ex}$ must be reduced until equation IV-294 is satisfied. The total number of multiplications is

$$N_{Ex} = \frac{2S^3\tau_1}{r_{Ex}} = \frac{2S\tau_1}{\Delta\tau_{Ex}} \qquad \text{(IV-304)}$$

Two other important points about the explicit method are that a non-oscillatory solution occurs for an r_{Ex} of $\frac{1}{4}$ and that for an r of 1/6 the γ_o given by equation IV-262 is 0. This means that the damping factors obtained for an r of 1/6 are close to being the most accurate damping factors possible for a given S and, if the accuracy criteria are not satisfied, S probably must be increased. Also, this means that implicit calculations should be considered only for r's larger than 0.58 under any circumstances.

The implicit calculation should be studied only for r's greater than those given in equation IV-299 and in

$$r_{Im} > 3.5\, r_{Ex} \qquad \text{(IV-305)}$$

Since any implicit calculation can be made non-oscillatory by increasing γ the oscillatory restriction does not limit the size of r_{Im} or $\Delta\tau_{Im}$ directly. However, increasing both γ and $\Delta\tau_{Im}$ reduces the accuracy of the damping factors and the inequality IV-294 for damping factor accuracy actually limits the maximum $\Delta\tau_{Im}$ or r_{Im} to be used. A way to find the maximum r_{Im} which satisfies this inequality, IV-294, is first to find the maximum r_{Im} which satisfies the oscillatory restriction and the relationship for γ_o, equation IV-262. Substituting for γ_o from equation IV-262 into the oscillatory restriction, equation IV-295, we obtain:

$$\frac{V}{\Phi} >> \left[\frac{3S^2\Delta\tau - 1}{3S^2\Delta\tau + 1}\right]^{\tau_0/\Delta\tau} \qquad \text{(IV-306)}$$

Again, a trial and error calculation is required to obtain the largest $\Delta\tau_{Im}$ that satisfies this inequality. Once the largest $\Delta\tau_{Im}$ that satisfies the above equation IV-306 is found, the ratio $e^{-v_J^2}/q_J^{1/\Delta\tau}$ is calculated and the accuracy of the Jth damping factor is checked. If the accuracy is not satisfied, a smaller time increment and corresponding γ_o should be selected until the accuracy requirement is satisfied. On the other hand, if the accuracy requirement is satisfied, the time increment or r possibly can be increased if the weighting γ is also increased, so that the increased time increment

does not introduce excessive oscillations. A time increment $\Delta\tau_{Im}$ which can be used then occurs at a point where both the accuracy and the oscillatory restrictions limit any further increase in the time increment. This limiting time increment can be found by selecting a large suitable value of r, finding the minimum weighting γ that allows the oscillatory restriction to be satisfied (this requires a trial and error calculation of equation IV-295) and then checking the accuracy of damping factors for this combination of S, γ, and r. Repeating this procedure for several values of r should allow a good estimate of the maximum r_{Im} or $\Delta\tau_{Im}$ which satisfies both the oscillatory and accuracy limits. The number of non-zero multiplications for the implicit method is:

$$N_{Im} = \frac{7S^3\,1}{r_{Im}} = \frac{7S\tau\,1}{\Delta\tau_{Im}} \qquad \text{(IV-307)}$$

The calculation giving the minimum N of N_G, N_{Ex}, or N_{Im} is then used for the approximate solution. If the graphical solution N_G is smallest, graphical method A should be used; if N_{Ex} or N_{Im} is the smallest, method C should be used. In some cases, not all these N's would be computed because the several calculations can eliminate further consideration of either the explicit or even the implicit calculation.

For some problems, the maximum time increment which can be

used is not fixed by either the oscillatory behavior, equation IV-295, or the accuracy of the damping factors, equation IV-294. If the forcing functions are oscillatory, the time increment must be smaller than one-half the smallest period of oscillation to obtain an adequate description of the forcing function (see Chapter III, section C). In this case, when the maximum time increment is reached by the above procedure, the best selection of parameters has been made. If this time increment is smaller than any obtained in the calculations above it should be used with an explicit method directly.

Although the above procedure is probably too time consuming to be carried out in detail, it should prove to be a useful guide in selecting these parameters. The exact maximum values of r_{Ex} and r_{Im} do not have to be found as one is not usually interested in 5 or even 10 per cent larger values of r, but is trying to double or triple this value. Therefore, only estimates of these quantities are required. Further, after finding the differencing parameters for many problems and different accuracies, a table of these quantities could be developed, and further generalizations and rules for selecting these values should become apparent.

(F) CONCLUSIONS AND SUMMARY--CHAPTER IV

The z-transforms have been successfully used to find the analytic solutions of the one-dimensional partial difference equation of diffusion for several different methods of approximating the continuous boundary equations. These solutions have been compared with the analytic solutions of the approximate partial differential equations for the same problems. The methods for approximating the boundary equations are based on a mesh with adjacent points on the boundary (mesh $\Delta\xi$), methods G and A, and on a mesh with the adjacent points located a distance $\Delta\xi/2$ from the boundary (mesh $\Delta\xi/2$), method C. Use is made of both the usual differencing for the approximations at the boundaries (methods G and C) and a backward difference for the known fluid temperature for method A. Consideration has been given to all selections of differencing parameters from graphical solutions to implicit calculations. The z-transform solutions have proved very useful for studying and understanding quantitatively the errors and oscillatory behavior of these approximate solutions. Not only can the effect of the differencing parameters be studied quantitatively, but the consistent superiority of the accuracy of generalized method C and graphical method A of the graphical methods has been shown analytically. However, these complete analytic solutions, as found by the

z-transforms, are not the most practical way to find the sufficient conditions on the differencing parameters to obtain a solution with satisfactory oscillatory behavior, and are not the best way to study the bound for $|q_{min}|$ which is Y. Simply calculated matrix norms and the techniques of Chapter III should be used for this purpose.

In addition, the following observations and conclusions are made:

(1) When the particular solution is a true steady-state solution, the steady-state solution for the approximate solution is equal to that for the continuous solution for all methods. For problems where a quasi-steady-state solution occurs, the same comments made about the eigenvectors and damping factors for the transient solution probably also apply to the particular solution.

(2) The error in the intercept $g_j c_{mj}$ of a slowly decaying term in the approximate transient solution, on a semi-logarithmic plot versus time, is approximately proportional to $1/S^2$. The error in the slope $[(1/\Delta\tau)\ln q_j]$, of a slowly decaying term in an approximate transient solution on such a plot is approximately proportional to $1/S^2$ and to $\Delta\tau$; the proportionality constant for $\Delta\tau$ is a function of the weighting γ. The errors, both in the intercept and in the slope, are of importance in obtaining accurate transient solutions. These two quantitative statements above are of much more practical importance in selection of the differencing parameters, or in estimating the error,

than the statements that the truncation error for the difference used to approximate the differentials is of the order of $1/S^2$ and of the order of $\Delta\tau$. Further, the error propagation analysis (2) based on the truncation error does not usually lead to such useful or precise quantitative statements.

(3) The trigonometric characteristic roots for the approximate solution are not equal to those of the continuous solutions for problems with a finite non-zero heat-transfer coefficient for any of the methods; the difference between the squares of these roots goes to zero with $1/S^2$. Errors in these roots affect the accuracy of each of the parts of the approximate solution. For methods G, A, and C, for problems with only zero and/or infinite H's, these roots are equal.

(4) The important effects on the accuracy of the slope of a semi-logarithmic graph of a transient term for the slower decaying positive damping factors in the approximate solution caused by changes in the damping factors are: first, for methods and problems where the approximate trigonometric roots are equal to those of the continuous solution, increasing S does not improve the accuracy of the slope as much as making appropriate changes in the time increment $\Delta\tau$ and weighting γ, providing S is five or larger; the weighting γ which gives the most accurate slopes or damping factors is:

$$\gamma_o = \frac{1}{2} - \frac{1}{12S^2\Delta\tau} = \frac{6r-1}{12r} \qquad \text{(IV-308)}$$

Second, for methods and problems where the approximate trigonometric roots are not equal to the continuous roots the best selection of γ is not given by γ_o but additional research must be carried out to determine quantitatively how much different γ must be from γ_o for a given H and S and a given approximate method. Third, for any methods and problems selecting γ significantly larger than γ_o causes the transient solution to decay too rapidly; selecting γ significantly smaller than γ_o causes the transient solution to decay too slowly. Since γ_o is always between zero and one-half, inclusive, the slopes of the transient terms for the backward difference implicit calculation with a γ of 1 always are the least accurate.

(5) The error contributed to the intercepts by the elements of the eigenvector matrix for the approximate method is proportional to the difference between the trigonometric roots for the approximate solution and those for the continuous solution. Thus, for methods and problems where these roots are equal, the error in the intercept is caused only by the error in the initial vector components.

(6) Method C, with no points located on the boundary, has the most accurate trigonometric roots, eigenvectors, and initial vectors of any of the approximations discussed here. Consequently, in general, for a given selection of differencing parameters, the approximate

solution for method C is expected to be the most accurate solution of any discussed here, and should be used when accuracy is of primary importance. Three other important observations about method C are:

a) The accuracy of method C is a direct result of using mesh $\Delta\xi/2$ which does not locate points on the boundary but locates the points a distance $\Delta\xi/2$ away from the boundaries. This leads to a difference formulation that is closer both physically and mathematically to the continuous problem than for any other method. This is because only for methods based on mesh $\Delta\xi/2$ do the following three properties occur simultaneously, (i) the boundary equations for this mesh are a direct discretization of the continuous boundary equations; (ii) no heat capacity is associated with the point on the boundary which is consistent with the continuous boundary equation; (iii) no heat capacity of the solid is neglected. Further, of the possible methods based on mesh $\Delta\xi/2$ or mesh $\Delta\xi$, method C is the only method that has a symmetric Y/A matrix which must have orthogonal eigenvectors; consequently, method C has an orthogonal relationship that is closer mathematically to the orthogonal relationship of the continuous solution, and method C is expected to have the most accurate initial vectors for any initial temperature distribution.

b) The bound for $|q_{min}|$ for method C is independent of the type of linear boundary equations or size of the heat-transfer

coefficients at both surfaces; thus, only one relationship between the bound and the differencing parameters r and γ is necessary. However, for problem I with a large heat-transfer coefficient or problem II, method C gives a large weighting to the damping factors which can be negative. This large weighting causes the very oscillatory behavior in Figure IV-1 for a graphical method C for problem II, and it is the reason method C is not suitable, in general, for graphical methods. Indeed, to obtain the high potential accuracy of method C, the time differencing parameters, $\Delta\tau$ and γ, must be selected so that any oscillatory transient terms are negligible.

c) Although method C is expected to give the most accurate approximations in general, for a specific problem with a specific set of differencing parameters, another method might give a more accurate solution because of compensating errors in the intercept and in the slope in the transient terms for the other method. Unless the direction and an estimate of the size of the error in the intercepts is known (the direction and size of the error in the slope can be found from a prior conclusion) this compensation cannot be predicted.

(7) The most accurate approximate solutions for problems in one, two, or three dimensions in any coordinate system are probably obtained using regular meshes with the points located a distance of $\Delta\xi/2$ away from the boundary. This conclusion is based on the more

accurate solutions obtained for method C, and on the fact that to obtain the physical and mathematical advantages mentioned in (6a) above, the adjacent points must be located away from the boundary. For irregular geometries and/or irregular meshes, locating the points away from the surfaces is probably also advantageous.

(8) Of the graphical methods, A is usually the most accurate. Using this method if no more than 8 points are required, an approximate solution of sufficient accuracy can be obtained by a graphical construction for times when no more than two of the terms of the transient solution are significant. If more than 8 points are required, the solution should be calculated numerically, but, because of the simplicity of the difference equation, this can be done readily on a desk calculator; thus, graphical method A is suitable for use in problems where a moderately accurate solution at intermediate and long times is desired, and where the solution is desired quickly and cheaply. Several additional comments should be made about graphical methods:

a) Graphical method A, along with the other graphical methods, requires about twice the number of points to give as accurate an intercept as does method C for problem II because of the direct or indirect effect of the large negative root. If this is a characteristic of graphical methods for other problems, they are not competitive in the number of calculations required to obtain a given accuracy at a

constant time, and a generalized method C should be used for accurate solutions.

b) The reason that graphical method A is the most accurate of the graphical methods is that several errors tend to compensate. As the trigonometric roots α_j are symmetric about $\pi/2$ for all the problems considered here, the damping factors always occur in pairs with equal magnitude but opposite sign. Thus, the slopes for the sum of the two paired damping factor terms give straight lines on semi-logarithmic paper only if the points at alternate time intervals are used. Further, this method appears to have the tendency that the amplitude for the negative of the pair of damping factors is small and about the size of the error in the initial vector component-eigenvector product for the positive damping factor, and the intercept is more accurate for (n+m) odd in these problems. Additional compensation is obtained when the intercept error is positive as it is in these problems using graphical method A, because the slope is too steep for graphical solutions and the transient terms in the approximate solution cross the corresponding continuous terms on a semi-logarithmic graph. Also for problem I, the error in the approximate slope for graphical method A is smaller than usual for γ of zero and r of $\frac{1}{2}$, because the difference in trigonometric roots compensates part of the usual error.

c) The forward average improves the accuracy of graphical method A by reducing the size of the oscillatory component and by

reducing the error in the initial vector components. However, this latter improvement is dependent upon the fact that the error in the initial vector for graphical method A in all problems studied here is always positive, and under these circumstances, the forward average decreases this error.

(9) The technique of averaging the approximate solution at the beginning and end of a time interval, and applying the average at either end or some time during the time interval, was shown to be equivalent under certain conditions to interpolating as defined by Longwell (9) for graphical solutions, and also equivalent to replacing a step change in fluid temperature with a one-half step initially, and the full step change occurring at the end of the first increment. Averaging reduces the amplitude of any large negative damping factors, and if the average is applied at the center of the time increment, the initial vector component of the slower-decaying positive damping factors is not changed. If the average is applied at the beginning or end of the time increment, the error in their initial vectors can either be reduced or increased, depending upon the direction of that error before averaging; therefore, averaging other than the central average probably should not be used unless the direction of the error in the initial vector is known. The main application of averaging appears to be to graphical methods.

(10) The quantitative relationships derived for the errors as a function of both the problem and the differencing parameters, have

given a procedure that allows the selection of the differencing parameters to obtain a certain accuracy, and at the same time, tends to minimize the amount of calculations. Although this procedure requires estimates of certain quantities in the continuous solution, and is fairly complicated, the following observations and conclusions can be made about the selection of differencing parameters.

a) The number of points S or S_G to be used is fixed by the allowable error in the intercept.

b) Graphical solutions, if they are to be calculated numerically, must not require many more points than the general explicit or implicit method.

c) The only generalized explicit calculations that need to be considered are those where $r_{Ex} > 2(S/S_G)^3$, and in most cases the oscillatory behavior will limit the size of r_{Ex}. The explicit calculation of r of 1/6 is the most accurate calculation possible for a fixed value of S.

d) The oscillatory behavior need never limit the r_{Im} for an implicit calculation, as increasing γ can always remove this restriction. However, a limiting combination of r_{Im} and γ exists for a fixed S that satisfies the restrictions governing both the oscillatory behavior and the accuracy of the damping factors. For an implicit calculation to be considered, it must have an r_{Im} that is greater than 0.58, or the explicit calculation with r of 1/6 is as accurate and requires the same

number of multiplications. In general, the r_{Im} must be greater than $3.5\ r_{Ex}$ or $3.5\ (S/S_G)^3$.

(G) SUGGESTIONS FOR FUTURE WORK

In order to extend, generalize, and check the conclusions reached in this discussion of the approximations to the one-dimensional diffusion equation, several further numerical and analytic studies are suggested.

First, for problem I, a study of the trigonometric roots for method C and graphical method A should be made to develop a relationship either in the form of a series or a graphical correlation for the difference $(\nu_j^2-\mu_j^2)$ as a function of S, H, and j for at least the first several j's. This type of relationship is necessary to allow use of equation IV-287 to estimate the error in the damping factors. Also the characteristic roots for method C and problem I should be further studied to develop a modified relationship corresponding to equation IV-308 for γ_o which allows one to find the optimum γ_o for problems with a finite heat-transfer coefficient. Again, this relationship could be a graphical correlation of H, r, and γ_o. Along these lines, the numerical solution of generalized method A with r of 1/6 could be stepped out to find if significant improvement is possible when the μ_j and ν_j are brought closer together.

Second, a problem where the heat-transfer coefficients at both boundaries are between zero and infinity should be studied both analytically and numerically to be sure that the methods which are accurate for problem I are also more accurate for this more general

problem. The analytic study could follow the work here and the numerical study should follow that suggested for problem I.

Third, the approximate solutions for problems where the boundary forcing functions change with time should be studied. This should indicate if any advantage in accuracy accrues to any of the methods in the particular solution. Also, the effect of the particular solution on the initial vector would be shown. Probably the easiest functions and problems to study first would be a sinusoid surface temperature for problem II and a sinusoid flux for problem III. These results might be generalized to problem I without actually deriving its solution. These several solutions might be sufficient to indicate that the comments made about the eigenvectors and damping factors for the several methods also apply to the particular solution as indicated by equation IV-213.

Fourth, the solutions for the one-dimensional problem found here can be extended to two-dimensional problems (two Cartesian space coordinates). This can be done because by subtracting the continuous solutions from 1 for problems I, II, and V, the solutions to the problem with a zero fluid temperature and initial temperature of 1 are obtained; for these problems, the product solution theorem as described in Carslaw and Jeager (1), section 1.15, can be used for the two-dimensional problem to obtain the complete solution of initial vector, eigenvector matrix, and eigenvalue exponentials directly. For an

approximate solution, the eigenvalues and the eigenvectors for the Y/A matrices for square meshes can be found by using matrix theorems proved by Rutherford (23). These theorems are analogous to the product solution theorem for the continuous solution and, by multiplying the two solutions together, the complete solution including the initial vector for the two-dimensional problem results.

Fifth, the approximate solutions for problem V, the semi-infinite solution, should be studied in more detail with the possibility of understanding the errors in the approximate solutions for short times and infinite solids. This could be approached by seeing if the approximate solutions can be expressed in terms of a "binomial error function" and then comparing this to the normal error function term in the continuous solution, as mentioned previously.

Sixth, as the initial vector has not been studied previous to this work several useful additional studies can be made. Of these, probably the most important is further investigation, both analytic and numerical, of the conclusion that the most accurate initial vectors are obtained when the A_i in the orthogonality relationship for g_j correspond to an integration of the weighting in the orthogonality relationship for a_j. This could be done by deriving the approximate solutions for meshes corresponding to mesh $\Delta\xi/2$ for cylindrical and spherical coordinates. Several useful studies involving method C based on mesh $\Delta\xi/2$ also can be made. The z-transform solutions for other problems using this

method should be found to try to improve the relationship used to estimate the number of points S for an approximation of a certain accuracy. In particular, the number 1/12 appears as the coefficient of $1/S^2$ in the expansions for the errors in method C; and possibly a study of the orthogonality relationship and the discretization error near the surface might give a relationship that would allow us to use this 1/12 as a proportionality constant in the intercept error bound in equation IV-286.

In this study, the finite sum in the orthogonality relationship for the approximate solution might be treated as an estimate of the integral for the continuous solution. Also, in studying the space discretization at the surface, care should be taken since a discontinuity occurs at the boundary. These additional studies of the initial vectors are also warranted, because if a simple way could be found of determining the sign alone of (g_j-a_j) considerable improvement in accuracy could be obtained by cancelling errors. For example, if (g_j-a_j) is known to be positive, then the time differencing parameters can be selected so that the damping factors are small, $\gamma < \gamma_o$, and the approximate transient term crosses a continuous transient term on a semi-logarithmic graph. Also, a forward average could then be used, which would reduce the error in the initial vector. If (g_j-a_j) is negative, then the γ and $\Delta\tau$ are selected so that the damping factors are too large ($\gamma > \gamma_o$) or the backward average is used. If the coefficient of

$1/S^2$ in the expansion for (g_j-a_j) were also known, spectacular improvement in accuracy could be made by selecting r and the averaging parameter k so that this term would be zero and (g_j-a_j) would be proportional to $1/S^4$. These same studies should also be carried out for graphical method A. For this graphical method, the effective initial vectors should be derived and studied. These vectors would then show not only the size and sign of (g_j-a_j) for the slower decaying damping factors, but would also indicate the relative size of the g_j for the oscillatory components. Hopefully, these expansions could show that (g_j-a_j) is always positive, so that the forward average could always be associated with graphical method A, and that the g_j for the oscillatory components are always about the same size as (g_j-a_j) so that for either odd or even values of n, compensation occurs. Further, a simple way should be developed to find at which values of n odd or even this compensation occurs; again, hopefully, this would be at (n+m) odd as for the problems already studied.

Table IV-1

GENERALIZED METHOD G

Mesh $\Delta\xi$, Equation IV-10, $\beta = \phi = \gamma$

Boundary Equation $\quad 0 \leq h < \infty$

$$t_{0,n+1}\left[1+2r\gamma(1+H_0/S)\right] - 2r\gamma t_{1,n+1} = \left[1-2r(1-\gamma)(1+H_0/S)\right]t_{0,n} + 2r(1-\gamma)t_{1,n} + (2rH_0/S)\left[\gamma t_{f_0,n+1} + (1-\gamma)t_{f_0,n}\right]$$

$$n \geq 0$$

Interior Points

$$-r\gamma t_{m-1,n+1} + (1+2r\gamma)t_{m,n+1} - r\gamma t_{m+1,n+1} = r(1-\gamma)t_{m+1,n} + \left[1+2r(1-\gamma)\right]t_{m,n} + r(1-\gamma)t_{m-1,n}$$

$$m = 1, 2, \ldots, S-1$$

$$n \geq 0$$

Matrices

$$(\Delta\xi)^2 \frac{Y}{A} = \begin{bmatrix} -2(1+H_0/S) & 2 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 2 & -2(1+H_S/S) \end{bmatrix}$$

$$(\Delta\xi)^2 \frac{Y_B}{A} = \begin{bmatrix} 2H_0/S & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 2H_S/S \end{bmatrix}$$

Table IV-1 (Cont.)

GENERALIZED METHOD G

Initial Condition $\quad 0 \leq h < \infty$

$$t_{m,0} = T(m\Delta\xi,\ 0+) \qquad m = 0,\ 1,\ \ldots,\ S$$

Fluid Temperature

$$t_{f,n} = T_f(n\Delta\tau) \qquad n \geq 0$$

Stability $\quad 0 \leq h < \infty$

$$r(1-2\gamma)M \leq 2 \qquad M = \min\left(\left\| \frac{Y}{A} \right\|_I \text{ or } \left\| \frac{Y}{A} \right\|_{II} \right)$$

$$\left\| \frac{Y}{A} \right\|_I = 2(2 + H_{MAX}/S)$$

$$\left\| \frac{Y}{A} \right\|_{II} = MAX\left\{ 5 \text{ or } (3 + 2H_{MAX}/S) \right\}$$

Infinite h $\quad H = \infty$

$$t_{f,n} = t_{0,n} \qquad n \geq 0$$

Delete row and column containing infinite H in $(\Delta\xi)^2 Y/A$ and row containing infinite H in $(\Delta\xi)^2 Y_B/A$; in $(\Delta\xi)^2 Y_B/A$ put a one in place of element located at the intersection of column containing infinite H and row adjacent to deleted row.

Stability: $\quad r(1-2\gamma) \leq \frac{1}{2}$

Initial Condition: Fits boundary condition at n = 0 for boundaries where $H = \infty$

Table IV-2

GENERALIZED METHOD A

Mesh $\Delta\xi$, Equation IV-10, $\beta = 1$, $\phi = 0$

Boundary Equation $\quad 0 \leq h < \infty \quad$ Explicit only

$$t_{0,n+1} = \frac{(1-2r)t_{0,n} + 2rt_{1,n} + (2rH_0/S)t_{f_0,n+1}}{1 + 2rH_0/S}$$

$$n \geq 0$$

Interior Points

$$t_{m,n+1} = rt_{m-1,n} + (1-2r)t_{m,n} + rt_{m+1,0}$$

$$m = 1, 2, \ldots, S-1$$

$$n \geq 0$$

Matrices (Note Y/A matrix is a function of r or $\Delta\tau$.)

$$(\Delta\xi)^2 \frac{Y}{A} = \begin{bmatrix} -\frac{2(1+H_0/S)}{1+2rH_0/S} & \frac{2}{1+2rH_0/S} & 0 \cdots & 0 & 0 \\ 1 & -2 & 1 \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 \cdots & \frac{2}{1+2rH_S/S} & -\frac{2(1+H_S/S)}{1+2rH_S/S} \end{bmatrix}$$

$$(\Delta\xi)^2 \frac{Y_B}{A} = \begin{bmatrix} \frac{2H_0/S}{1+2rH_0/S} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \frac{2H_S/S}{1+2rH_S/S} \end{bmatrix}$$

Table IV-2 (Cont.)

GENERALIZED METHOD A

Initial Distribution $0 \leq h < \infty$

$t_{m,0} = T(m\Delta\xi, 0-)$ $\quad m = 0, 1, 2, \ldots, S$

Fluid Temperature $0 \leq h < \infty$

$t_{f,n} = T_f(n\Delta\tau)$ $\quad n \geq 0$

Stability

$r \leq \frac{1}{2}$ $\quad$ all h $\quad 0 \leq h \leq \infty$

Infinite h $\quad H = \infty$

$t_{0,n+1} = t_{f_0,n+1}$ $\quad n \geq 0$

In $(\Delta\xi)^2 Y/A$ in row containing infinite H set diagonal element equal to 1/r ; other elements in row to zero. In $(\Delta\xi)^2 Y_B/A$ put 1/r in position of infinite H.

Initial Condition:

$t_{m,0} = T(m\Delta\xi, 0-)$ $\quad m = 0, 1, 2, \ldots, S$

$T(m\Delta\xi, 0-)$ = Initial condition before fluid temperature changes

Note: As $t_{0,0}$ is not required to equal $t_{f_0,0}$ initially, the initial distribution need not necessarily give the surface temperature at 0 or S initially.

Special Case:

If one H is infinite other H zero and $t_{f,n}$ constant, then define implicit method A as implicit method G delayed one $\Delta\tau$:

$$t_{m,n\ A} = t_{m,n-1\ G}$$

Table IV-3

GENERALIZED METHOD C

Mesh $\Delta\xi/2$, Equation IV-29, $\beta = \phi = \gamma$

Boundary Equation $\quad 0 \leq h \leq \infty$

$$t_{\frac{1}{2},n+1}\left[1+ r\left(\frac{1+3H_0/2S}{1+H_0/2S}\right)\right] - r\gamma t_{\frac{3}{2},n+1} = t_{\frac{1}{2},n}\left[1-r(1-\gamma)\left(\frac{1+3H_0/2S}{1+H_0/2S}\right)\right]$$

$$+ r(1-\gamma)t_{\frac{3}{2},n} + \left[\frac{rH_0/S}{1+H_0/2S}\right]\left[(1-\gamma)t_{f_0,n} + t_{f_0,n+1}\right]$$

$$n \geq 0$$

Interior Points

$$-r\gamma t_{m-1,n+1} + (1-2r\gamma)t_{m,n+1} - r\gamma t_{m+1,n+1} = r(1-\gamma)t_{m-1,n}$$

$$+ \left[1-2r(1-\gamma)t_{m,n}\right] + r(1-\gamma)t_{m+1,n}$$

$$m = \frac{3}{2}, \frac{5}{2}, \ldots, S-\frac{3}{2}$$

$$n \geq 0$$

Matrices

$$(\Delta\xi)^2 \frac{Y}{A} = \begin{bmatrix} -\dfrac{1+3H_0/2S}{1+H_0/2S} & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\dfrac{1+3H_S/2S}{1+H_S/2S} \end{bmatrix}$$

$$(\Delta\xi)^2 \frac{Y_B}{A} = \begin{bmatrix} \dfrac{H_0/S}{1+H_0/2S} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \dfrac{H_S/S}{1+H_S/2S} \end{bmatrix}$$

Table IV-3 (Cont.)

GENERALIZED METHOD C

Initial Condition

$$t_{m,0} = T(m\Delta\xi,\ 0) \qquad m = \frac{1}{2},\ \frac{3}{2},\ \ldots,\ S-\frac{1}{2}$$

Fluid Temperature

$$t_{f,n}\ T_f(n\Delta\tau)$$

Stability

$$r(1-2\gamma) \leq \frac{1}{2} \qquad 0 \leq h \leq \infty$$

Infinite h

Boundary Equation:

$$t_{\frac{1}{2},n+1}(1+3r\gamma) - r\gamma t_{\frac{3}{2},n+1} = t_{\frac{1}{2},n}\left[1-3r(1-\gamma)\right] + r(1-\gamma)t_{\frac{3}{2},n} + 2r\left[(1-\gamma)t_{f_0,n} + t_{f_0,n+1}\right]$$

Table IV-4

GENERALIZED METHOD F

Mesh $\Delta\xi/2$, Equation IV-29, $\beta = \frac{1}{2}$, $\phi = 0$, Explicit $\gamma = 0$

Boundary Equation $0 \leq h \leq \infty$

$$t_{\frac{1}{2},n+1} = \frac{t_{\frac{1}{2},n}\left[1-r\left(\frac{1+H_0/S}{1+H_0/2S}\right)\right] + rt_{\frac{3}{2},n} + \left[\frac{rH_0/2S}{1+H_0/2S}\right]\left[t_{f_0,n} + t_{f_0,n+1}\right]}{1 + r\left(\frac{H_0/2S}{1+H_0/2S}\right)}$$

$$n \geq 0$$

(If H = 0, same equation as explicit method C)

Interior Points

$$t_{m,n+1} = rt_{m-1,n} + (1-2r)t_{m,n} + rt_{m+1,n}$$

$$m = \frac{3}{2}, \frac{5}{2}, \ldots, S-\frac{3}{2}$$

$$n \geq 0$$

Matrices (Note: Equivalent Y/A is a function of r)

$$(\Delta\xi)^2 \frac{Y}{A} = \begin{bmatrix} -\left\{\frac{1+3H_0/2S}{1+(1+r)H_0/2S}\right\} & \left\{\frac{1+H_0/2S}{1+(1+r)H_0/2S}\right\} & 0 \ldots & & \\ 1 & -2 & 1 \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 \ldots & \left\{\frac{1+H_S/2S}{1+(1+r)H_S/2S}\right\} & -\left\{\frac{1+3H_S/2S}{1+(1+r)H_S/2S}\right\} \end{bmatrix}$$

$$(\Delta\xi)^2 \frac{Y_B}{A} = \begin{bmatrix} \frac{H_0/2S}{1+(1+r)H_0/2S} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \frac{H_S/2S}{1+(1+r)H_S/2S} \end{bmatrix}$$

Table IV-4 (Cont.)

GENERALIZED METHOD F

Initial Condition

$$t_{m,0} = T(m\Delta\xi,\ 0) \qquad m = \frac{1}{2},\ \frac{3}{2},\ \ldots,\ S-\frac{1}{2}$$

Fluid Temperature

$$t_{f,n} = T_f(n\Delta\tau) \qquad n \geq 0$$

Stability

$$r \leq \frac{1}{2}$$

Infinite h $\qquad H = \infty$

$$t_{\frac{1}{2}} = \frac{(1-2r)t_{\frac{1}{2},n} + rt_{\frac{3}{2},n} + r(t_{f_0,n} + t_{f_0,n+1})}{1 + r}$$

Table IV-5

GRAPHICAL METHODS

General Definition

Generalized methods with

$$\gamma = 0$$

$$r = \frac{1}{2}$$

Graphical Method G

Only defined for:

H's infinite or zero

(Unstable other H's and simple construction not available)

If both H's are zero:

$q_{min} = -1$

Averaged graphical method G, $q_{min} = -1$ does not appear

Graphical Method A

Defined all H

If both H's are zero, $q_{min} = -1$

Averaged graphical method A, $q_{min} = -1$ does not appear

Graphical Method C

Not recommended for $H/S > 2$ (stable all H)

Graphical Method F

Usually used only for both H's = oo, or one 0, the other oo

Construction may be difficult or impossible other H's

Table IV-6

AVERAGED METHODS

General Definition

$$t_{n+k\ Ave} = \frac{1}{2}(t_n + t_{n+1}) \qquad \text{Applies at time } (n+k)\Delta\tau \qquad n \geq 0$$

Forward Average, k=0:

$$t_{n\ Ave} = \frac{1}{2}(t_n + t_{n+1}) \qquad \text{at time } n\Delta\tau \qquad n \geq 0$$

(Changes initial condition)

Central Average, k=1/2:

$$t_{n+\frac{1}{2}} = \frac{1}{2}(t_n + t_{n+1}) \qquad \text{at time } (n+\tfrac{1}{2})\Delta\tau \qquad n \geq 0$$

Backward Average, k=1:

either

$$t_{n+1} = \frac{1}{2}(t_n + t_{n+1}) \qquad \text{at time } (n+1)\Delta\tau \qquad n \geq 0$$

or

$$t_n = \frac{1}{2}(t_{n-1} + t_n) \qquad \text{at time } n\Delta\tau \qquad n \geq 1$$

In definitions t_n's on right side are found by stepping out the solution for a method using unaveraged temperature vector to advance to (n+1); i.e. $t_{n+k\ Ave}$ is not used to advance. However, if calculation matrices are constant (not a function of n or t_n) using average $t_{n+k\ Ave}$ and an average boundary temperature vector to step to the next time increment gives a solution identical to an averaged solution (see equation IV-44).

Definition Averaged Methods

Averaged Method G -- Backward Average

Averaged Method A -- Forward Average

Averaged Method C -- Backward Average

Averaged Method F -- Forward Average

Table IV-7

z-TRANSFORM OPERATION PAIRS

	Difference Function f_n	z-Transform $\bar{f}(z)$
1.	f_{n+1}	$z\bar{f}(z) - zf_0$
2.	f_{n+2}	$z^2 \bar{f}(z) - z^2 f_0 - zf_1$
3.	f_{n+3}	$z^3 \bar{f}(z) - z^3 f_0 - z^2 f_1 - zf_2$
4.	f_{n+a} $\quad a \geq 0$	$z^a \bar{f}(z) - z^a f_0 - \ldots - zf_{a-1}$
5.	f_{n-b} $\quad b \geq 0$ where $f_n = 0 \quad n < 0$	$z^{-b} \bar{f}(z)$
6.	$\Delta f_n = f_{n+1} - f_n$	$(z - 1)\bar{f}(z) - zf_0$
7.	nf_n	$- z \frac{d}{dz} \bar{f}(z)$
8.	$\frac{1}{n} f_n$	$-\int \frac{\bar{f}(z)}{z} dz$
9.	$\frac{1}{n + a} f_n$	$- z^a \int \frac{\bar{f}(z)}{z^{a+1}} dz$
10.	$a^n f_n$	$\bar{f}\left(\frac{z}{a}\right)$
11.	$\sum_{p=0}^{n} f_p$	$\frac{z}{z - 1} \bar{f}(z)$
12.	$\sum_{p=0}^{n} f_p \, u_{n-p}$	$\bar{f}(z) \, \bar{u}(z)$

Table IV-7 (Cont.)

z-TRANSFORM OPERATION PAIRS

	Difference Function f_n	z-Transform $\bar{f}(z)$
13.	$\sum_{j=1}^{\infty} \rho_{j,n}$ where $\rho_{j,n}$ = residues of $z^{n-1}\bar{f}(z)$ For simple poles: $\rho_{j,n} = \lim_{z \to q_j} (z-q_j) f(z) z^{n-1} = \lim_{z \to q_j} \left[\frac{N(z)}{\frac{d}{dz} D(z)} \right]$	$\frac{N(z)}{D(z)z^{n-1}}$ (No Branch Cuts)
14.	Initial Value Theorem	$f_0 = \lim_{z \to \infty} \bar{f}(z)$
15.	Final Value Theorem	$f_\infty = \lim_{z \to 1} (z-1)\, \bar{f}(z)$ (Assumes True Steady State Exists)

Table Modified After Aseltine (25), p. 260

Table IV-8

z-TRANSFORM FUNCTION PAIRS

	Difference Function f_n	z-Transform $\bar{f}(z)$
1.	A = constant	$\frac{z}{z-1} A$
2.	n	$\frac{z}{(z-1)^2}$
3.	n^2	$\frac{z(z+1)}{(z-1)^3}$
4.	n^3	$\frac{z(z^2+4z+1)}{(z-1)^4}$
5.	q^n	$\frac{z}{z-q}$
6.	nq^n	$\frac{qz}{(z-q)^2}$
7.	$\sin \omega_0 n$	$\frac{z \sin \omega_0}{z^2 - 2z\cos \omega_0 + 1}$
8.	$\cos \omega_0 n$	$\frac{z(z - \cos \omega_0)}{z^2 - 2z\cos \omega_0 + 1}$
9.	$\sinh \omega_0 n$	$\frac{z \sinh \omega_0}{z^2 - 2z\cosh \omega_0 + 1}$
10.	$\cosh \omega_0 n$	$\frac{z(z - \cosh \omega_0)}{z^2 - 2z\cosh \omega_0 + 1}$

Table IV-8 (Cont.)

z-TRANSFORM FUNCTION PAIRS

Difference Function f_n	z-Transform $\bar{f}(z)$
11. $e^{-\alpha n} \sin \omega_0 n$	$\dfrac{ze^{-\alpha} \sin \omega_0}{z^2 - 2ze^{-\alpha} \cos \omega_0 + e^{-2\alpha}}$
12. $e^{-\alpha n} \cos \omega_0 n$	$\dfrac{z(z - e^{-\alpha} \cos \omega_0)}{z^2 - 2ze^{-\alpha} \cos \omega_0 + e^{-2\alpha}}$
13. $\dfrac{1}{n!}$	$e^{0.5}$
14. $\dfrac{1}{(2n)!}$	$\cosh(z^{-0.5})$
15. $\delta_{np} = \begin{matrix} 0 & n \neq p \\ 1 & n = p \end{matrix}$	z^{-p}

Table Modified After Aseltine (25), p. 259.

Table IV-9

PROBLEMS TO BE APPROXIMATED

Partial Differential Equation

$$\frac{\partial T}{\partial \tau} = \frac{\partial^2 T}{(\partial \xi)^2} \qquad 0 \leq \xi \leq 1$$

Problem I - Finite h and Zero h

Initial	$T(\xi,0) = 0$	$0 \leq \xi \leq 1$
Left Boundary	$-q\left[1 - T(0,\tau)\right] = \frac{\partial T}{\partial \xi}(0,\tau)$	$\tau \geq 0$
Right Boundary	$\frac{\partial T}{\partial \xi}(1,\tau) = 0$	$\tau \geq 0$

Problem II - Infinite h and Zero h

Initial	$T(\xi,0) = 0$	$0 < \xi \leq 1$
Left Boundary	$T(0,\tau) = 1$	$\tau \geq 0$
Right Boundary	$\frac{\partial T(1,\tau)}{\partial \xi} = 0$	$\tau \geq 0$

Problem III - Both h's Zero

Initial	$T(\xi,0) =$	$0 \leq \xi \leq 1$
Left Boundary	$\frac{\partial T}{\partial \xi}(0,\tau) = 0$	$\tau \geq 0$
Right Boundary	$\frac{\partial T}{\partial \xi}(1,\tau) = 0$	$\tau \geq 0$

Problem IV - Both h's Infinite

Initial	$T(\xi,0) = 0$	$0 \leq \xi < 1$
Left Boundary	$T(0,\tau) = 0$	$\tau \geq 0$
Right Boundary	$T(1,\tau) = 1$	$\tau \geq 0$

Problem V - Semi-Infinite Solid

Initial	$T(\xi,0) = 0$	$0 \leq \xi \leq \infty$
Left Boundary	$T(0,\tau) = 1$	$\tau \geq 0$
Right Boundary	$\frac{\partial T}{\partial \xi}(\infty,\tau) = 0$	$\tau \geq 0$

Table IV-10

CONTINUOUS, ANALOG, AND DIFFERENCE SOLUTIONS IN DIFFERENCE FORM

One-Dimensional Diffusion Equation in Cartesian Co-ordinates

Continuous

$$T(m\Delta\xi,\ n\Delta\tau) = T_P(m\Delta\xi,\ n\Delta\tau) - \sum_{j=1}^{\infty} a_j\, b_j(m\Delta\xi)\left[e^{-\psi_j^2 r}\right]^n$$

where ψ_j defined by $F_C(\psi_j) = 0$ Characteristic Equation

Approximate

$$t_{m,n} = t_{P\ m,n} - \sum_{j=1}^{S} g_j\, c_{mj}\, q_j^{\,n}$$

for difference solution $q_j^{\,n} = \left[\dfrac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r(1-\gamma)(1-\cos\alpha_j)}\right]^n$

for analog solution (only for methods G and C) replace

$q_j^{\,n}$ with $e^{-2S^2(1-\cos\alpha_j)n\tau}$

where α_j defined by $F_D(\alpha_j) = 0$ Characteristic Equation

Notes

All Equations and Methods	$n = 0, 1, \ldots, \infty$	
Continuous Solution either	$m = 0, 1, 2, \ldots, S$	Mesh $\Delta\xi$
or	$m = \frac{1}{2}, \frac{3}{2}, \ldots, S-\frac{1}{2}$	Mesh $\Delta\xi/2$
Approximate Method G or A	$m = 0, 1, 2, \ldots, S$	
Approximate Method C or F	$m = \frac{1}{2}, \frac{3}{2}, \ldots, S-\frac{1}{2}$	

For approximate methods summation on j is to number of variable temperature points

Table IV-10 (Cont.)

CONTINUOUS, ANALOG, AND DIFFERENCE SOLUTIONS IN DIFFERENCE FORM

Corresponding Quantities

	Continuous	Approximate
Particular Solution - Steady State or Quasi-Steady State	$T_P(m\Delta\xi,\ n\Delta\mathcal{T})$	$t_{P\ m,n}$
Fourier Coefficient - Component of Initial Vector Weighting Given to j^{th} Eigenvector-Eigenvalue Product	$a_j = a_j(\psi_j)$	$g_j = g_j(\alpha_j)$
Component in Row for m^{th} Temperature Point and j^{th} Column of Eigenvector Matrix	$b_j(m\Delta\xi) = b_j(m\Delta\xi, \psi_j)$	$c_{mj} = c_{mj}(\alpha_j)$
Damping Factors	$e^{-\psi_j^2 r}$ or $e^{-\nu_j^2 \Delta\mathcal{T}}$	Difference: $\frac{1-2r(1-\gamma)(1-\cos\alpha_j)}{1+2r\gamma(1-\cos\alpha_j)}$ Analog: $e^{-2r(1-\cos\alpha_j)}$
Characteristic Equation	$F_C(\psi_j) = 0$	$F_D(\alpha_j) = 0$
Eigenvalue Parameter or Trigonometric Root	$\psi_j = \frac{\nu_j}{S}$	$\alpha_j = \frac{\mu_j}{S}$

Table IV-11

PARTICULAR OR STEADY-STATE SOLUTIONS

Problem I

$$T_P(m\Delta\xi) = t_{P\ m,n} = 1$$

Approximate Methods G, A, and C

Problem II

$$T_P(m\Delta\xi) = t_{P\ m,n} = 1$$

Approximate Methods G, A, and C, and Graphical Method F

Problem III

$$T_P(m\Delta\xi,\ n\Delta\tau) = t_{P\ m,n} = \frac{1}{2}$$

Approximate Methods G, A, and C

Problem IV

$$T_P(m\Delta\xi,\ n\Delta\tau) = t_{P\ m,n} = m\Delta\xi = \frac{m}{S}$$

Approximate Method G

Problem V

$$T_P(m\Delta\xi,\ n\Delta\tau) = t_{P\ m,n} = 1$$

Approximate Method G

($t_{P\ m,n}$ and $T_{P\ m,n}$ steady state for $m < \infty$; steady state for $m = \infty$ is zero as found by adding integral, Table IV-29, to particular solution.)

Table IV-12

CHARACTERISTIC EQUATIONS - PROBLEM I

One h Finite; One h Zero

(Applies in Limits $h \to \infty$, Problem II; and $h \to 0$, Problem III)

	Characteristic Equations	Number of Roots	Notes on Roots
Continuous *			
	$(S\psi_j)(\tan S\psi_j) = H$	∞	ψ_j real and positive $\quad 0 \leq \psi_j \leq \infty$ $j = 1, 2, \ldots, \infty$
Approximate Methods			All real α_j's fall within $0 \leq \alpha_j \leq \pi$
G**	$S(\sin\alpha_j)(\tan S\alpha_j) = H$	S+1	$\alpha_1, \alpha_2, \ldots, \alpha_S$ real α_{S+1} complex; $\alpha_{S+1} = \pi + \sqrt{-1}\,\mathcal{J}(\alpha_{S+1})$ $0 \leq \mathcal{J}(\alpha_{S+1}) < \infty$
A	$\dfrac{S(\sin\alpha_j)(\tan S\alpha_j)}{1 - 2r(1-\cos\alpha_j)} = H$	S+1	If $r < \frac{1}{4}$, α_j's as in method G If $r \geq \frac{1}{4}$, S+1 roots real
Graphical A, r=1/2			
	$S(\tan\alpha_j)(\tan S\alpha_j) = H$	S+1	All α_j's real
C**	$2S(\tan\frac{\alpha_j}{2})(\tan S\alpha_j) = H$	S	All α_j's real

Notes: 1. Only for h=∞ (Problem II) or h=0 (Problem III) are the S or (S+1) α_j's equal to first S or (S+1) ψ_j's for these approximate methods.

2. Characteristic equations depend only on H, not on how fluid temperature changes with time or the initial distribution.

*Continuous characteristic equation in difference form **Characteristic equation independent of γ and r

Table IV-13

CHARACTERISTIC EQUATIONS - PROBLEMS II, III, AND IV

Problem II $h_0 = \infty$ and $h_S = 0$ (Limiting Case Problem I, $H \longrightarrow \infty$)

Continuous $\cos S\psi_j = 0$ $\psi_j = \frac{(2j-1)\pi}{2S}$ $j = 1, \ldots, \infty$

Approximate

Methods G, A, and C

$\cos S\alpha_j = 0$ $\alpha_j = \frac{(2j-1)\pi}{2S}$ $j = 1, \ldots, S$

$\alpha_j = \psi_j$ for j's $1 \leq j \leq S$

Graphical Method F

$\left[\tan(S-\frac{1}{2})\alpha_j\right] \tan \alpha_j = 2$ $j = 1, \ldots, S$

Problem III Both h's Zero (Limiting Case Problem II, $H \longrightarrow 0$)

Continuous $\sin S\psi_j = 0$ $\psi_j = \frac{(j-1)\pi}{S}$ $j = 1, \ldots, \infty$

Approximate

Methods G, A, and C

$\sin S\alpha_j = 0$ $\alpha_j = \frac{(j-1)\pi}{S}$ $j = 1, \ldots, S, S+1$

Method C $j \neq S+1$

$\alpha_j = \psi_j$ for j less than S or S+1

(Note that for j=1, q_1=1, which, although an eigenvalue of the matrices, gives the steady state solution and is not included in transient part of solution.)

The α_{S+1} which equals π corresponds to q_{min} for methods G and A

Problem IV Both h's Infinite

Continuous $\sin S\psi_j = 0$ $\psi_j = \frac{j\pi}{S}$ $j = 1, 2, \ldots, \infty$

Approximate

Method G $\sin S\alpha_j = 0$ $\alpha_j = \frac{j\pi}{S}$ $j = 1, 2, \ldots, S-1$

Problem V No Characteristic Equation

(In limit of characteristic equation of problem II the fact that as $S \longrightarrow \infty$ $\Delta\xi \not\longrightarrow 0$ shows that ψ and α become continuous because any values of ψ and α satisfy the equation.)

Table IV-14

DAMPING FACTORS

DAMPING FACTORS - EIGENVALUES

Continuous $\qquad j = 1, 2, \ldots, \infty$

$$e^{-\nu_j^2 \Delta\tau} = 1 - \nu_j^2 \Delta\tau + \frac{\nu_j^4 (\Delta\tau)^2}{2} - \frac{\nu_j^6 (\Delta\tau)^3}{6} \cdots$$

Approximate All Methods $\qquad j = 1, 2, \ldots, S \text{ or } S+1$

$$q_j = \frac{1-2S^2\Delta\tau(1-\gamma)(1-\cos\frac{\mu_j}{S})}{1+2S^2\Delta\tau(\gamma)(1-\cos\frac{\mu_j}{S})} = 1 - \mu_j^2\Delta\tau + \frac{\mu_j^4(\Delta\tau)^2}{2} 2\gamma + \left[\frac{1}{6S^2\Delta\tau}\right]$$

$$- \frac{\mu_j^6(\Delta\tau)^3}{6}\left[6\gamma^2 + \frac{\gamma}{S^2\Delta\tau} + \frac{1}{60S^4(\Delta\tau)^2}\right] \cdots$$

DAMPING FACTORS TO n^{th} POWER AS A FUNCTION OF TIME

Continuous

$$e^{-\nu_j^2 n\Delta\tau} = 1 - \nu_j^2 n(\Delta\tau) + \frac{\nu_j^4 (n\Delta\tau)^2}{2} - \frac{\nu_j^6 (n\Delta\tau)^3}{6} \cdots$$

Approximate All Methods

Difference (Note $\tau = n\Delta\tau$):

$$q_j^n = 1 - \mu_j^2(n\Delta\tau) + \frac{\mu_j^4(n\Delta\tau)^2}{2}\left[1 + \frac{1}{6S^2\tau} - \frac{(1-2\gamma)}{n}\right]$$

$$- \frac{\mu_j^6(n\Delta\tau)^3}{6}\left[1 + \frac{1}{2S^2\tau} - \frac{3(1-2\gamma)}{n} + \frac{1}{60S^4\tau^2} + \frac{2(1-3\gamma+3\gamma^2)}{n^2} - \frac{(1-2\gamma)}{2S^2\tau n}\right] \cdots$$

Analog: $\Delta\tau \longrightarrow 0 \qquad n\Delta\tau \longrightarrow \tau \qquad n \longrightarrow \infty$

$$e^{-2S^2(1-\cos\frac{\mu_j}{S})n\Delta\tau} = 1 - \mu_j^2(n\Delta\tau) + \frac{\mu_j^4(n\Delta\tau)^2}{2}\left[1 + \frac{1}{6S^2\tau}\right]$$

$$- \frac{\mu_j^6(n\Delta\tau)^3}{6}\left[1 + \frac{1}{2S^2\tau} + \frac{1}{60S^4\tau^2}\right] \cdots$$

Table IV-15

DIFFERENCE IN LOGARITHMS OF DAMPING FACTORS

$$\frac{1}{\Delta\tau}\ln q_j = -\mu_j^2 + \mu_j^4\left[\frac{1}{12S^2} - (\tfrac{1}{2}-\gamma)\Delta\tau\right]$$

$$- \mu_j^6\left[(\tfrac{1}{3}-\gamma+\gamma^2)(\Delta\tau)^2 - \frac{(\frac{1}{2}-\gamma)\Delta\tau}{6S^2} + \frac{1}{360S^4}\right] \quad \cdots$$

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = \mu_j^2 - \nu_j^2 - \mu_j^4\left[\frac{1}{12S^2} - (\tfrac{1}{2}-\gamma)\Delta\tau\right]$$

$$+ \mu_j^6\left[(\tfrac{1}{3}-\gamma+\gamma^2)(\Delta\tau)^2 - \frac{(\frac{1}{2}-\gamma)\Delta\tau}{6S^2} + \frac{1}{360S^4}\right] \quad \cdots$$

Special Cases for γ

Explicit, $\gamma = 0$

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = \mu_j^2 - \nu_j^2 - \mu_j^4\left[\frac{1}{12S^2} - \frac{\Delta\tau}{2}\right]$$

$$+ \mu_j^6\left[\frac{1}{360S^4} - \frac{\Delta\tau}{12S^2} + \frac{(\Delta\tau)^2}{3}\right] \quad \cdots$$

Implicit, $\gamma = \frac{1}{2} - \frac{1}{12S^2\Delta\tau}$

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = \mu_j^2 - \nu_j^2 + \mu_j^6\left[\frac{(\Delta\tau)^2}{12} - \frac{1}{240S^2}\right] \quad \cdots$$

Implicit, $\gamma = \frac{1}{2}$

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = \mu_j^2 - \nu_j^2 - \mu_j^4\left[\frac{1}{12S^2}\right] + \mu_j^6\left[\frac{1}{360S^4} + \frac{(\Delta\tau)^2}{2}\right] \quad \cdots$$

Table IV-15 (Cont.)

DIFFERENCE IN LOGARITHMS OF DAMPING FACTORS

Special Cases for γ

Implicit, $\gamma = 1$

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = \mu_j^2 - \nu_j^2 - \mu_j^4\left[\frac{1}{12S^2} + \frac{\Delta\tau}{2}\right] + \mu_j^6\left[\frac{1}{360S^4} + \frac{\Delta\tau}{12S^2} + \frac{(\Delta\tau)^2}{3}\right] \cdots$$

Special Cases for γ and $\Delta\tau$ or r

Graphical Explicit, $\gamma = 0$, $r = \frac{1}{2}$

$$\ln \frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = \mu_j^2 - \nu_j^2 + \mu_j^4\left[\frac{1}{6S^2}\right] + \mu_j^6\left[\frac{2}{45S^4}\right] \cdots$$

Analog, $\Delta\tau \longrightarrow 0$

$$\ln \frac{e^{-\nu^2}}{e^{+\lambda}} = -\nu^2 - \lambda = -\nu^2 + 2S^2\left(1 - \cos\frac{\mu_j}{S}\right)$$

$$= \mu_j^2 - \nu_j^2 - \mu_j^4\left[\frac{1}{12S^2}\right] + \mu_j^6\left[\frac{1}{360S^2}\right] \cdots$$

Table IV-16

EIGENVECTOR MATRICES

Element in Row Corresponding to m^{th} Point,* j^{th} Column

Continuous Eigenvector Applies Either Mesh $\Delta\xi$ or $\Delta\xi/2$

For Mesh $\Delta\xi$	$m = 0, 1, 2, \ldots, S$
For Mesh $\Delta\xi/2$	$m = \frac{1}{2}, \frac{3}{2}, \ldots, S-\frac{1}{2}$

Approximate Eigenvector

Methods G and A, Mesh $\Delta\xi$	$m = 0, 1, 2, \ldots, S$
Method C, Mesh $\Delta\xi/2$	$m = \frac{1}{2}, \frac{3}{2}, \ldots, S-\frac{1}{2}$

Problem I Finite h and zero h $\alpha_j \neq \psi_j$

Continuous**

$$b_j(m\Delta\xi) = \cos(S-m)\psi_j \qquad j = 1, 2, \ldots, \infty$$

Approximate Methods G, A, and C

$$c_{mj} = \cos(S-m)\,\alpha_j$$

Methods G and A	$j = 1, 2, \ldots, S+1$
Method C	$j = 1, \ldots, S$

Problem II Infinite h and zero h $j = 1, \ldots, \infty$

Continuous**

$$b_j(m\Delta\xi) = \sin m\psi_j \qquad \psi_j = \frac{(2j-1)\pi}{2S} \qquad j = 1, \ldots, \infty$$

Methods G and A

$$c_{mj} = \sin m\psi_j \qquad \alpha_j = \psi_j \qquad j = 1, \ldots, S$$

Method C

$$c_{mj} = \sin m\psi_j \qquad \alpha_j = \psi_j \qquad j = 1, \ldots, S$$

Graphical Method F

$$c_{mj} = \cos(S-m)\,\alpha_j \qquad \alpha_j \neq \psi_j \qquad j = 1, \ldots, S$$

Table IV-16 (Cont.)

EIGENVECTOR MATRICES

Problem III*** Both h's zero

Continuous**

$$b_j(m\Delta\xi) = \cos m\psi_j \qquad \psi_j = \frac{(j-1)\pi}{S} \qquad j = 1, \ldots, \infty$$

Methods G and A

$$c_{mj} = \cos m\psi_j \qquad \alpha_j = \psi_j \qquad j = 1, \ldots, S+1$$

Method C

$$c_{mj} = \cos m\psi_j \qquad \alpha_j = \psi_j \qquad j = 1, \ldots, S$$

Problem IV Both h's infinite

Continuous **

$$b_j(m\Delta\xi) = \cos m\psi_j \qquad \psi_j = \frac{j\pi}{S} \qquad j = 1, 2, \ldots, \infty$$

Method G

$$c_{mj} = \cos m\psi_j \qquad \alpha_j = \frac{j\pi}{S} \qquad j = 1, 2, \ldots, S-1$$

* m refers to equation for m^{th} point, which is (m+1) row in conventional matrix notation for methods G and A based on mesh $\Delta\xi$ or $(m + \frac{1}{2})$ row in conventional matrix notation for methods C and F.

** Continuous "eigenvector" is actually an eigenfunction if m is allowed to vary continuously from 0 to S.

*** Vector corresponding to $\psi_1 = 0$ is considered as steady state solution, although its corresponding eigenvector is eigenvector of Y/A matrix.

Table IV-17

COMPONENTS OF INITIAL VECTOR - PROBLEM I

One h Finite; Other h Zero; Fluid Temperature Zero; Initial Condition Zero

$$T(m\Delta\xi,\ 0\ -) = T(m\Delta\xi,\ 0\ +) = 0 \qquad 0 \leq m\Delta\xi \leq 1$$

Continuous $\qquad j = 1, 2, \ldots, \infty$

$$a_j = \frac{2H}{(\nu_j)\left[(H+1)(\sin \nu_j) + (\nu_j)(\cos \nu_j)\right]}$$

Method G $\qquad j = 1, 2, \ldots, S+1$

$$g_j = \frac{2H}{\left[(2S)(\tan \frac{\mu_j}{2S})\right]\left[(H+\cos \frac{\mu_j}{S})(\sin \mu_j) + (S)(\sin \frac{\mu_j}{S})(\cos \mu_j)\right]}$$

Generalized Method A, Explicit $\qquad j = 1, 2, \ldots, S+1$

$$g_j = \frac{2H}{\left[2S \tan\frac{\mu_j}{2S}\right]\left[\left\{H\left[1-2r(1-\cos\frac{\mu_j}{S})\right]+\cos\frac{\mu_j}{S}\right\}(\sin \mu_j)+(S+\frac{2rH}{S})(\sin\frac{\mu_j}{S})(\cos \mu_j)\right]}$$

Graphical Method A, r = 1/2 $\qquad j = 1, 2, \ldots, S+1$

$$g_j = \frac{2H}{\left[2S \tan \frac{\mu_j}{2S}\right]\left[(H+1)(\cos \frac{\mu_j}{S})(\sin \mu_j) + (S+ \frac{H}{S})(\sin \frac{\mu_j}{S})(\cos \mu_j)\right]}$$

Method C $\qquad j = 1, 2, \ldots, S$

$$g_j = \frac{2H}{\left[2S \sin \frac{\mu_j}{2S}\right]\left[(H+1)(\sin \mu_j) + (2S+ \frac{H}{2S})(\tan \frac{\mu_j}{2S})(\cos \mu_j)\right]}$$

Averaged Methods

G and C multiply by	$\frac{q_j + 1}{2q_j}$	Backward
A multiply by	$\frac{q_j + 1}{2}$	Forward

Table IV-18

COMPONENTS OF INITIAL VECTOR - PROBLEM II

$h = \infty$

Initial Condition $T(m\Delta\xi, 0-) = 0 \qquad 0 \leq m\Delta\xi \leq 1$

$$T(m\Delta\xi, 0+) = \begin{cases} 0 & 0 < m\Delta\xi \leq 1 \\ 1 & 0 = m\Delta\xi \end{cases}$$

Continuous $\qquad \nu_j = \dfrac{(2j-1)\pi}{2} \qquad j = 1, 2, \ldots, \infty$

$$a_j = \frac{2}{\nu_j}$$

Approximate $\qquad j = 1, 2, \ldots, S$

g_j	Expansion for $(g_j - a_j)$ Small $\frac{\nu_j}{S}$, $0 \leq \frac{\nu_j}{S} < \pi$
Method G $\quad \dfrac{1}{S \tan \frac{\nu_j}{2S}}$	$-\dfrac{\nu_j}{6S^2} - \dfrac{\nu_j^3}{360S^4} \cdots$
Generalized Method A, Explicit or Implicit $\quad \dfrac{1}{S(\tan \frac{\nu_j}{2S})q_j}$	$\dfrac{\nu_j(12r-1)}{6S^2}$ $- \dfrac{\nu_j^3}{360}\left[1+120r\{1-6r(1-\gamma)\}\right] \cdots$
Graphical Method A $\quad \dfrac{1}{S(\tan \frac{\nu_j}{2S})(\cos \frac{\nu_j}{S})}$	$\dfrac{5\nu_j}{6S^2} - \dfrac{\nu_j^3}{360}(119) \cdots$
Method C $\quad \dfrac{1}{S \sin \frac{\nu_j}{2S}}$	$\dfrac{\nu_j}{12S^2} + \dfrac{7\nu_j^3}{2880S^4} \cdots$
Graphical Method F $\quad \dfrac{4}{\tan\frac{\mu_j}{2S}\left[3(S+\frac{1}{2})(\sin\frac{\{2S+1\}\mu_j}{2S})+(S-\frac{3}{2})(\sin\frac{\{2S-3\}\mu_j}{2S})\right]}$ where μ_j are the roots of the characteristic equation, Table IV-13.	

Table IV-19

COMPONENTS OF INITIAL VECTOR - PROBLEM III

Both h's Zero

Initial Condition $T(m\Delta\xi, 0+) = T(m\Delta\xi, 0-) = m\Delta\xi$

Continuous

$$a_j = \frac{2\left[1 - (-1)^j\right]}{\nu_j^2} \qquad \text{where } \nu_j = j\pi \qquad j = 1, 2, \ldots, \infty$$

Approximate $\qquad j = 1, 2, \ldots, S-1 \text{ or } S$

g_j	Expansion for $(g_j - a_j)$ Small $\frac{\nu_j}{S}$
Methods G and A Equivalent $\frac{\left[1 - (-1)^j\right]}{S^2(1-\cos\frac{\nu_j}{S})}$	$\left[1 - (-1)^j\right]\left[\frac{1}{6S^2} + \frac{\nu_j^2}{120S^4} \cdots\right.$
Method C $\frac{\left[1 - (-1)^j\right]}{2S^2(\tan\frac{\nu_j}{2S})(\sin\frac{\nu_j}{2S})}$	$-\left[1 - (-1)^j\right]\left[\frac{1}{12S^2} + \frac{7\nu_j^2}{960S^2} \cdots\right.$

Table IV-20

COMPONENTS OF INITIAL VECTOR - PROBLEM IV

Both h's Infinite

Initial Condition $T(m\Delta\xi,\ 0\ -) = 0 \qquad 0 \leq m\Delta\xi \leq 1$

$$T(m\Delta\xi,\ 0\ +) = \begin{pmatrix} 0 & m\Delta\xi = 0 \\ 0 & 0 < m\Delta\xi < 1 \\ 1 & m\Delta\xi = 1 \end{pmatrix}$$

Continuous

$$a_j = \frac{-2(-1)^j}{\nu_j} \qquad \text{where } \nu_j = j\pi$$

Approximate $\qquad j = 1, 2, \ldots, S-1$

g_j	Expansion for $(g_j - a_j)$ Small ν_j
Method G $\dfrac{-(-1)^j \sin \frac{\nu_j}{S}}{2S(\sin \frac{\nu_j}{2S})^2}$	$(-1)^j\left[\dfrac{\nu_j}{6S^2} + \dfrac{\nu_j^3}{360S^4} \cdots\right.$
Averaged Method G $\dfrac{-(-1)^j (\sin \frac{\nu_j}{S})(q_j + 1)}{4S(\sin \frac{\nu_j}{2S})^2 q_j}$	$(g_{j\ \text{Ave}} - a_j)$ $(-1)^j\left[\dfrac{\nu_j}{6S^2}(1-6r) + \dfrac{\nu_j^3}{360S^4}\left\{1-360r^2(1-\gamma)\right\}\right.$ $\cdots$

Table IV-21

EFFECTIVE INITIAL VECTOR COMPONENTS

Graphical Solutions - Problem II

Continuous

$$a_j = \frac{2}{\nu_j} \qquad \nu_j = \frac{(2j-1)\pi}{2} \qquad j = 1, 2, \ldots, \infty$$

Approximate Graphical

The following expressions result from pairs of q_j's of equal magnitude but opposite in sign, combined in a single q, with the corresponding initial vector components combined into effective weightings.

$g_{j\,E} = g_j + (-1)^n g_{S+1-j} \frac{c_{m\,S+1-j}}{c_{mj}}$ $1 \leq j < \frac{S}{2}$	Expansion for $(g_{j\,E} - a_j)$ Small ν_j $\left\lvert \frac{\nu_j}{S} \right\rvert < \frac{\pi}{2}$
Method G $\dfrac{\operatorname{ctn} \frac{\nu_j}{2S} - (-1)^{m+n}(\tan \frac{\nu_j}{2S})}{2S}$	$-\nu_j\left[\dfrac{1+3(-1)^{m+n}}{6S^2}\right] - \nu_j{}^3\left[\dfrac{1+15(-1)^{m+n}}{350S^4}\right]\cdots$
Method A $\dfrac{\operatorname{ctn} \frac{\nu_j}{2S} + (-1)^{m+n}(\tan \frac{\nu_j}{2S})}{S \cos \frac{\nu_j}{S}}$	$\nu_j\left[\dfrac{5+3(-1)^{m+n}}{6S^2}\right] + \nu_j{}^3\left[\dfrac{119+105(-1)^{m+n}}{360S^4}\right]\cdots$
Method C $\dfrac{1}{S \sin \frac{\nu_j}{2S}} + \dfrac{(-1)^{n+m-\frac{1}{2}}(\operatorname{ctn} \frac{m\nu_j}{S})}{\cos \frac{\nu_j}{2S}}$	$\dfrac{a_j(-1)^{n+m-\frac{1}{2}}}{2m} + \nu_j\left[\dfrac{2m+(-1)^{n+m-\frac{1}{2}}(3-8m^2)}{24mS^2}\right]\cdots$

Table IV-22

COMPARISON OF GRAPHICAL METHOD F

Problem II, S=5, $r=1/2$, $\gamma=0$

$\Delta\tau = 0.02$

Trigonometric Roots

Continuous, Methods G, A, and C	18°	54°	90°	126°	162°
Radians, ψ_j	0.3142	0.9426	1.5710	2.1994	2.8278
Graphical Method F	17.955°	52.62°	83.05°	111.5°	144.4°
α_j	0.3133	0.9183	1.4495	1.946	2.5199

Continuous Solution (First Five Damping Factors, Mesh $\Delta\xi/2$)

$$\begin{pmatrix} T_{\frac{1}{2}} \\ T_{\frac{3}{2}} \\ T_{\frac{5}{2}} \\ T_{\frac{7}{2}} \\ T_{\frac{9}{2}} \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.1992 & 0.1927 & 0.1800 & 0.1621 & 0.1397 \\ 0.5780 & 0.4192 & 0.1800 & -0.0285 & -0.1261 \\ 0.9003 & 0.3001 & -0.1800 & -0.1286 & 0.1000 \\ 1.1345 & -0.0664 & -0.1800 & 0.1797 & -0.0642 \\ 1.2576 & -0.3782 & 0.1800 & 0.0826 & 0.0221 \end{bmatrix} \begin{pmatrix} 0.9518^n \\ 0.6414^n \\ 0.2912^n \\ 0.0891^n \\ 0.0184^n \end{pmatrix}$$

Graphical Method F (Mesh $\Delta\xi/2$)

$$\begin{pmatrix} t_{\frac{1}{2}} \\ t_{\frac{3}{2}} \\ t_{\frac{5}{2}} \\ t_{\frac{7}{2}} \\ t_{\frac{9}{2}} \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.2084 & 0.2800 & 0.3346 & 0.1523 & 0.0243 \\ 0.5951 & 0.5100 & 0.1216 & -0.1676 & -0.0592 \\ 0.9236 & 0.3391 & -0.3051 & -0.0295 & 0.0720 \\ 1.1623 & -0.0982 & -0.1955 & 0.1892 & -0.0573 \\ 1.2876 & -0.4583 & 0.2578 & -0.1092 & 0.0220 \end{bmatrix} \begin{pmatrix} 0.9513^n \\ 0.6071^n \\ 0.1210^n \\ -0.3665^n \\ -0.8129^n \end{pmatrix}$$

Averaged Graphical Method F (Mesh $\Delta\xi/2$)

$$\begin{pmatrix} t_{\frac{1}{2}} \\ t_{\frac{3}{2}} \\ t_{\frac{5}{2}} \\ t_{\frac{7}{2}} \\ t_{\frac{9}{2}} \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.2033 & 0.2250 & 0.1875 & 0.0482 & 0.0023 \\ 0.5806 & 0.4098 & 0.6816 & -0.0531 & -0.0055 \\ 0.9011 & 0.2725 & -0.1710 & -0.0093 & -0.0067 \\ 1.1339 & -0.0789 & -0.1096 & 0.0599 & 0.0054 \\ 1.2556 & -0.3682 & 0.1450 & -0.0346 & -0.0021 \end{bmatrix} \begin{pmatrix} 0.9513^n \\ 0.6071^n \\ 0.1210^n \\ -0.3665^n \\ -0.8129^n \end{pmatrix}$$

Table IV-23

COMPARISON OF SOLUTIONS, GRAPHICAL METHOD A

Problem II, S=5, $r=1/2$, $\gamma=0$, $\Delta\tau=0.02$

Continuous Solution (First Five Damping Factors, Mesh $\Delta\xi$)

$$\begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.3934 & 0.3433 & 0.2546 & 0.1472 & 0.0437 \\ 0.7484 & 0.4036 & 0 & -0.1730 & -0.0832 \\ 1.0300 & 0.1311 & -0.2546 & 0.0562 & 0.1145 \\ 1.2109 & -0.2495 & 0 & 0.1069 & -0.1346 \\ 1.2732 & -0.4244 & 0.2546 & -0.1819 & 0.1415 \end{bmatrix} \begin{bmatrix} 0.9518^n \\ 0.6414^n \\ 0.2912^n \\ 0.0891^n \\ 0.0184^n \end{bmatrix}$$

Graphical Method A (Mesh $\Delta\xi$)

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.4103 & 0.5403 & -^* & -0.1402 & -0.0103 \\ 0.7804 & 0.6352 & - & 0.1649 & 0.0196 \\ 1.0741 & 0.2064 & - & -0.0536 & -0.0269 \\ 1.2627 & -0.3925 & - & -0.1019 & 0.0317 \\ 1.3277 & -0.6679 & - & 0.1734 & -0.0333 \end{bmatrix} \begin{bmatrix} 0.9511^n \\ 0.5878^n \\ -^* \\ -0.5878^n \\ -0.9851^n \end{bmatrix}$$

(* Zero roots not shown)

Effective Solution Graphical Method A (Mesh $\Delta\xi$)

n even

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.4000 & 0.4000 \\ 0.8000 & 0.8000 \\ 1.4472 & 0.1528 \\ 1.2944 & -0.4944 \\ 1.2944 & -0.4945 \end{bmatrix} \begin{bmatrix} 0.9511^n \\ 0.5878^n \end{bmatrix}$$

n odd

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.4206 & 0.6805 \\ 0.7608 & 0.4703 \\ 1.1010 & 0.2600 \\ 1.2310 & -0.2906 \\ 1.3610 & 0.8413 \end{bmatrix} \begin{bmatrix} 0.9511^n \\ 0.5878^n \end{bmatrix}$$

Averaged Graphical Method A (Mesh $\Delta\xi$)

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.4002 & 0.4289 & -^* & -0.0289 & -0.0002 \\ 0.7614 & 0.5043 & - & 0.0340 & 0.0005 \\ 1.0478 & 0.1638 & - & -0.0110 & -0.0006 \\ 1.2016 & -0.3116 & - & -0.0210 & 0.0008 \\ 1.2953 & -0.5303 & - & 0.0358 & -0.0009 \end{bmatrix} \begin{bmatrix} 0.9511^n \\ 0.5878^n \\ -0.5878^n \\ -0.9511^n \end{bmatrix}$$

(* Zero roots not shown)

Effective Solution Averaged Graphical Method A

n even

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.4000 & 0.4000 \\ 0.7619 & 0.5383 \\ 1.0472 & 0.1528 \\ 1.2024 & -0.3346 \\ 1.2944 & -0.4945 \end{bmatrix} \begin{bmatrix} 0.9511^n \\ 0.5878^n \end{bmatrix}$$

n odd

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix}_n = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 0.4004 & 0.4578 \\ 0.7608 & 0.4703 \\ 1.0484 & 0.1748 \\ 1.2310 & -0.2906 \\ 1.2962 & -0.5661 \end{bmatrix} \begin{bmatrix} 0.9511^n \\ 0.5878^n \end{bmatrix}$$

Table IV-24

SERIES TO MULTIPLY BY "UNAVERAGED" INITIAL VECTOR COMPONENTS TO OBTAIN SERIES REPRESENTATION OF INITIAL VECTOR COMPONENTS FOR AVERAGED METHODS

(Only for Use on Non-Oscillatory Part of Effective Initial Vectors Components)

General Average Applied at Time $(n+k)\Delta\tau$ $\qquad t_{n+k\ Ave} = \frac{1}{2}(t_n + t_{n+1})$

$$\frac{q_j^{+1}}{2\omega_j^k} = 1 + \frac{(k-\frac{1}{2})r\mu_j^2}{S^2} + \frac{r^2\mu_j^4}{S^4}\left[(\frac{1}{12r} + \gamma)(\frac{1}{2}-k) + \frac{k^2}{2}\right]$$
$$- \frac{r^3\mu_j^6}{S^6}\left[(\gamma^2 + \frac{\gamma}{6r} + \frac{1}{360r^2})(\frac{1}{2}-k) + \frac{k^2}{12}(\frac{1}{r} + 12\gamma) - \frac{(2k+1)(k+1)k}{12}\right] \cdots$$

Forward Average $\quad k=0 \quad$ Time $n(\Delta\tau)$

$$\frac{q_j^{+1}}{2} = 1 - \frac{r\mu_j^2}{2S^2} + \frac{r\mu_j^4}{24S^4}(1+12r\gamma) - \frac{r\mu_j^6}{720S^6}\left[60r\gamma(6r\gamma + 1) + 1\right] \cdots$$

Central Average $\quad k=1/2 \quad$ Time $(n+1/2)\Delta\tau$

$$\frac{q_j^{+1}}{2\sqrt{q_j}} = 1 + \frac{r^2\mu_j^4}{8S^4} - \frac{r^2\mu_j^6}{48S^6}\left[1 + 6r(2\gamma - 1)\right] \cdots$$

Backward Average $\quad k=1 \quad$ Time $(n+1)\Delta\tau$

$$\frac{q_j^{+1}}{2q_j} = 1 + \frac{r\mu_j^2}{2S^2} + \frac{r\mu_j^4}{24S^4}\left[12r(1-\gamma) - 1\right] + \frac{r\mu_j^6}{720S^6}\left[1+60r(1-\gamma)(6r\{1-\gamma\}-1)\right] \cdots$$

Graphical Methods $\qquad r=1/2 \qquad \gamma = 0$

Forward: $$\frac{q_j^{+1}}{2} = 1 - \frac{\mu_j^2}{4S^2} + \frac{\mu_j^4}{48S^4} - \frac{\mu_j^6}{1440S^6} \cdots$$

Central: $$\frac{q_j^{+1}}{2\sqrt{q_j}} = 1 + \frac{\mu_j^4}{32S^4} + \frac{\mu_j^6}{96S^6} \cdots$$

Backward: $$\frac{q_j^{+1}}{2q_j} = 1 + \frac{\mu_j^2}{4S^2} + \frac{5\mu_j^4}{48S^4} + \frac{61\mu_j^6}{1440S^6} \cdots$$

Table IV-25

AVERAGED INITIAL VECTOR COMPONENTS - PROBLEM II

Continuous

$$a_j = \frac{\nu_j}{2} \qquad \nu_j = \frac{(2j-1)\pi}{2} \qquad j = 1, 2, \ldots, \infty$$

Approximate*

$g_{j\ Ave}$	Expansion $(g_{j\ Ave} - a_j)$ Small ν_j
Averaged Methods A and G (Equivalent) $\dfrac{q_j + 1}{2q_j\ S \tan \frac{\nu_j}{2S}}$	$\dfrac{\nu_j(6r-1)}{6S^2} + \dfrac{\nu_j^3\left[60r\{6r(1-\gamma)-1\}-1\right]}{360S^4} \ldots$
Averaged Method C $\dfrac{q_j + 1}{2q_j\ S \sin \frac{\nu_j}{2S}}$	$\dfrac{\nu_j(12r+1)}{12S^2} + \dfrac{\nu_j^3\left[7+120r\{24r(1-\gamma)-1\}\right]}{2880S^4} \ldots$

* Approximate initial vectors for averaged solutions found by multiplying unaveraged initial vector by

$\dfrac{q_j + 1}{2q_j}$ (Backward) for methods G and C

or by $\dfrac{q_j + 1}{2}$ (Forward) for method A

Table IV-26

AVERAGED INITIAL VECTOR COMPONENTS - PROBLEM III

Continuous

$$a_j = \frac{2\left[1-(-1)^j\right]}{\nu_j^2} \qquad \nu_j = j\pi \qquad j = 1, 2, \ldots, \infty$$

Approximate $\qquad j = 1, 2, \ldots, S \text{ or } S+1$

$g_{j\ Ave}$	Expansion for $(g_{j\ Ave} - a_j)$ Small ν_j
Averaged Method G $\frac{\left[1-(-1)^j\right](q_j+1)}{2q_jS^2(1-\cos\frac{\nu_j}{S})}$	$\left[1-(-1)^j\right]\left[\frac{(6r+1)}{6S^2} + \frac{\nu_j^2}{120S^4}\left\{1+120r(1-\gamma)\right\} \ldots\right.$
Averaged Method A (Explicit) $\frac{\left[1-(-1)^j\right](q_j+1)}{2S^2(1-\cos\frac{\nu_j}{S})}$	$-\left[1-(-1)^j\right]\left[\frac{6r-1}{6S^2} + \frac{\nu_j^2}{120S^4} \quad \ldots\right.$
Averaged Method C $\frac{\left[1-(-1)^j\right](q_j-1)}{2q_j(\tan\frac{\nu_j}{2S})(\sin\frac{\nu_j}{2S})}$	$\left[1-(-1)^j\right]\left[\frac{12r-1}{12S^2} + \frac{\nu_j^2}{960S^4}\left\{7+120r\left[4r(1-\gamma)+1\right]\right\} \ldots\right.$

Table IV-27

SERIES TO MULTIPLY BY OSCILLATORY SERIES IN EFFECTIVE INITIAL VECTOR TO OBTAIN OSCILLATORY SERIES IN AVERAGED SOLUTION

(Graphical Solution, $\gamma=0$, $r=1/2$ Only)

General Average at Time $(n+k)\Delta\tau$

$$\frac{1-q_j}{2q_j^{\,k}} = \frac{\mu_j^2}{4S^2} + \frac{\mu_j^4}{48S^4}(4k-1) \quad \cdots$$

Forward Average $k=0$ Time $n\Delta\tau$

$$\frac{1-q_j}{2} = \frac{\mu_j^2}{4S^2} - \frac{\mu_j^4}{48S^4} \quad \cdots$$

Central Average $k=1/2$ Time $n\Delta\tau$

$$\frac{1-q_j}{2\sqrt{q_j}} = \frac{\mu_j^2}{4S^2} + \frac{\mu_j^4}{48S^4} \quad \cdots$$

Backward Average $k=1$ Time $(n+1)\Delta\tau$

$$\frac{1-q_j}{2q_j} = \frac{\mu_j^2}{4S^2} + \frac{\mu_j^4}{16S^4} \quad \cdots$$

To find expansions for effective initial vector components for averaged graphical methods: write series for effective initial vector component into the sum of a series of non-oscillatory terms and a series of oscillatory terms containing $(-1)^n$ as a factor; multiply non-oscillatory series by appropriate series on Table IV-22 for $\gamma=0$ and $r=1/2$; multiply oscillatory series by appropriate series on this table and recombine series to obtain series for the effective initial vector components for an averaged initial vector.

Table IV-28

AVERAGED EFFECTIVE INITIAL VECTOR COMPONENTS - PROBLEM II

Continuous

$$a_j = \frac{2}{\nu_j} \qquad \nu_j = \frac{(2j-1)\pi}{2} \qquad j = 1, 2, \ldots, \infty$$

Approximate Effective Initial Vectors

$1 \leq j \leq \frac{S}{2}$ $g_{j\ \text{Ave}\ E}$	Expansion for $(g_{j\ \text{Ave}\ E} - a_j)$ $0 \leq \frac{\nu_j}{S} \leq \frac{\pi}{2}$
Averaged Methods A and G (Equivalent) $\frac{1}{2S}\left[\operatorname{ctn}\frac{\nu_j}{2S} + \frac{\operatorname{ctn}\frac{\nu_j}{2S}}{\cos\frac{\nu_j}{S}} + \frac{(-1)^{m+n}(\tan\frac{\nu_j}{2S})(1-\cos\frac{\nu_j}{S})}{\cos\frac{\nu_j}{S}}\right]$	$\frac{\nu_j}{3S^3} + \frac{\nu_j^3\left[59+45(-1)^{n+m}\right]}{360S^4}\ldots$
Averaged Method C $\frac{1}{2S\cos\frac{\nu_j}{S}}\left[\frac{1+\cos\frac{\nu_j}{S}}{\sin\frac{\nu_j}{2S}} + \frac{(-1)^{n+m-\frac{1}{2}}(\cos\frac{\nu_j}{S}-1)(\operatorname{ctn}\frac{m\nu_j}{S})}{\cos\frac{\nu_j}{2S}}\right]$	$\frac{\nu_j\left[7m-3(-1)^{n+m-\frac{1}{2}}\right]}{12mS^2}\ \ldots$

Table IV-29

COMPLETE SOLUTION - PROBLEM V

Semi-Infinite Solid

Infinite h at $m\Delta\xi = 0$

Initial $T(m\Delta\xi) = 0 \qquad 1 \leq (m\Delta\xi) \leq \infty$

Continuous

$$T(m\Delta\xi,\ n\Delta\tau) = 1 - \operatorname{erf}\frac{m}{2\sqrt{rn}} = 1 - \frac{2}{\pi}\int_0^{\infty} \frac{(\sin m\psi)e^{-\psi^2 rn}\, d\psi}{\psi}$$

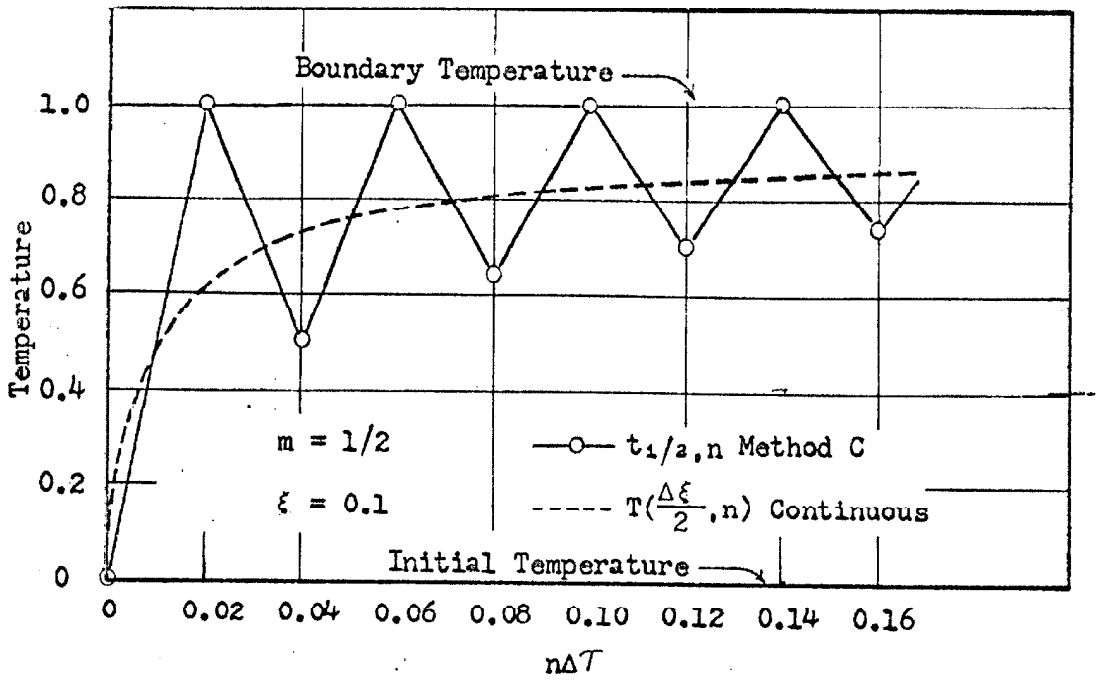
$$= 1 - \operatorname{erf}\frac{m\Delta\xi}{2\sqrt{n\Delta\tau}} = 1 - \frac{2}{\pi}\int_0^{\infty} \frac{\left[\sin(m\Delta\xi)\nu\right]e^{-\nu^2\Delta\tau n}\, d\nu}{\nu}$$

Approximate

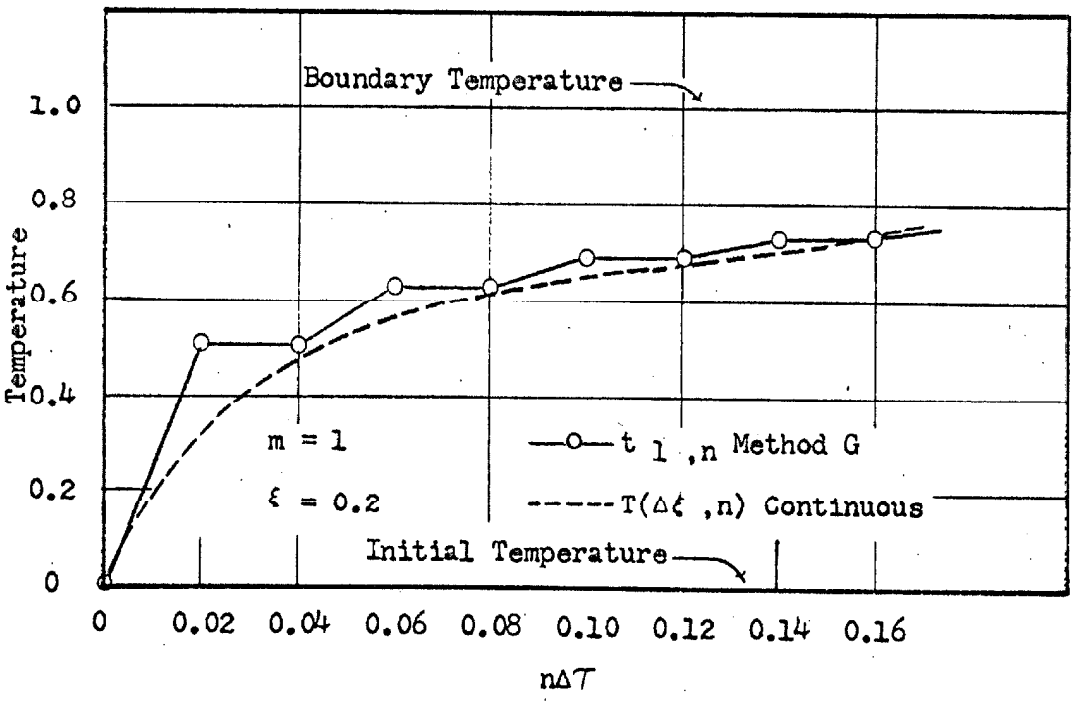
Method G:

$$t_{m,n} = 1 - \frac{1}{\pi}\int_0^{\pi} \frac{\left[\sin m\alpha\right]\left[1-2r(1-\cos\alpha)\right]^n d\alpha}{\tan\frac{\alpha}{2}}$$

$$= 1 - \frac{1}{\pi}\int_0^{\frac{\pi}{\Delta\xi}} \frac{\Delta\xi\left[\sin m(\Delta\xi)\mu\right]\left[1-2r(1-\cos\mu\,\Delta\xi)\right]^n d\mu}{\tan\frac{(\Delta\xi)\mu}{2}}$$

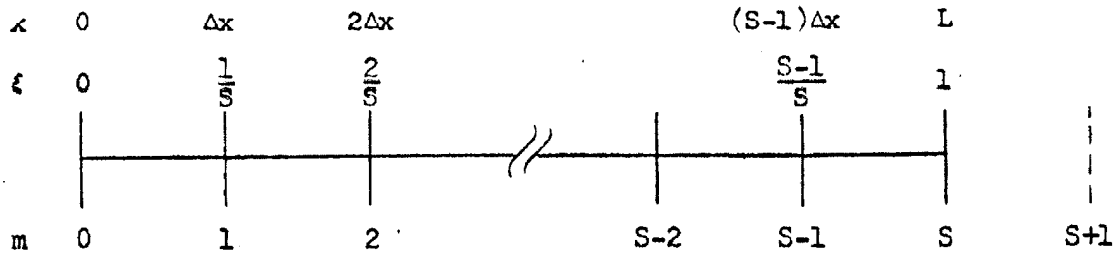


a. Method C, Mesh $\Delta\xi$ /2

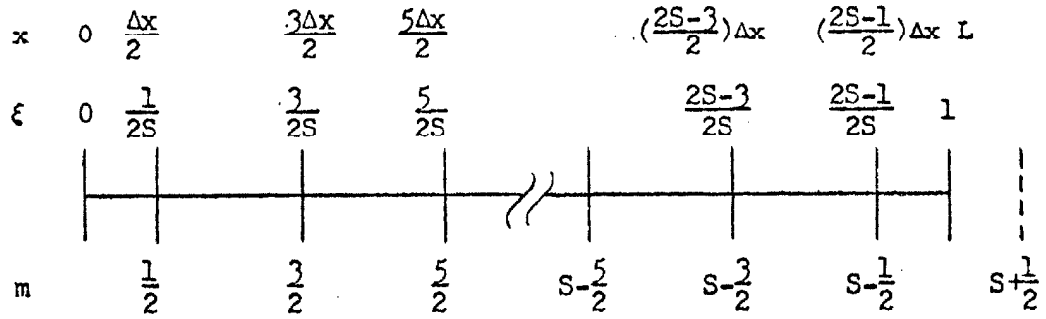


b. Method G, Mesh $\Delta\xi$.

Figure IV-1. Comparison of graphical methods G and C.



a. MESH $\Delta\xi$



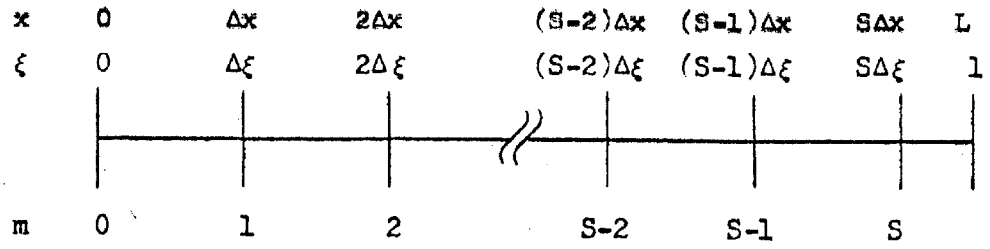
b. MESH $\Delta\xi/2$

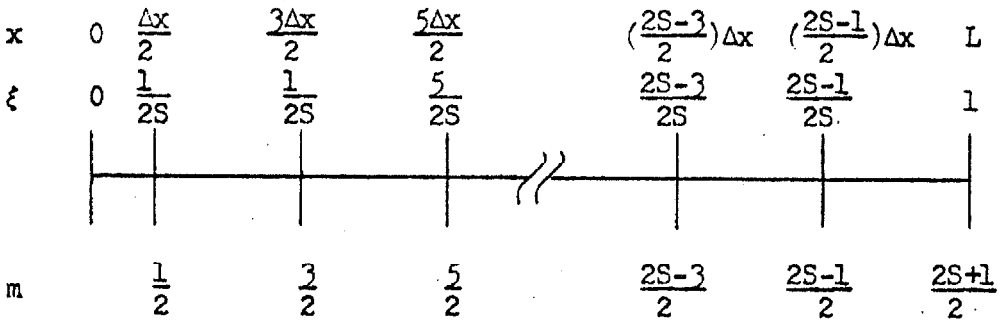
Relationships $\xi = \frac{x}{L}$ $\Delta\xi = \frac{\Delta x}{L} = \frac{1}{S}$

$\xi = m\Delta\xi = \frac{m}{S}$

$x = m\Delta x = \frac{m}{S}L$

Figure IV-2. Meshes $\Delta\xi$ and $\Delta\xi/2$.





Relationships $\xi = \frac{x}{L}$ $\Delta\xi = \frac{\Delta x}{L} = \frac{1}{S+\frac{1}{2}}$

$$\xi = m\Delta\xi = \frac{m}{S+\frac{1}{2}}$$

$$x = m\Delta x = \left(\frac{m}{S+\frac{1}{2}}\right)L$$

Figure IV-3. Other meshes.

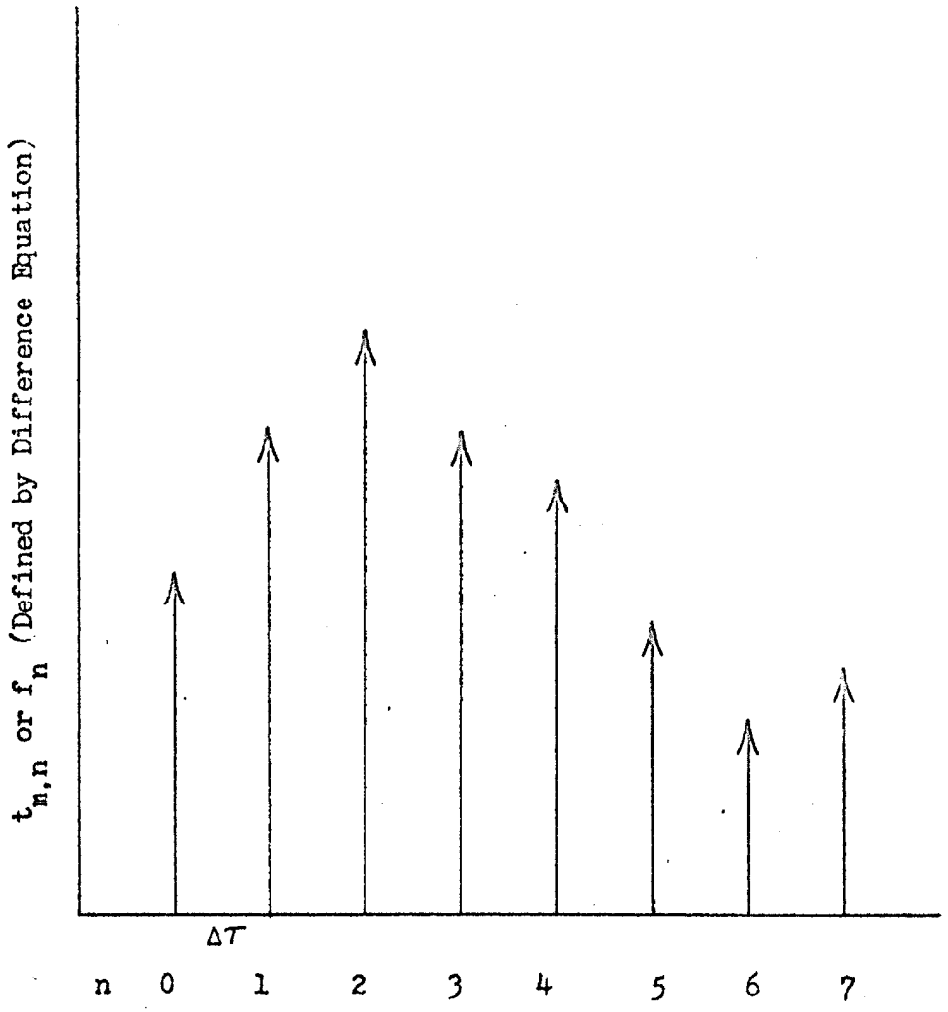
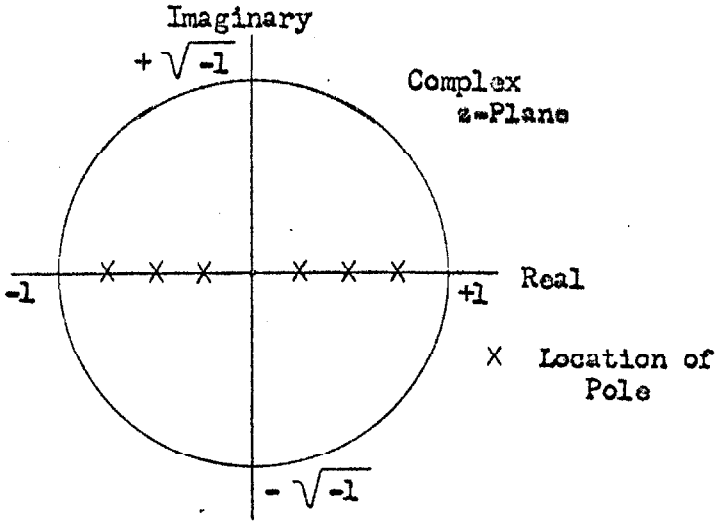
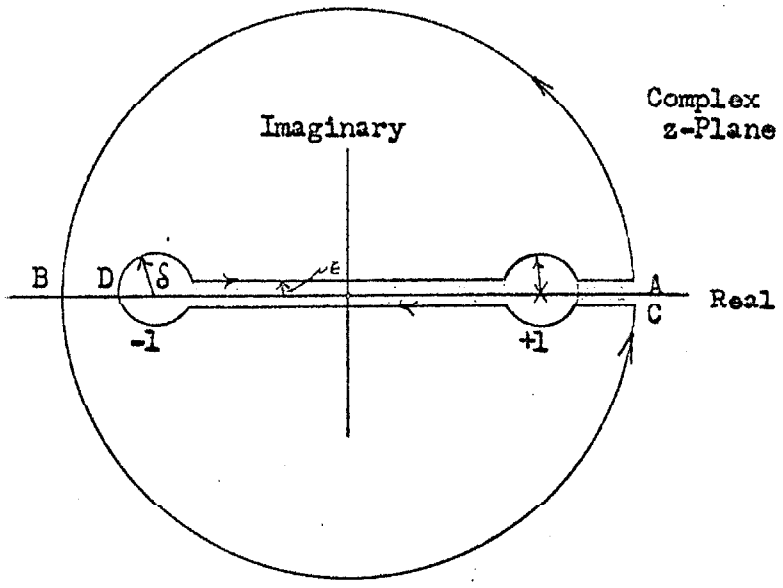


Figure IV-4. Example of a sampled or difference signal.



a. Integration Contour for Inverting Transform with Poles Inside Unit Circle (Stable Solution).



b. Integration Contour for Inverting Transform with Pole at +1 and Branch Cut from -1 to +1.

$$\bar{t}_m = \left[\frac{z}{z-1}\right]\left[\frac{1}{z+\sqrt{z^2-1}}\right]^m$$

Figure IV-5. Contour integrations for inverting z-transforms.

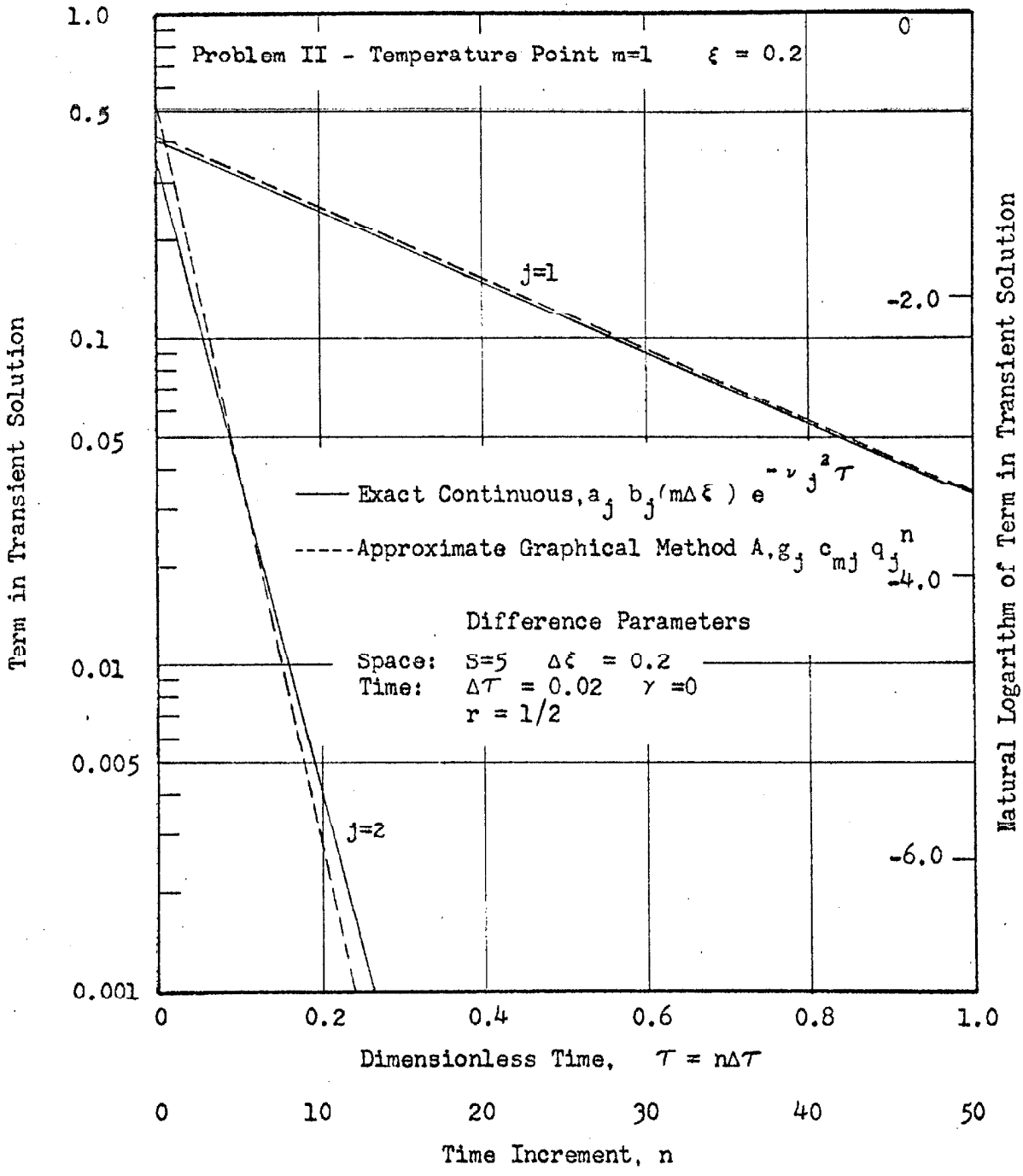


Figure IV-6. Semi-logarithmic graph of term in transient solution versus time.

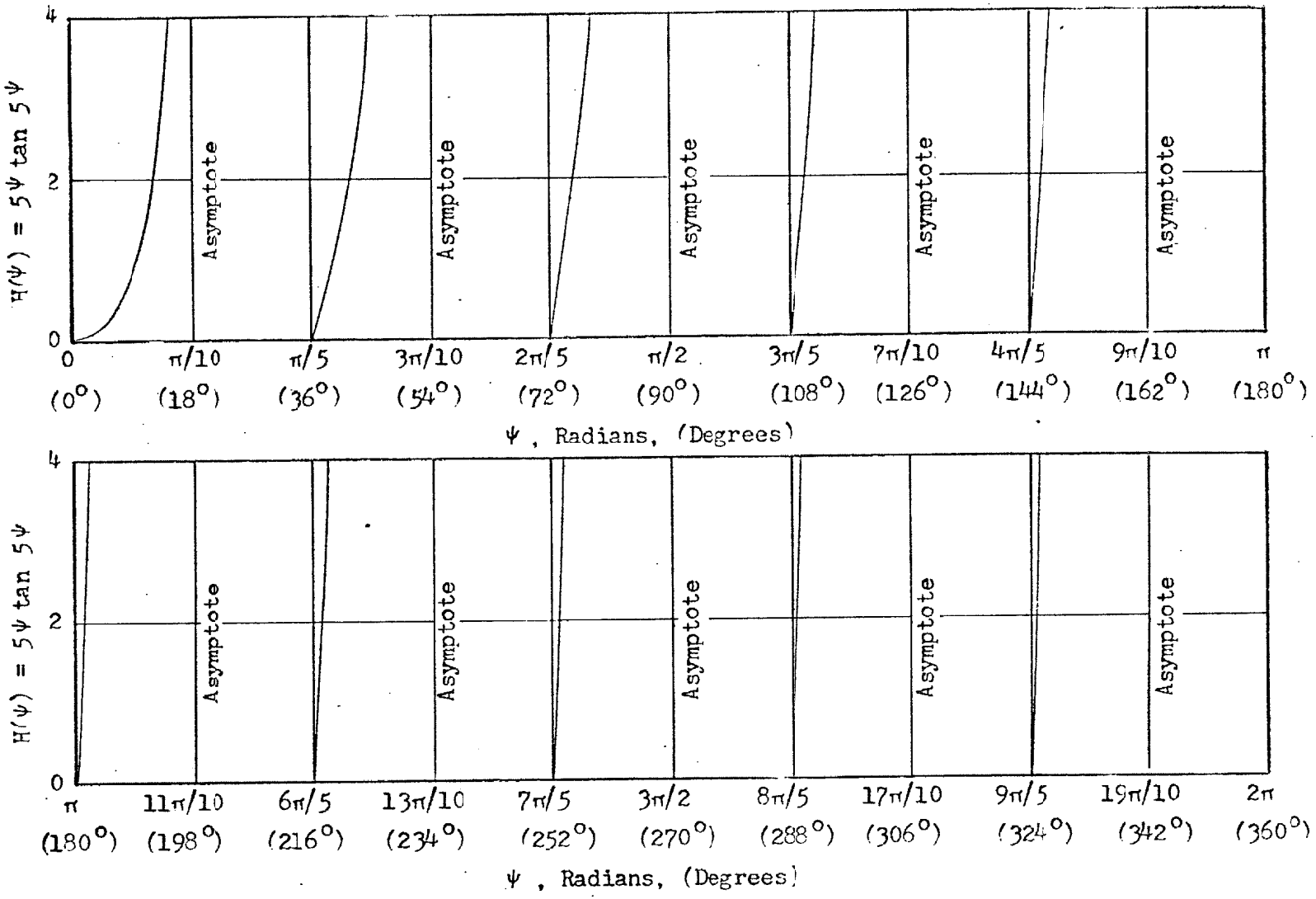


Figure IV- 7. Characteristic equation, continuous solution, first ten roots.
Note: $\nu = 5\psi$

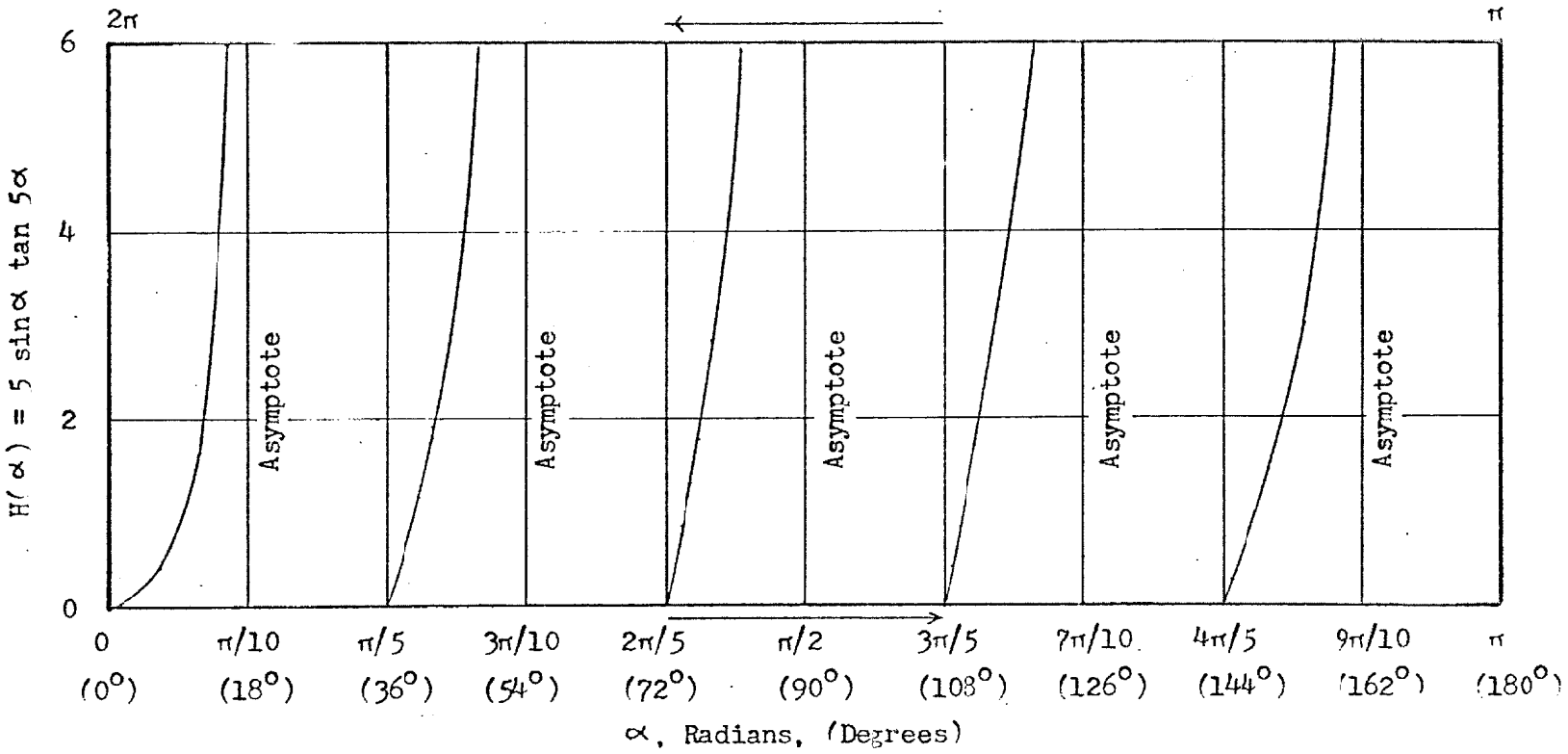


Figure IV-8 . Characteristic equation, method G, real roots for S=5, first five roots.
Note: $\mu = 5\alpha$

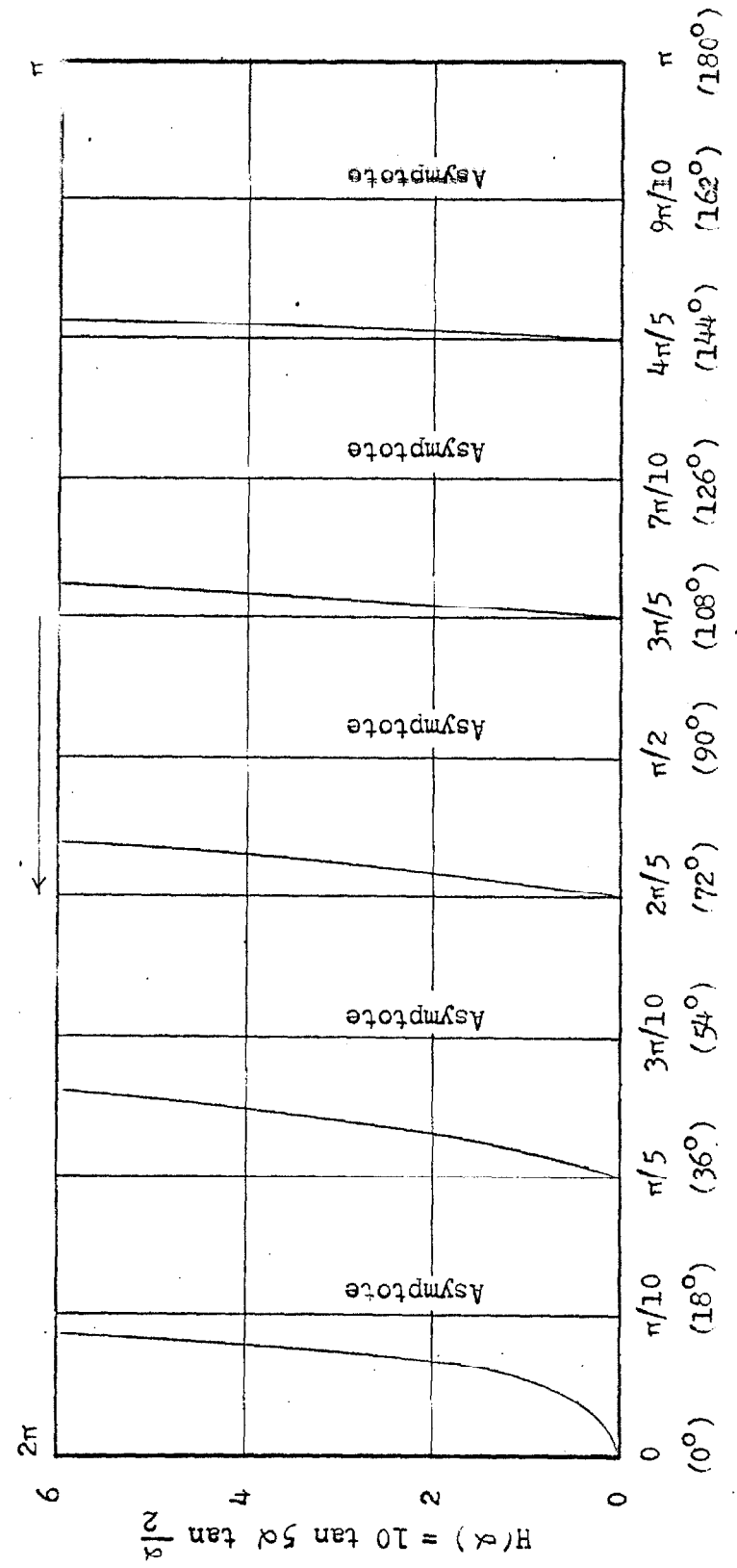


Figure IV-9. Characteristic equation, method C, S=5, five roots.
Note: $\mu = 5\alpha$

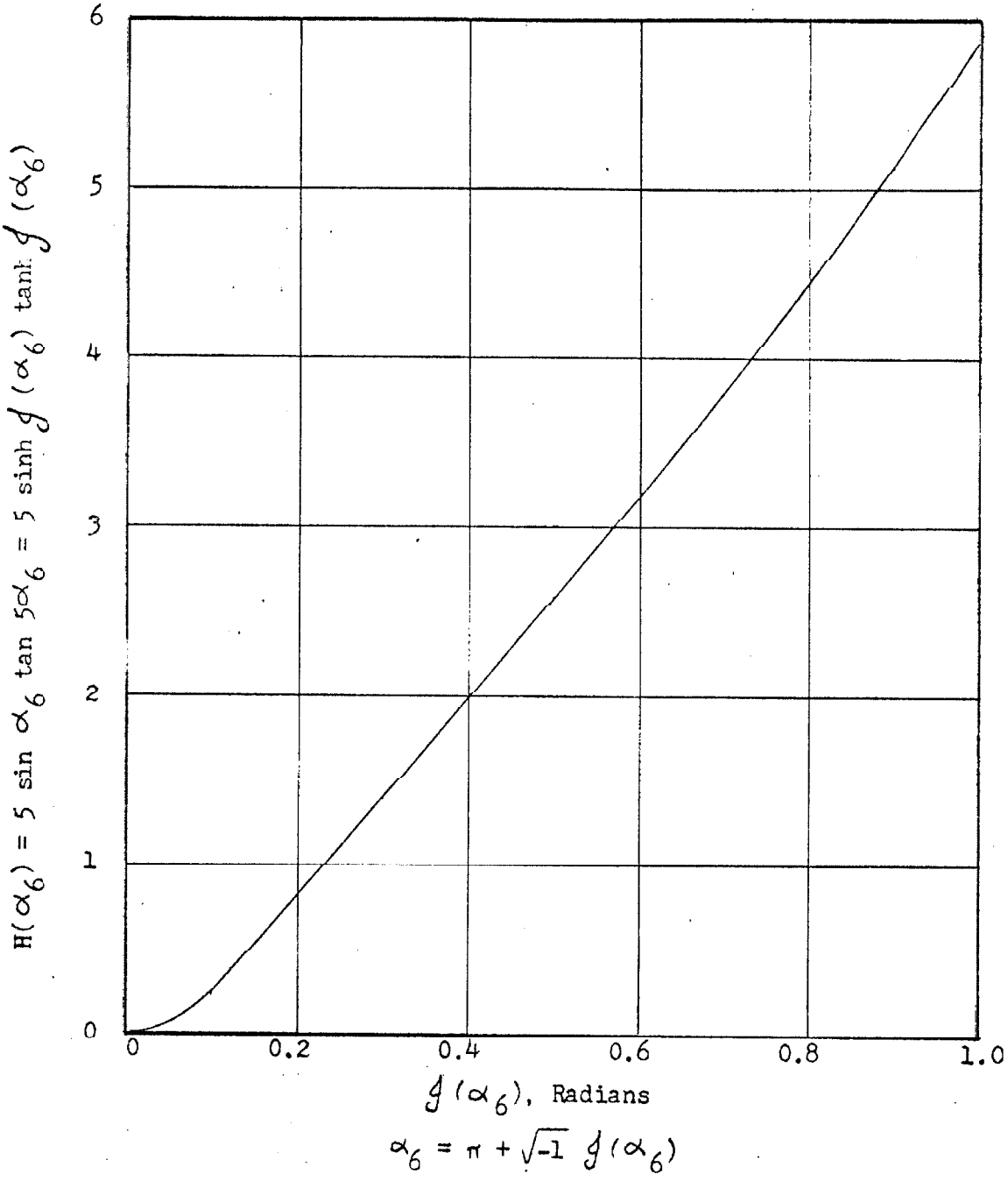


Figure IV-10. Characteristic equation, method G, imaginary root (sixth root) for S=5.
Note: $\mu = 5\alpha$

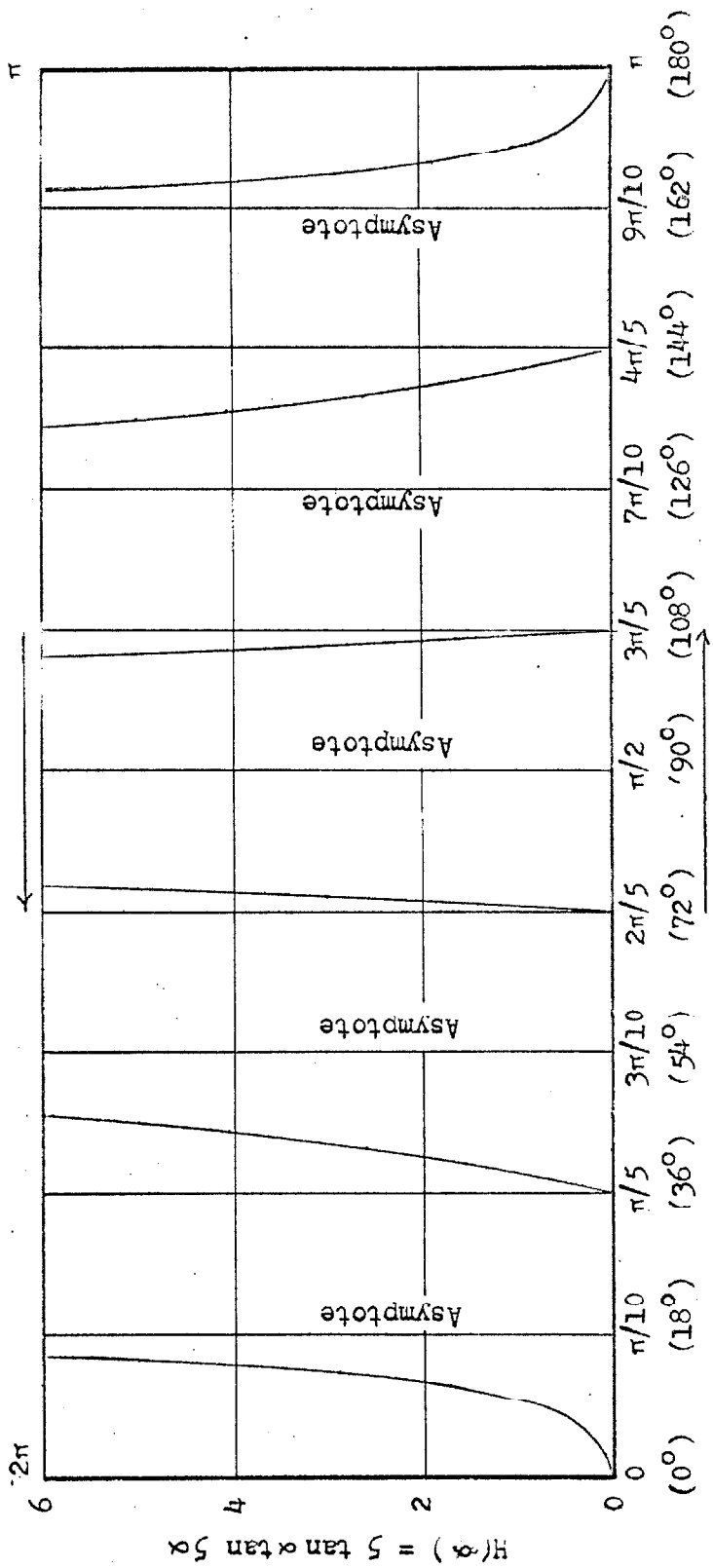


Figure IV-11. Characteristic equation graphical method A, S=5, all six roots
Note: $\mu = 5\alpha$

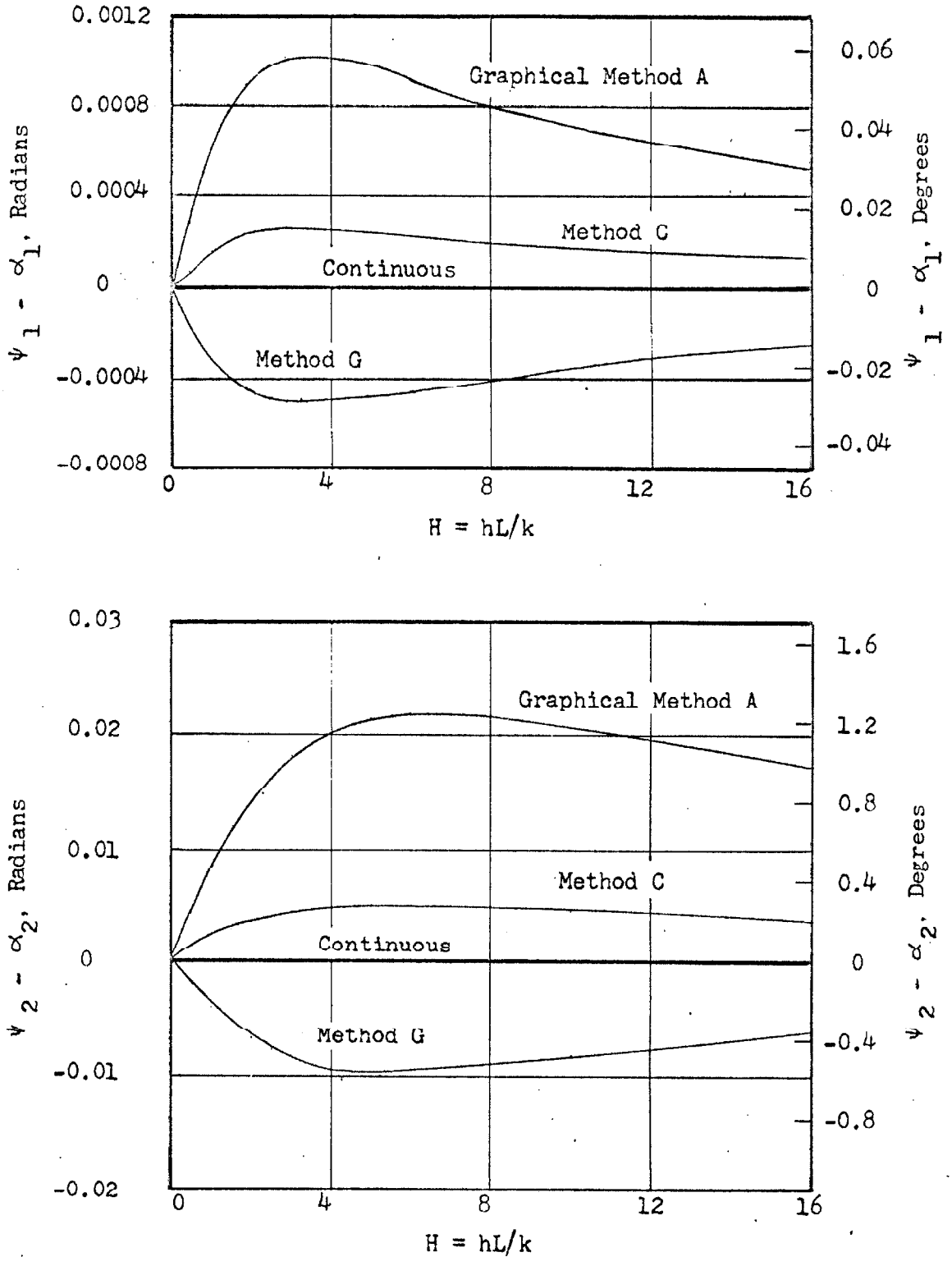


Figure IV-12. Error in first two roots, method G, graphical method A, and method C, S=5.

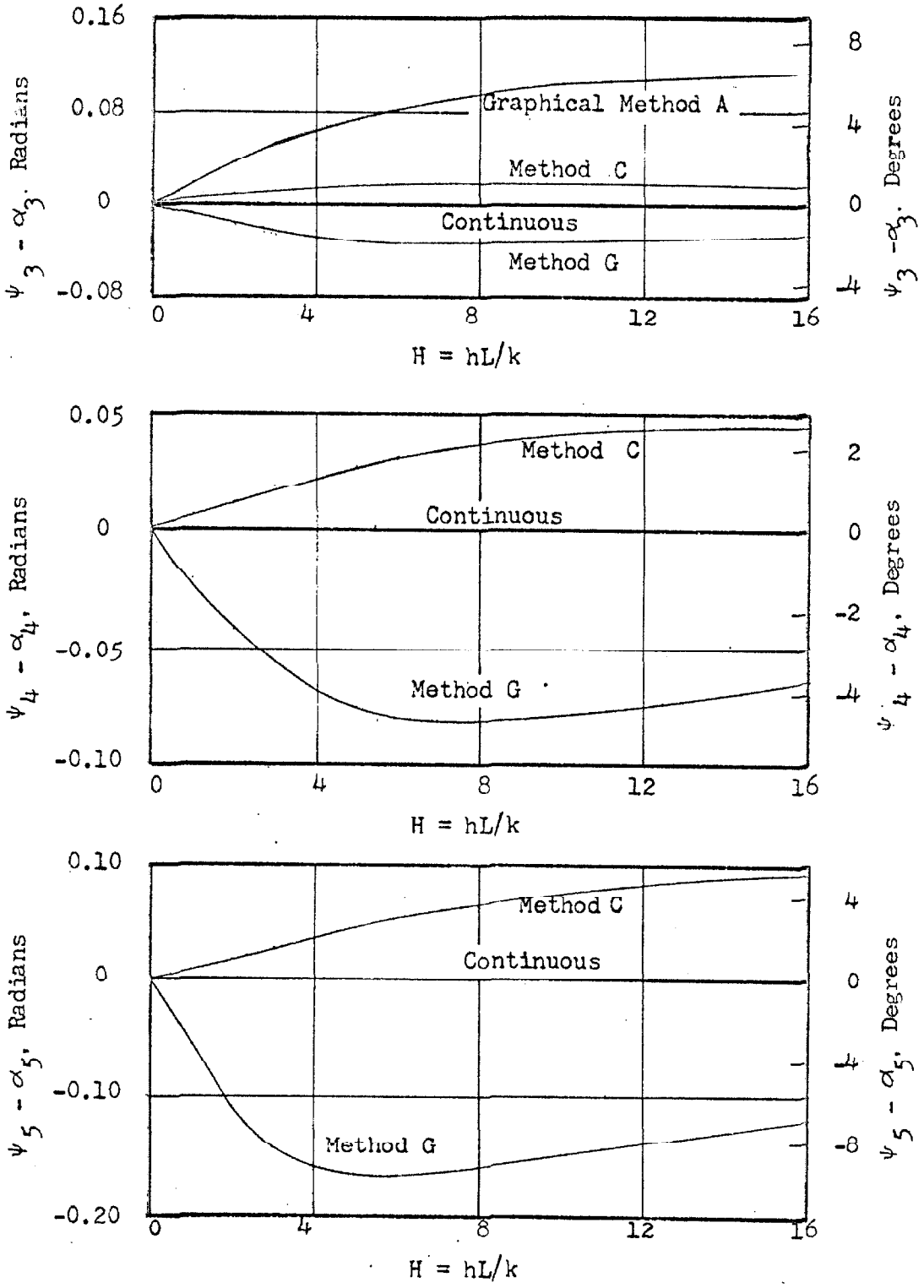


Figure IV-13. Error in roots three, four, and five, S=5.

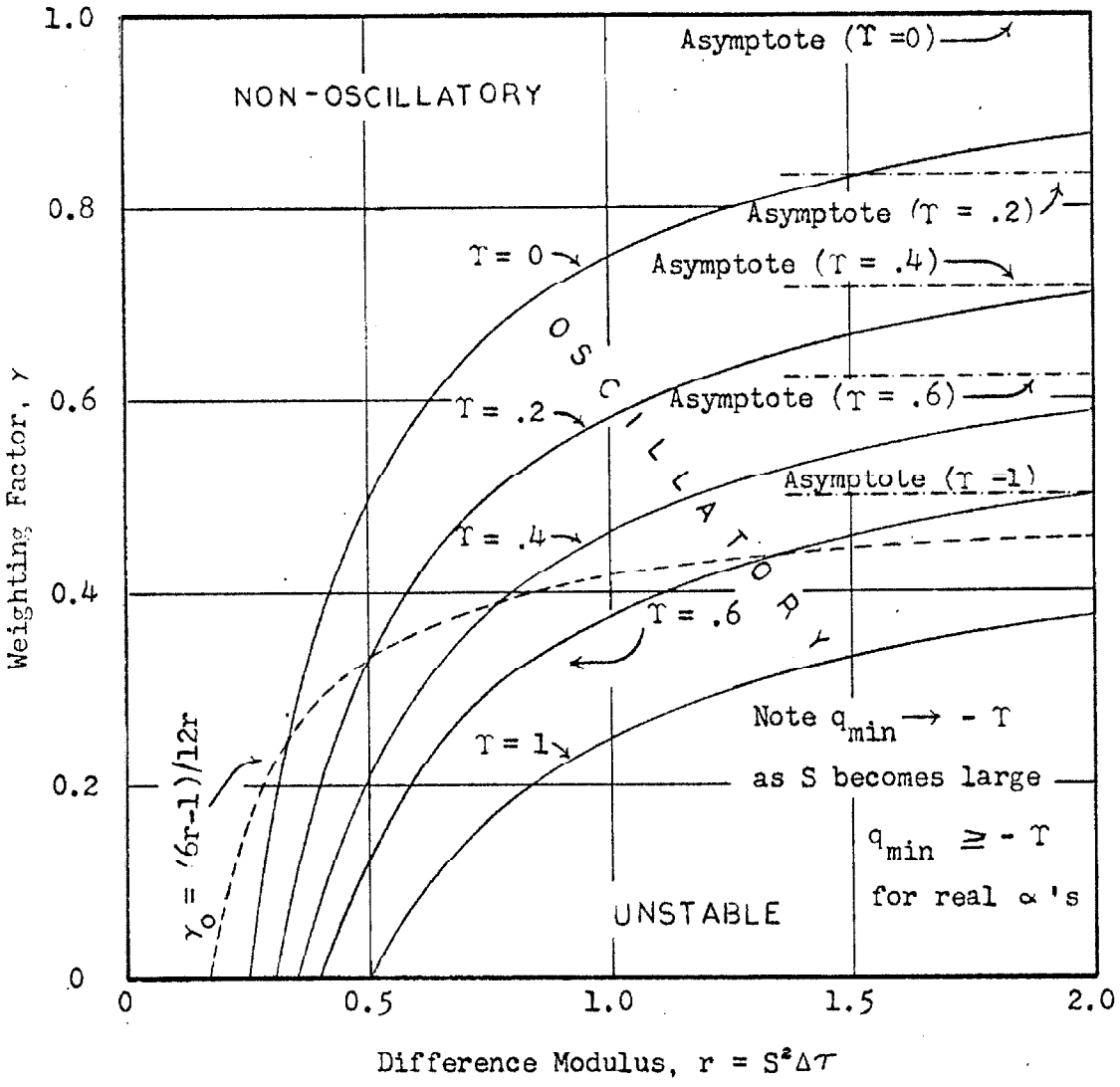


Figure IV-14. Sufficient conditions such that $q_{min} \geq -\tau$ for methods where minimum norm of Y/A matrix is equal to or less than $4S^2$, or all trigonometric roots α_j are real.

$$\gamma = \frac{4r-(1+\tau)}{4r(1+\tau)}$$

Selection of r and γ above and to the left of a line for constant τ assures $q_{min} \geq -\tau$.

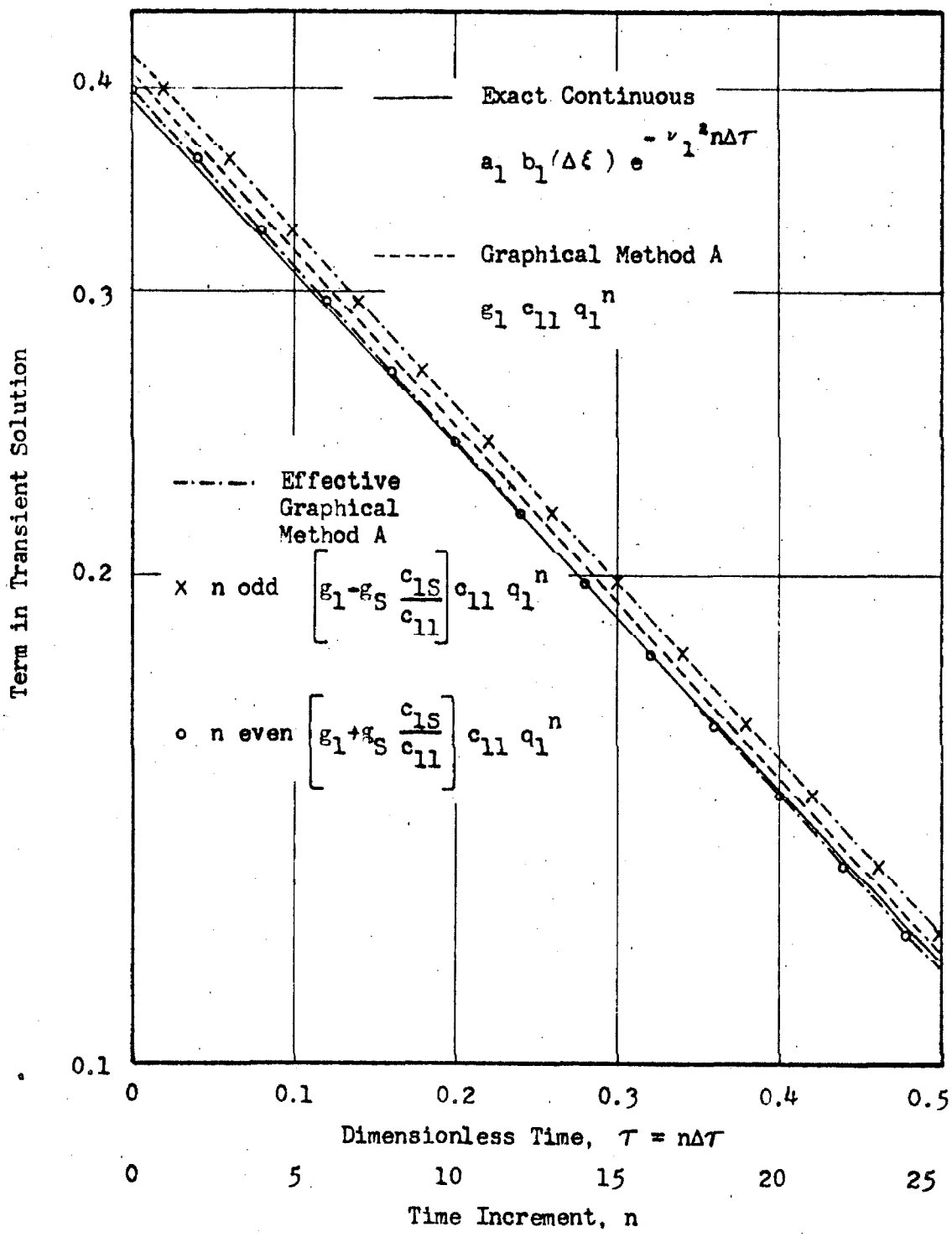


Figure IV-15. Semi-logarithmic graph of term in transient solution versus time showing effective initial vector component for graphical method A.

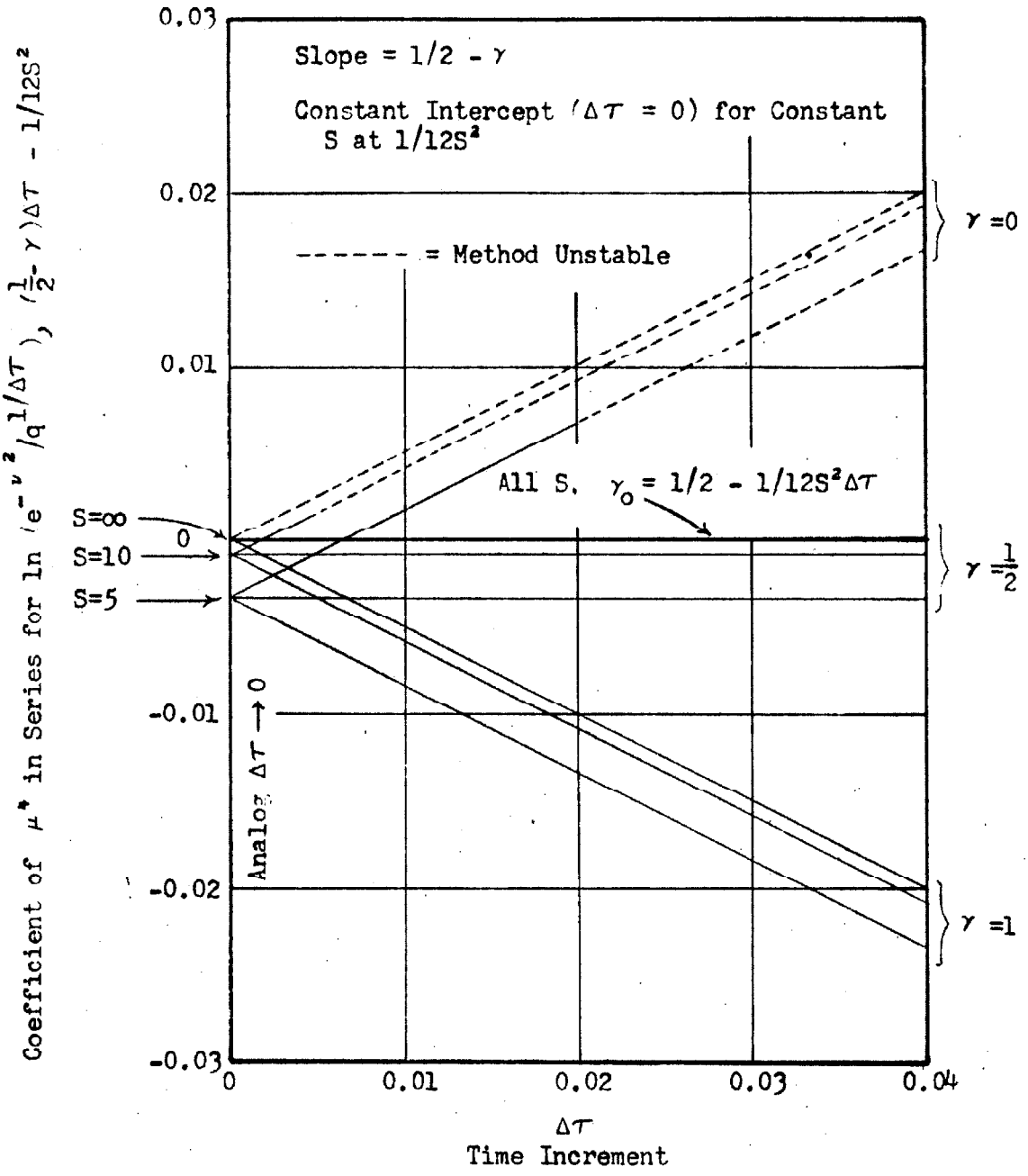


Figure IV-16. Coefficient of μ^4 term in series for the logarithm of $e^{-\nu^2}/q^{1/\Delta\tau}$ as a function of the differencing parameters.

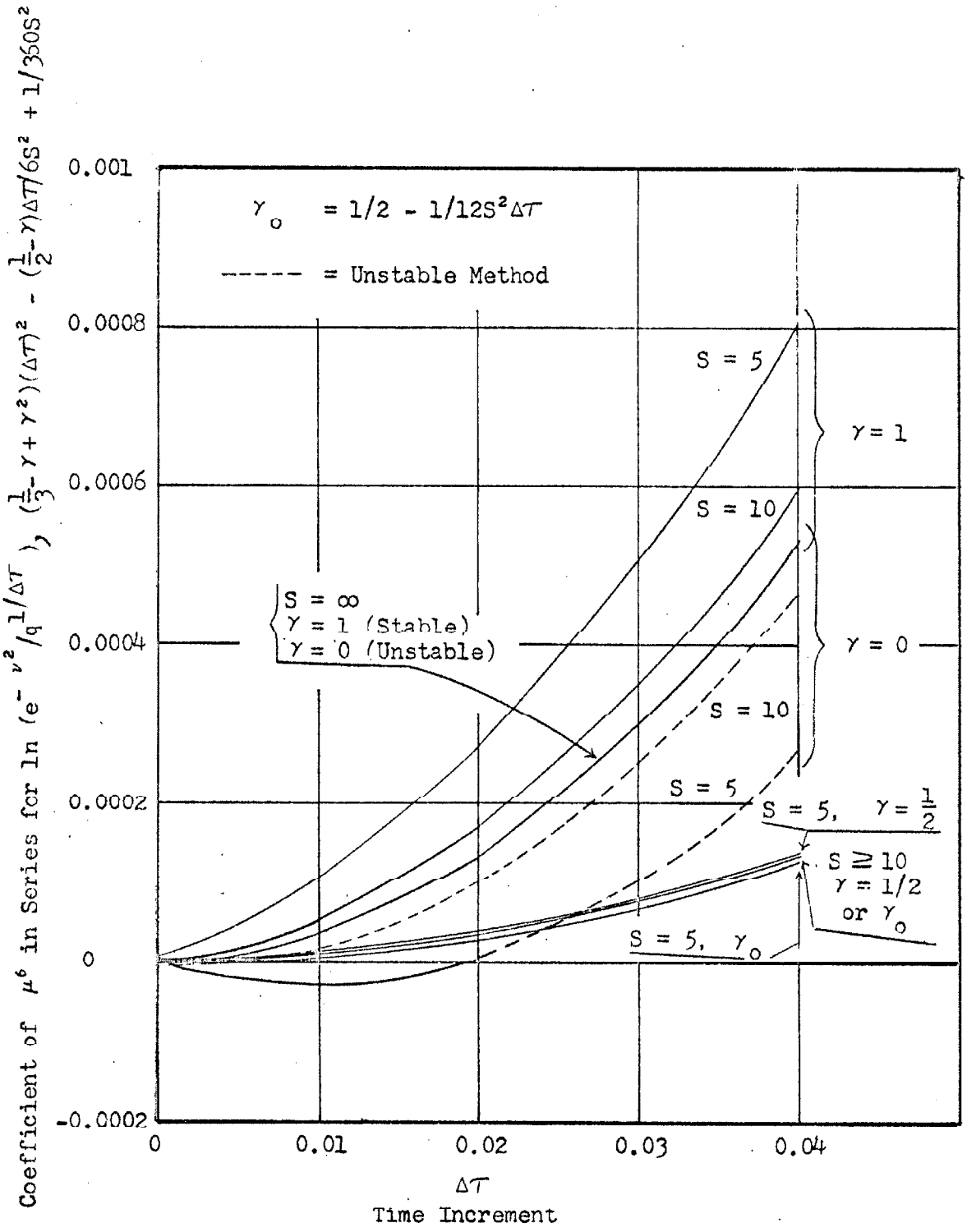


Figure IV-17. Coefficient of μ^6 term in series for the logarithm of $e^{-\nu^2/q^{1/\Delta\tau}}$ as a function of the differencing parameters.

CHAPTER V

ACCURACY OF TRANSIENT SOLUTIONS USING AN ASYMMETRIC NETWORK OF NODES

(A) INTRODUCTION

As mentioned in the introductory Chapter I, and as pointed out by MacNeal (8), locating the temperature points on the corners of rectangles or other regular polygons such as equilateral triangles, either of which defines a regular grid, is not a satisfactory arrangement for a solid with a curved or irregular boundary. For this type of solid a regular arrangement of nodes usually requires an extremely find grid with a very large number of points to follow the curved boundary closely. A moderate-size regular mesh that does not follow the boundary has been suggested by others (32, 10), by assuming that the nodes closest to the boundary surface actually fall on the boundary. For problems where the temperature on the boundary surface is specified (infinite heat-transfer coefficient) a somewhat better approximation is to interpolate linearly for this temperature (32). However, when the boundary equations have a finite non-zero heat-transfer coefficient, many techniques are available (6, 10, 17, 32), although none of them appears to be completely satisfactory (8), or to be widely used.

MacNeal suggested that by not restricting the location of the points to the corners of regular polygons, but by locating the points within and on the solid boundary conforming to a much less restrictive set of rules, an irregular boundary could be followed exactly, with a fairly coarse network. The equations for each node can then be found by an energy balance, and in general the equation for each node is different. These points are discussed in more detail in Chapter II, section C. A boundary condition involving a heat-transfer coefficient is easily included in the energy balance for a surface node by adding a term representing the energy in from the fluid of the form

$$(t_{f_i,n} - t_{i,n})h\,A_{f_i} = \text{Heat into surface node from fluid at time } n \qquad (V\text{-}1)$$

where i is allowed to take on values of the nodes on the boundary surface; t_{f_i} is the average fluid temperature opposite the ith surface node ; and A_{f_i} is the surface area through which the energy flows. For the two-dimensional problem this area is the length of surface (assuming unit thickness) terminated by the perpendicular bisectors of the ith node (see Figure V-1). The dimensionless conductance between the ith surface node and the ith fluid node is

$$y_{i,f_i} = \frac{h\,A_{f_i}}{k_0} \qquad (V\text{-}2)$$

The volume for heat capacity associated with the surface node for the two-dimensional case is bounded by the perpendicular bisectors of the ith node and the boundary of the solid. Therefore, an approximate solution for an irregular solid, using a well designed fairly coarse network of points, based on MacNeal's rules for locating the points and calculating the conductances and capacitances of the nodes, would be expected to be more accurate than a regular mesh using the same number of points but not following the boundary exactly.

However, further study of the equations for the space truncation error associated with asymmetric nodes, as given in MacNeal, indicates that an approximate solution to the diffusion equation based upon an asymmetric network of nodes does not necessarily follow Richtmyer's (3) definition of consistency (see Chapter III, section A). This is true because Richtmyer has shown that, for an approximation to be consistent, the truncation error must go to zero as the network is refined; however, as shown by MacNeal's equations (but not pointed out or discussed), the series expansion for the contribution to this truncation error of a general asymmetric node shows that terms occur in this difference that do not go to zero as the neighboring nodes are moved closer. This also can be shown to be caused by the asymmetry in the node location. Since, according to the Richtmyer theory, the approximation must be both consistent and stable to be convergent to the continuous solution as the mesh is refined, the accuracy of an

approximate solution found using asymmetric nodes is in doubt. Consequently, the question is for what types of problems is it appropriate to use an asymmetric network, or to include asymmetric nodes in regular networks, and how should these nodes be located.

It should be pointed out that one of the distinguishing features of an asymmetric network of points is that the equation for each of the nodes is in general different from those of the other nodes because the thermal capacities of each of the points are different and because the thermal conductances between points within and on the boundary of the solid (excluding conductances to fluid temperature points) are not the same. Further, these capacities and conductances cannot be predicted by a recurrence relationship. However, for almost every network used in practice nodes with different heat capacities and/or different conductances occur, and these nodes can be considered in this sense asymmetric or irregular. Several examples are: (1) For the simple one-dimensional diffusion problem in Cartesian coordinates in a solid with uniform thermal properties, only for methods based on mesh $\Delta\xi/2$ are all the conductances (excluding those to fluid temperatures) and capacities equal; for method G based on mesh $\Delta\xi$ the two surface nodes at the boundaries have heat capacities that are different from the interior nodes, and further the heat capacities of these nodes cannot be found from a simple recurrence relationship. Hence, in the sense of heat capacities, method G is based on a regular

network with two irregular nodes at the surface. (2) For a one-dimensional problem in radial coordinates, for either spherical or cylindrical coordinates, approximated with a network of points spaced $\Delta\rho$ apart, the capacities and conductances are different for each node and, therefore, each node has a different equation, which is also the case of a general asymmetric network. However, for the regular radial network the capacities and conductances (and thus each equation) can be found from a recurrence relationship. (3) For an approximation based on a regular rectangular mesh spacing in a two-dimensional uniform solid with a curved boundary at a known temperature, the technique of linear interpolation mentioned previously can be shown to be equivalent to assuming that the points which have conductances to the surface nodes are asymmetric nodes and follow MacNeal's rules. These nodes then contribute to the truncation error quantities of the order of $\Delta\xi$ and $\Delta\eta$; while the other nodes contribute a truncation of the order of $(\Delta\xi)^2$ and $(\Delta\eta)^2$.

Thus, these examples show that the use of networks which contain irregular nodes, or for which the equations for each node are different, are common. Each of the irregular nodes in the above three examples can be shown to be consistent, however.

In order to study the accuracy of a general asymmetric network or the effect on accuracy of several irregular nodes in a regular network the classical method is used to study the truncation error and

propagation throughout the calculation. This is based on a method devised by Von Neumann and reported in most works on numerical solution of partial differential equations (2, 3, 4, 10, 17, 18, 21). The following study first defines the truncation error and then shows how this error affects the difference between the continuous and approximate solution. A detailed discussion of the space and time discretization errors is made. This includes a practical discussion of the meaning of consistency and indicates a possible method of making asymmetric nodes consistent. From these discussions some general rules can be formulated about when asymmetric nodes can be used and how they should be located. Some conclusions can also be made about the selection of the time differencing parameters. This investigation uses both analytic and numerical results. Since these are closely related, they are usually discussed together. The emphasis is primarily on the two-dimensional (space) problem, but many of the relationships also apply to one-dimensional problems, and some one-dimensional references are made. Unfortunately, the analytic and numerical studies of the truncation error and its propagation do not usually give as precise information about accuracy as does the comparison of the analytic expressions for the continuous and approximate solutions. However, it is hoped that eventually the results from the two types of error analysis may be related so that by examining the truncation error precise statements can be made

about the errors in eigenvectors, damping factors and initial vector. Consequently, a precise analysis, such as was made for a regular network for the one-dimensional problem, can be made from a knowledge of only the truncation errors.

In the following discussion the various shapes are associated with both networks and nodes, for example, a triangular network or hexagonal nodes. The shape associated with a network is the polygon formed by the branch lengths ℓ_{ij} when the points are interconnected. The shape associated with a node is the polygon formed by the perpendicular bisectors r_{ij} of the branch lengths and is the shape of the area used to calculate the heat capacity.

(B) DEFINITION OF DISCRETIZATION ERROR

A useful definition of the discretization or truncation error that is equivalent to the usual definition of this error can be developed by considering equation II-39 which defines the stepping out of the calculation. This vector equation, rearranged, is:

$$\frac{1}{\Delta\tau}(t_{n+1}-t_n) - \frac{Y}{A}[\gamma t_{n+1}+(1-\gamma)t_n] - \frac{Y_B}{A}[\gamma t_{B,n+1}+(1-\gamma)t_{B,n}] = 0 \qquad (V-3)$$

If the continuous temperature vector T_n is substituted for the approximate solution vector t_n, the right side of the equation is no longer equal to zero:

$$\frac{1}{\Delta\tau}(T_{n+1}-T_n) - \frac{Y}{A}[\gamma T_{n+1}+(1-\gamma)T_n] - \frac{Y_B}{A}[\gamma T_{B,n+1}+(1-\gamma)T_{B,n}] \equiv d_n \qquad (V-4)$$

This equation V-4 defines the discretization error vector d_n; the elements of this vector $d_{i,n}$ constitute the discretization error associated with the ith node. These components are

$$d_{i,n} \equiv \frac{T_{i,n+1}-T_i}{\Delta\tau} - \frac{1}{A_i}\sum_{j=1}^{R} y_{ij}[\gamma(T_{j,n+1}-T_{i,n+1})+(1-\gamma)(T_{j,n}-T_{i,n})] \qquad (V-5)$$

$$i = 1, 2, \ldots, S$$

As this definition of the discretization error vector is based on either the vector equation, or, equivalently, the component equation, it defines the discretization error for both the interior nodes and nodes which have conductances containing a heat-transfer coefficient to fluid temperature points outside the boundaries.

The reason for using this definition for the discretization error rather than the usual definition is that the effect of this discretization error vector upon the error in the approximate solution is easily derived. Further, for the interior points with no conductances to a fluid temperature the discretization error component $d_{i,n}$ is equivalent to the usual definition. This can be easily seen by writing the diffusion equation for the ith node at the nth and $(n+1)$ time increments in the following form,

$$\frac{\partial T_{i,n}}{\partial \tau} - \frac{1}{K_0(C_p\sigma)_i} \nabla' k_i \nabla' T_{i,n} = 0 \qquad \text{(V-6)}$$

$$\frac{\partial T_{i,n+1}}{\partial \tau} - \frac{1}{K_0(C_p\sigma)_i} \nabla' k \nabla' T_{i,n+1} = 0 \qquad \text{(V-7)}$$

and then subtracting an appropriate linear combination of these equations from equation V-5. The discretization error for the points with conductances across an interface to a fluid cannot be found by subtracting the diffusion equation, as those equations are actually approximations to both the diffusion equation and the boundary equations.

Since other works have not usually defined, discussed, or used the discretization error for nodes with conductances involving a heat-transfer coefficient, the above definition is not in conflict. In section D of this chapter the discretization error for both the interior points and points adjacent to fluid temperatures is discussed in more detail.

(C) PROPAGATION OF DISCRETIZATION ERROR

A difference equation for the error in the approximate solution can be derived from the above definition of the discretization error. The solution of this equation indicates how the discretization error affects the error in the approximate solution. The knowledge of how the discretization error affects the accuracy of the approximate solution should then allow a better practical understanding of section D, on how the selection of the differencing parameters affects the size of the discretization error.

In order to make the following derivation as general as possible equation V-3 is rewritten to include the effect of round-off error. This can be done by noting that if the approximate solution vectors t_{n+1} and t_n, as found by stepping out the solution, are substituted in equation V-3 the sum of vector-matrix products is not the zero vector, but is a vector e_n made up of small components $\epsilon_{i,n}$ caused by using only a finite number of decimals in the stepping calculation. This equation is, in practice, then

$$\frac{1}{\Delta\tau}(t_{n+1}-t_n)-\frac{Y}{A}\left[\gamma t_{n+1}+(1-\gamma)t_n\right]-\frac{Y_B}{A}\left[\gamma t_{B,n+1}+(1-\gamma)t_{B,n}\right]=\epsilon_n \qquad (V-8)$$

Now subtracting equation V-8 from equation V-4 and substituting v_n for the difference between solution vectors (T_n-t_n), the difference equation for the error in the approximate solution is

$$[I-\gamma\Delta\tau\frac{Y}{A}]v_{n+1}=[I+(1-\gamma)\Delta\tau\frac{Y}{A}]v_n+\Delta\tau\frac{Y_B}{A}[(v_{B,n+1})\gamma+(1-\gamma)(v_B)_n]$$

$$+\Delta\tau(d_n+e_n) \qquad \text{(V-9)}$$

or solving for v_{n+1}:

$$v_{n+1}=[I-\gamma\Delta\tau\frac{Y}{A}]^{-1}[(I+\{1-\gamma\}\Delta\tau\frac{Y}{A})v_n+\Delta\tau\frac{Y_B}{A}(\gamma v_{B,n+1}+\{1-\gamma\}v_{B,n})$$

$$+\Delta\tau(d_n+e_n)] \qquad \text{(V-10)}$$

where $v_{B,n}=T_{B,n}-t_{B,n}$.

The quantity $\Delta\tau(Y_B/A)(\gamma v_{B,n+1}+\{1-\gamma\}v_{B,n})$ is usually zero, as the approximate boundary temperature vector is almost always taken equal to the continuous temperature vector, $T_{B,n}$. However, if an element of $t_{B,n}$ is not equal to the corresponding element in $T_{B,n}$ the effect, as seen from equation V-10, is the same as changing the corresponding element in the discretization error vector. Assuming that no error is made in the boundary vector we have

$$v_{n+1}=[I-\gamma\Delta\tau\frac{Y}{A}]^{-1}[(I+\{1-\gamma\}\Delta\tau\frac{Y}{A})v_n+\Delta\tau(d_n+e_n)] \qquad \text{(V-11)}$$

The similarity between this difference equation and the difference equation II-39 that defines the calculation procedure should be noted. The major difference is that the forcing function vector $[\Delta\tau(Y_B/A)][\gamma t_{B,n+1}+(1-\gamma)t_{B,n}]$ which has non-zero elements only

for nodes with conductances to the fluid temperature is now replaced with the vector $\Delta\tau(d_n + e_n)$ which, in general, has all non-zero elements. Thus, equation V-11 can be considered to be a difference approximation to the partial differential equation of diffusion with a source term:

$$\frac{\partial V}{\partial \tau} = \frac{1}{K_0 C_p \sigma} \nabla' k \nabla' V + D(\xi, \eta, \tau) \qquad (V-12)$$

where $D(\xi, \eta, \tau)$ = known distributed source term expressed in degrees per dimensionless time corresponding to elements in $(d_n + e_n)$

$V(\xi, \eta, \tau)$ = continuous temperature function corresponding to elements in v_n.

This shows that the effect of the discretization and round-off error is that it acts equivalently to an energy source (or sink) distributed throughout the solid.

The solution of the difference equation V-11 for the error, as found by classical means, is:

$$v_n = CQ^n C^{-1} v_0 + \sum_{p=1}^{n} CQ^{n-p} C^{-1} (I - \gamma \Delta\tau \frac{Y}{A})^{-1} (d_{p-1} + e_{p-1}) \Delta\tau \qquad (V-13)$$

where $v_0 = T_0 - t_0$

and where the C and Q matrices are the eigenvector and eigenvalue matrices as defined in equations II-43, II-46, and II-47.

The solution for the error v_n can also be expressed in terms of the calculation matrix $[I-\gamma\Delta\tau(Y/A)]^{-1}[I+(1-\gamma)\Delta\tau(Y/A)]$ by using the eigenvector-eigenvalue matrix identity:

$$C\,Q^{n-p}\,C^{-1} = \left[I-\gamma\Delta\tau\,\frac{Y}{A}\right]^{p-n}\left[I+(1-\gamma)\Delta\tau\,\frac{Y}{A}\right]^{n-p} \qquad \text{(V-14)}$$

The solution equation V-13 has been checked in equation V-11 using the identity in equation V-14, and it is analogous to the solution for a system of ordinary differential equations with constant coefficients as given in Bellman (33), p. 169.

From the above equation, one sees that round-off error is not the cause of instability, but can only aggravate the basic instability caused by the eigenvalues q_j of the calculation matrix. At this point, for simplicity, the round-off error is assumed to be much smaller than the discretization error and it is neglected in the remainder of the discussion:

$$|e_{i,n}| << |d_{i,n}| \qquad i = 1, \ldots, S \qquad n \geq 0 \qquad \text{(V-15)}$$

For a thorough quantitative discussion of the effect of round-off error the references can be consulted (3, 10, 17). The main conclusion is, as given in Richtmyer (3), p. 25, that if the rounding is unbiased and if the quantities are kept in scale, round-off errors accumulate roughly in proportion to the square root of the number of increments n.

In most approximate solutions the stepping-out calculation is started with the continuous temperature vector and v_0 is zero. In this case the error in the approximate solution can be written as a weighted cumulative sum:

$$v_n = \sum_{p=1}^{n} C\,Q^{n-p}\,C^{-1}\left[I-\gamma\Delta\tau\,\frac{Y}{A}\right]^{-1} d_{p-1}\Delta\tau$$

$$= C\,Q^{n-1}\,C^{-1}\left[I-\gamma\Delta\tau\,\frac{Y}{A}\right]^{-1} d_0\Delta\tau + C\,Q^{n-2}\,C^{-1}\left(I-\gamma\Delta\tau\,\frac{Y}{A}\right) d_1\Delta\tau + \ldots$$

$$+ C\,Q^{n-p}\,C^{-1}\left[I-\gamma\Delta\tau\,\frac{Y}{A}\right]^{-1} d_{p-1}\Delta\tau + C\,Q^{n-p-1}\,C^{-1}\left(I-\gamma\Delta\tau\,\frac{Y}{A}\right)^{-1} d_p\Delta\tau + \ldots$$

$$+ C\,Q\,C^{-1}\left(I-\gamma\Delta\tau\,\frac{Y}{A}\right)^{-1} d_{n-2}\Delta\tau + \left(I-\gamma\Delta\tau\,\frac{Y}{A}\right)^{-1} d_{n-1}\Delta\tau \qquad \text{(V-16)}$$

where d_p is the discretization error in stepping from increment p to p+1.

As only stable solutions are considered here, equation V-16 shows that at time n only the discretization errors that occur for several time increments immediately previous to the nth increment are significant.

The above solution equation is a rigorous solution based on no estimates or series expansions. Unfortunately, unless the discretization error d_n is known as a function of time, the summation cannot be expressed in a simpler form; even if it were known as a function of time, in all probability it would be too complicated a function to obtain a simple expression for the summation. Thus the solution is

of limited usefulness. However, an averaged discretization error vector $\bar{d}_n$ can be defined such that the error vector v_n is given by the following equation.

$$v_n = \sum_{p=1}^{n} CQ^{n-p}C^{-1}(I-\gamma\Delta\tau\frac{Y}{A})^{-1} d_{p-1}\Delta\tau \equiv \left[\sum_{p=1}^{n} CQ^{n-p}C^{-1}\right](I-\gamma\Delta\tau\frac{Y}{A})^{-1}\Delta\tau\,\bar{d}_n \tag{V-17}$$

where $\bar{d}_n$ is a vector constant with time that gives the same error at time n, and where the summation $\left[\sum_{p=1}^{n} CQ^{n-p}C^{-1}\right]$ now is a matrix. This vector contains time-averaged components $\bar{d}_{i,n}$. By using the matrix identity in equation V-14, the matrix equations in Fadeeva (16), p. 62, for matrix series, together with other matrix manipulations, the product of the summation matrix and the other matrix is exactly:

$$\left[\sum_{p=1}^{n} C\,Q^{n-p}\,C^{-1}\right]\left[I-\gamma\Delta\tau\frac{Y}{A}\right]^{-1}\Delta\tau = -\left[I - C\,Q^{n}C^{-1}\right]Y^{-1}A \tag{V-18}$$

The solution becomes

$$v_n = -\left[I - C\,Q^{n}\,C^{-1}\right]Y^{-1}A\,\bar{d}_n$$

$$= \left[(I-\gamma\Delta\tau\frac{Y}{A})^{-n}\left(I+\left\{1-\gamma\right\}\Delta\tau\frac{Y}{A}\right)^{n} - I\right]Y^{-1}A\,\bar{d}_n \tag{V-19}$$

If the boundary forcing functions are such that a true steady state exists, the steady-state error is given by

$$v_{SS} = -Y^{-1}\,A\,d_{SS} \tag{V-20}$$

where d_{SS} is given by

$$Y\, T_{SS} + Y_B\, T_B = \Lambda\, d_{SS} \qquad \text{(V-21)}$$

In this case the steady-state discretization error vector d_{SS} is only a function of the number and location of the points, i.e., the Y and Y_B matrices, and obviously it cannot be a function of the time differencing parameters $\Delta\tau$ and γ.

A study of equation V-19 and its special steady-state case, equation V-20, leads to several important practical conclusions about the effect of the discretization error on the accuracy of the solution or smallness of the elements in v_n, even without a specific knowledge of the averaged discretization error vector $\bar{d}_n$. However, in order to relate the information in the next section about the actual discretization error d_n to the following discussion, several general properties of the averaged vector $\bar{d}_n$ should be mentioned. The averaged discretization error vector should for most problems be approximately equal to the discretization error for the several increments just previous to n and the nth increment. This can be seen from writing out the summations in equation V-17, which defines the average $\bar{d}_n$ and by remembering the damping or decaying nature of the Q^n matrix for stable solutions. Moreover, the elements of the actual discretization error vector d_n involve the second- and higher-order derivatives of the continuous temperature solution with space variables, and since

this temperature function is a smooth function for most problems, for a short period of time spanning only a few time increments, the discretization error vectors should not change much. Consequently, the averaged discretization error vector $\bar{d}_n$ can be considered to be about the same as the discretization error vectors that occur just previous to and during the nth increment, and comments made about this discretization error vector d_n in the next section should apply equally to the averaged vector $\bar{d}_n$. The averaged vector $\bar{d}_n$ is a function of time or of n. Further, in general, the elements of $\bar{d}_n$ or d_n are not the same size; however, in the following study the assumption is made that the elements in $\bar{d}_n$ are the same size. Conclusions derived on the basis of this assumption can be easily modified.

The matrix $[I-C\Omega^nC^{-1}]$ which multiplies the vector $(Y^{-1}A\,\bar{d}_n)$ to give v_n in equation V-19 can be taken as the identity matrix in this analysis. This is permitted because (1) it becomes essentially the identity matrix for sufficiently large n, because of the decaying nature of $(C\Omega^nC^{-1})$; and (2) only a very rough estimate of the relative elements in v_n is to be made. Thus, for values of n, where $[I-C\Omega^nC^{-1}]$ is only very approximately the identity matrix, it can still be assumed to be the identity matrix, as we are usually interested in estimating the position of the decimal point and possibly the first digit for the elements in v_n. For all but very short times it should be possible to obtain this rough estimate by assuming that

$$v_n \simeq - Y^{-1} A \bar{d}_n \qquad (V\text{-}22)$$

Note that v_n is still a function of time as $\bar{d}_n$ changes with time. To quantitatively analyze the accuracy for very short times when $[I\text{-}CQ^nC^{-1}]$ differs considerably from the identity matrix, the effect of this matrix operator would have to be found by numerical calculation. Some insight into the relative size of the elements of $[I\text{-}CQ^nC^{-1}]$ can be gained by noting that it appears in the approximate transient temperature solution for a solid with a zero initial condition.

$$t_n = [I\text{-}CQ^n C^{-1}] t_{P,n} \qquad (V\text{-}23)$$

Thus the conclusions about the manner in which the number and location of points affects the vector $(Y^{-1} A \bar{d}_n)$ can be applied to the error vector v_n for most problems and for all but small values of n.

The first and most obvious of these conclusions is that the vector $A \bar{d}_n$ is formed by the diagonal thermal capacity matrix A and $\bar{d}_n$ as components of $A_i \bar{d}_{i,n}$; that is, the time-averaged discretization error for the ith node is weighted by the thermal capacity of that node. Thus, for a fixed network of nodes, the discretization error of the relatively large nodes is magnified and that for the relatively small nodes has a smaller effect.

However, the above conclusion should not be interpreted to mean that merely by reducing the area of a few or all of the nodes can the approximate solution be made arbitrarily accurate; i.e., the

elements of the error vector v_n do not necessarily go to zero as the area goes to zero. This is because, to reduce the area of the nodes, the number of nodes S, which is the dimension of the square Y^{-1} and A matrices, must be increased proportionately; consequently, the elements in the vector $(Y^{-1}A\,\bar{d}_n)$ are now summations over more elements and the elements in this Y^{-1} matrix remain about the same size. This argument can be put on a more firm basis by considering the norm of the $Y^{-1}A$ matrix for one- and two-dimensional problems based on regular rectangular meshes. For these networks the heat capacities of the nodes are equal and assuming all elements of the time-averaged $\bar{d}_n$ vector are the same, an element of $(Y^{-1}A\,\bar{d}_n)$ is

$$(Y^{-1}A\,\bar{d}_n)_i = (A_i\,\bar{d}_{i,n}) \sum_{j=1}^{S} v_{ij} \qquad (V\text{-}24)$$

$$i = 1, \ldots, S$$

where $(Y^{-1}A\,\bar{d}_n)_i = i^{th}$ element of the vector $(Y^{-1}A\,\bar{d}_n)$

v_{ij} = the element in i^{th} row, j^{th} column, of Y^{-1} matrix.

For both the one-and two-dimensional networks above the elements v_{ij} in the inverse conductance matrix Y^{-1} are shown to be negative using results in the references (6, 29), i.e., Y^{-1} is a non-positive matrix. Consequently, $v_{ij} \leq 0$ for all i and j, and

$$-\sum_{j=1}^{S} v_{ij} = +\sum_{j=1}^{S} |v_{ij}| \qquad (V\text{-}25)$$

and the row norm defined in equation III-10, which is the maximum of the absolute values of the row sums, is a direct measure of the summation in equation V-24, and the product of the norm of Y^{-1} and A_i is a measure of the coefficient of $\bar{d}_{n,i}$.

$$(Y^{-1}A\,\bar{d}_n)_i \leq A_i \|Y^{-1}\|\,\|\bar{d}_{n,i}\| \qquad \text{(V-26)}$$

For the one-dimensional problem with both boundaries specified (problem IV, Table IV-9), approximated with method G, the inverse of the conductance matrix Y is known (29) and the norm-heat capacity product can be shown to be:

$$A_i \|Y^{-1}\| = (\Delta\xi)\|Y^{-1}\| = \frac{1}{8} \qquad \text{S-odd} \qquad \text{(V-27)}$$

$$A_i \|Y^{-1}\| = \frac{1}{8}\left[1 - (\Delta\xi)^2\right] \qquad \text{S-even} \qquad \text{(V-28)}$$

where S is the number of points for which the temperature is to be approximated as a function of time and where $\Delta\xi = \frac{1}{S+1}$. (Note that this S is not the same as would usually be used for method G in Chapter IV, but is S'-1 where S' is equivalent to S, as used in Chapter IV.) In Figure V-2 the product $\|Y^{-1}\|A_i$ and the number of variable temperature points are related to the area A_i based on equations V-27 and V-28. The extrapolation of $A_i\|Y^{-1}\|$ to an infinite number of points or zero nodal area A_i for a uniform unit square solid gives a finite, non-zero result of 1/8. The two-dimensional problem with a known boundary temperature has been studied numerically by calculating

the inverse conductance matrix Y^{-1} as the mesh is refined. In this calculation the network is made up of regular rectangles with constant spacing $\Delta\eta$ and $\Delta\xi$ with points located on the boundary surface. The area of each variable node is then

$$A_i = (\Delta\xi)(\Delta\eta) = \frac{1}{S + S_1 + S_2 + 1} \tag{V-29}$$

where $\Delta\xi = 1/(S_1+1)$

and $\Delta\eta = 1/(S_2+1)$ and S_1 and S_2 are the number of variable points in the ξ and η directions, respectively, and the conductance between nodes connected in the η direction is

$$y_{ij} = \frac{\Delta\xi}{\Delta\eta} \tag{V-30}$$

and in the ξ direction is

$$y_{ij} = \frac{\Delta\eta}{\Delta\xi} \tag{V-31}$$

The first mesh used contained nine variable points (S=9) with $\Delta\xi$ and $\Delta\eta$ equal to one-fourth (S_1 and S_2 equal to 3). The network was refined by first adding a row of points and then a column of points so that the second calculation was based on an S of 12 with an S_2 of 4 and an S_1 of 3, and the third on an S of 16 with both S_1 and S_2 equal to 4. The last network studied contained 64 points, 8 in each row and 8 in each column of the network. The numerical results for $A_i\|Y^{-1}\|$ are related to the area A_i and to the number of variable points S in Figure V-3. These results are somewhat like those for the one-dimensional

problem in that for selection of S resulting in a perfect square ($\Delta\xi$ and $\Delta\eta$ equal) two different lines can be drawn for S odd and S even. These lines are approximately linear and can be extrapolated to a constant value, at zero A_i or infinite S, of about 0.074. The products $A_i \| Y^{-1} \|$ for meshes with unequal $\Delta\xi$ and $\Delta\eta$ lie within the two lines.

From these two studies of the approximations to one- and two-dimensional problems for regular networks, we conclude that the components of the averaged discretization error must be arbitrarily small for the components of the error vector v_n to be made arbitrarily small for a stable solution. This is true because, as the regular network is refined, the coefficients, as measured by $A_i \| Y^{-1} \|$, of the time-averaged discretization error vector $\bar{d}_n$ do not go to zero as the nodal area A_i is reduced to zero. The above conclusion can also be seen to be equivalent to Richtmyer's consistency requirement, as, in order to obtain an arbitrarily accurate approximate solution, the solution must be both stable, $|q| < 1$, and the norm $\|\bar{d}_n\|$ of the discretization error vector must go to zero as the mesh is refined to make the norm $\|v_n\|$ of the error vector go to zero.

The above conclusion has been shown here only for regular networks, where the inverse conductance matrix Y^{-1} has been shown to be a non-positive matrix, for certain methods of refining the network, and for a problem with a specified boundary temperature. However,

since that conclusion is equivalent to the stability-consistency results of Richtmyer, a like conclusion can probably also be made for the asymmetric network. This conclusion is that the norm of the inverse conductance-capacitance matrix, $\|Y^{-1}A\|$, does not go to zero as all nodal areas go to zero. There is no reason to assume that for the irregular location any difference in behavior should occur; indeed the inverse conductance matrices Y^{-1} for four asymmetric networks shown in Tables V-1, V-2, V-3, and V-4 show that this matrix is a non-positive matrix, as the inverse conductance matrices must be for all problems with regular rectangular networks. Although no proof has been attempted that the inverse conductance matrix Y^{-1} for an asymmetric network is a non-positive matrix, the Y matrix has properties like those used in the proof for the matrices for rectangular networks (6).

Of more interest is how an asymmetric network is refined. Although this topic is discussed in more detail in the next section, one might consider adding more nodes in a certain vicinity of an asymmetric network leaving the rest of the network unchanged. These points would be located according to MacNeal's rules. The effect on the accuracy of the approximate solution for the nodes in the unchanged network probably would be insignificant unless the discretization error was also reduced. This is true, even though the area of some of the nodes is reduced, because there are now more nodes and each of the

new nodes probably has a weighting in the inverse conductance matrix Y^{-1} of about the same magnitude as the large nodes had before. Thus the sum of the weighting of the discretization error for this vicinity is not reduced. The error in the approximate solution for the new smaller nodes also probably is not improved unless the components of $\bar{d}_n$ are smaller.

The elements of the inverse conductance matrix Y^{-1} are the weightings for the discretization error-area products. That is, the element in the ith row and j^{th} column, v_{ij}, of the Y^{-1} matrix is the weighting given to the $A_j \bar{d}_{jn}$ product in the equation for the error for the ith node $v_{i,n}$. Because of the symmetry of the Y^{-1} matrix this coefficient v_{ij} is also v_{ji} and is the weighting of the product $A_i d_{i,n}$ in the equation for the error of the jth node. Consequently, conclusions made about the effect of the $A_j \bar{d}_{jn}$ on the ith node also apply to the effect of the $A_i \bar{d}_{i,n}$ on the jth node. These effects have been studied by numerically calculating the inverse conductance matrices, and correlating the size of the elements with the position of the nodes in the networks.

The inverse conductance matrix Y^{-1} is shown on Tables V-1, V-2, V-3, and V-4, for four asymmetric networks together with the Y matrix and a sketch of the network of points. Also shown are the matrices for two regular rectangular networks for 9 points. For the asymmetric networks the 9 x 9 matrix is for the numerical example

used in Chapter III and has a specified temperature along the horizontal boundary. The 8 x 8 matrices are for the same geometry but the hypotenuse boundary now also has a specified temperature. The shapes I and II are from solid-propellant curing problem used in the original Longwell (9) reference. The regular rectangular networks are for a square and a rectangular problem with specified boundary conditions.

A study of the inverse conductance matrix Y^{-1}, the conductance matrix Y, and the sketch allows the following conclusions to be made about the effect on the ith error $v_{i,n}$ of the components of the discretization error heat capacity vector $(A_i \bar{d}_{i,n})$ based on equation V-22, for both regular and asymmetric networks.

(1) The diagonal element v_{ii} of the inverse conductance matrix has a larger absolute value than the off-diagonal elements in the ith row or i^{th} column. Consequently the product $(A_i \bar{d}_{i,n})$ or the area-weighted discretization error associated with the ith node is weighted more in the error for the ith node, $v_{i,n}$, than any other area-weighted discretization error product. This property has been demonstrated to hold for rectangular networks, and inverse conductance matrices obtained for the asymmetric network indicate that it holds there also.

(2) The area products $(A_j \bar{d}_{j,n})$ for those nodes which have conductances to the ith node (non-zero elements in Y matrix, $y_{ij} \neq 0$) are usually weighted the next in the equation for the error at node i. The weighting is larger for those with the larger conductances to the ith node.

(3) The weighting for nodes that can be considered to have a high equivalent thermal resistance between them (usually nodes distant from each other) are small and the discretization error products for these nodes influence the error at i negligibly.

The conclusions about the relative effect of the area-weighted discretization error vector have been checked numerically using the asymmetric network shown in Figure III-1 to solve for the steady-state condition where the horizontal surface is at unity, the vertical boundary is adiabatic, and the hypotenuse is at a temperature of $\frac{1}{2}$. For the steady-state problem equation V-22 gives the error exactly and is not an approximation. This is equivalent to a steady-state problem for the unit square with its horizontal boundaries at one and its vertical boundaries at zero. The continuous solution as derived by suitable coordinate transformations and combinations from a solution in Carslaw and Jaeger (1) is:

$$T_{SS}(\xi,\eta) = \frac{4}{\pi}\sum_{j=0}^{\infty} \frac{[\sin(2j+1)\pi(\xi+0.5)]\,[\cosh(2j+1)\pi\eta]}{(2j+1)\,[\cosh(\frac{2j+1}{2})\pi]} \qquad (V-32)$$

The approximate solution was computed from

$$t_{SS} = -Y^{-1}Y_B t_B \qquad (V-33)$$

where the Y matrix is modified from that used in the transient problem in Chapter III, by assuming anti-symmetry of the nodes about the hypotenuse. This is equivalent to taking nodes on the hypotenuse with

a temperature of one half or to assuming that the temperature difference between the temperature of anti-symmetric nodes opposite 1, 3, 5, 7, and 8, and $\frac{1}{2}$ is equal to the temperature difference between the nodes on the hypotenuse ($\frac{1}{2}$) and nodes 1 3, 5, 7, and 8. Suitable modifications were made in the Y_B matrix. Note that the temperature at node 9 is now specified. Three approximate solutions were calculated using temperatures at node A of $\frac{1}{2}$, $\frac{3}{4}$, and 1. An argument for using each of these temperatures at A is that in the continuous problem the temperature at node A can be considered indeterminate. However, the continuous solution gives one half at this point; therefore, use of the temperatures of $\frac{3}{4}$ and 1 can be considered to be the equivalent of changing the area-weighted discretization error at point 1 (see equation V-10).

A summary of these calculations is in Table V-7. These show that increasing this boundary temperature from $\frac{1}{2}$ to 1 changes the area-weighted discretization error for node 1 from -0.0468 to 0.0740, while the other discretization errors remain constant. This change in discretization error changes the error at node 1 from 0.0067 to -0.0107, and also changes the error significantly at its neighbor nodes 2 and 3. However, no significant change is observed in nodes located at a distance, for example, nodes 6, 7, and 8. Thus the conclusions about the relative influence of the discretization error are confirmed. The accuracy of the approximate solution for the steady-state solution

is discussed in more detail later; however, it can be seen that the asymmetric network does give satisfactory results for this problem.

1. Conclusions

The important conclusions based on equation V-19 for the propagation of the discretization error which should be remembered in the following discussion on the discretization error and in locating points are:

(1) The error of an approximate solution cannot be made small merely by adding more points and reducing the area unless the discretization errors are decreased by the addition of the points. An important corollary to this conclusion is that the size of the discretization errors directly affects the error in the approximate solution, and one wishes to locate the points in such a manner that the discretization errors are small, particularly in a region of interest for the specific geometry, for the specific boundary conditions and forcing functions.

(2) For a fixed network of points the discretization error at a node is weighted by the heat capacity of that node; thus, the discretization error for a relatively large node is weighted more than that for a small node.

(3) The discretization error weighted with the nodal heat capacity ($A_i \bar{d}_{i,n}$) associated with the ith node affects the accuracy of the solution at the ith node more than it affects the accuracy of the

adjoining nodes. It affects the error at these adjoining nodes more than those located some distance away which have a high equivalent thermal resistance to the ith node. However, the sum of the discretization errors for many nodes on the ith node, even though each of these nodes does have a high equivalent resistance to the ith node, can give a significant contribution to the error at i.

(D) DISCRETIZATION ERROR

In this section series expansions are developed for the discretization error associated with the selection of the time discretization parameters of γ and $\Delta\tau$, and the space discretization error associated with the number and location of points. On the basis of these expansions several conclusions can be made about the accuracy of asymmetric networks, ways of improving the accuracy, and about the concept of consistency. Some practical conclusions also can be made about the selection of the differencing parameters.

1. Division of Discretization Error with Time and Space Differencing Parameters

In most discussions of the discretization error the total discretization error, as defined by equations V-4 or V-5, is divided into two parts. One part is associated with the time differencing parameters γ and $\Delta\tau$, and the other part with the space differencing parameters. This can be seen to be a logical procedure as the approximate formulation in Chapter II, section C, treated the two discretizations separately. However, the precise way in which this division is made affects both the size of the discretization associated with time and space and the series expansion for both discretizations. However, despite the fact that the size of the discretization errors associated

with time and space are affected by the precise division, the total discretization error which is the sum of the time and space errors is the same as defined by equation V-3.

Indeed, for the one-dimensional problem with a regular mesh, Richtmyer shows one series expansion for the total discretization error that can be derived by not dividing the discretization error. That expression is

$$d_{m,n} = \frac{\partial^4 T_{m,n}}{\partial \xi^4} \left[\Delta\tau \left(\frac{1}{2} - \gamma \right) - \frac{1}{12} (\Delta\xi)^2 \right] + \cdots \qquad \text{(V-34)}$$

$$\tfrac{1}{2} < m < S - \tfrac{1}{2}$$

where m is for an interior point.

Note that if the coefficient of $\partial^4 T_{m,n}/\partial \xi^4$ is set equal to zero equation IV-262 for γ_o results. Equation V-34 can be derived by writing equation V-5 in terms of m and $\Delta\xi$ for the one-dimensional problem and by subtracting the diffusion equation at point m at time n, which is

$$\frac{\partial T_{m,n}}{\partial \tau} - \frac{\partial^2 T_{m,n}}{\partial \xi^2} = 0 \qquad \text{(V-35)}$$

This gives

$$\frac{T_{m,n+1} - T_{m,n}}{\Delta\tau} - \frac{\partial T_{m,n}}{\partial \tau}$$

$$- \frac{1}{(\Delta\xi)^2} \left[(T_{m-1,n+1} - 2T_{m,n+1} + T_{m+1,n+1})\gamma + (T_{m-1,n} - 2T_{m,n} + T_{m+1,n})(1-\gamma) \right]$$

$$+ \frac{\partial^2 T_{m,n}}{\partial \xi^2} = d_{m,n} \qquad \text{(V-36)}$$

Then the quantities $(T_{m,n+1})$, $(T_{m+1,n+1})$, $(T_{m-1,n+1})$, $(T_{m+1,n})$, and $(T_{m-1,n})$ are written in terms of $T_{m,n}$ using a two-variable Taylor series expansion in $\Delta\xi$ and $\Delta\tau$ which, together with derivative identities based on the diffusion equation, gives equation V-34.

Although the development of a single series such as equation V-34 for the two-dimensional problem with a regular mesh would be desirable, unfortunately this would require a Taylor series expansion in three variables, $\Delta\tau$, $\Delta\xi$, and $\Delta\eta$. Even for a regular mesh, such an expansion would be very cumbersome, both to derive and to reduce to understandable expressions. Thus, the discretization error for the interior points is compared to the diffusion equation at times n and (n+1) weighted by γ and $(1-\gamma)$, respectively. Subtracting $(1-\gamma)$ times equation V-6 and γ times equation V-7 from equation V-5, there is obtained:

$$d_{i,n} = \left[\frac{T_{i,n+1} - T_{i,n}}{\Delta\tau} - \left\{ \gamma \frac{\partial T_{i,n+1}}{\partial\tau} + (1-\gamma) \frac{\partial T_{i,n}}{\partial\tau} \right\} \right]$$

$$- \gamma \left[\frac{1}{A_i} \sum_{j=1}^{l} y_{ij} (T_{j,n+1} - T_{i,n+1}) - \frac{1}{K_0 (C_p \sigma)_i} \nabla' k \nabla' T_{i,n+1} \right]$$

$$- (1-\gamma) \left[\frac{1}{A_i} \sum_{j=1}^{S} y_{ij} (T_{j,n} - T_{i,n}) - \frac{1}{K_0 (C_p \sigma)_i} \nabla' k \nabla' T_{i,n} \right] \qquad \text{(V-37)}$$

Now the discretization error associated with the time parameters can be defined as:

$$\varphi_{i,n} \equiv \frac{T_{i,n+1}-T_{i,n}}{\Delta\tau} - \left\{\gamma \frac{\partial T_{i,n+1}}{\partial \tau} + (1-\gamma)\frac{\partial T_{i,n}}{\partial \tau}\right\} \qquad \text{(V-38)}$$

and the space discretization error as:

$$\sigma_{i,n} = \frac{1}{A_i}\sum_{j=1}^{l} y_{ij}(T_{j,n}-T_{i,n}) - \frac{1}{K_0(C_p\sigma)_i}\nabla' k_i \nabla' T_{i,n} \qquad \text{(V-39)}$$

and

$$\sigma_{i,n+1} = \frac{1}{A_i}\sum_{j=1}^{R} y_{ij}(T_{j,n+1}-T_{i,n+1}) - \frac{1}{K_0(C_p\sigma)_i}\nabla' k_i \nabla' T_{i,n+1} \qquad \text{(V-40)}$$

Thus the total discretization error is:

$$d_{i,n+1} = \varphi_{i,n} - [\gamma\sigma_{i,n+1} + (1-\gamma)\sigma_{i,n}] \qquad \text{(V-41)}$$

This division of the discretization error has the disadvantage that the total contribution from the space error, $[(1-\gamma)\sigma_{i,n} + \gamma\sigma_{i,n+1}]$, is a function of the time differencing parameters γ and $\Delta\tau$, but it has the advantage that an expansion for $\sigma_{i,n}$ or $\sigma_{i,n+1}$ is only a two-variable series, and thus is significantly simpler than the three-variable expansion. This definition of the space discretization error also gives the usual definition for the steady-state problem.

Equations V-39 and V-40 apply for points or nodes which have conductances only to other nodes within or on the solid boundary, that

is, not to a fluid node through a finite non-zero heat-transfer coefficient. The difference equations for nodes, whether located on the boundary, as in method G, or away from the boundary, as in method C, that have a heat-transfer coefficient conductance to a fluid node, are an approximation to both the diffusion equation and to the boundary equation:

$$H[T_f(\tau) - T(\xi_B, \eta_B, \tau)] = - \frac{k \partial T}{\partial \eta} (\xi_B, \eta_B, \tau) \qquad (V\text{-}42)$$

Therefore, the discretization error of a node adjacent to a fluid node should be compared to the above equation in addition to the diffusion equation. However, no published discussion of this type of comparison has been found, and even for method C and the one-dimensional problem no useful expression or expansion has been derived for the discretization error. Very probably the difficulty in developing a meaningful expression for the discretization error at such an adjacent node is that, when h is finite, the series expansions in effect cross the solid boundary or a discontinuity.

For the special cases of zero or infinite heat-transfer coefficient, the adjacent nodes can be treated as interior nodes. For the zero heat-transfer coefficient condition the temperature distribution can be considered as continuous and symmetric about the boundary; for the infinite heat-transfer coefficient condition, where the surface is at a constant specific temperature, the temperature distribution can be considered as continuous and anti-symmetric about the boundary. To

consider the adjacent node as an interior node in these circumstances, additional points are located across the boundary such that if the boundary has no flux across it, the temperature at each of the additional points would be the same as at a corresponding point in the interior of the solid; for the infinite heat-transfer coefficient, the temperature difference between the additional point and the constant surface temperature would be equal to the temperature difference between the surface and the corresponding point within the solid. The adjacent node then can be treated as an interior node with neighbors corresponding to the original points within the solid and the necessary initial points as neighbors.

The above procedure, although valid in principle for curved boundaries, is useful only for straight-line boundaries. This is true because only for straight-line boundaries can the location of the additional temperature points be found easily, by reflecting the network about the boundary. Further, this technique for adjacent nodes under these conditions should only be considered as a temporary method of studying their discretization errors because eventually an expansion should be developed which would contain the heat-transfer coefficient as a parameter, and which would give results for zero and infinite coefficients as special cases. Moreover, it should also be noted that the series expansions for the errors as developed above would be identical for infinite and zero h, and that the difference in discretization

error for the two different conditions would be caused in these expansions by widely different values of the derivatives under the two conditions.

The adjacent nodes for methods G and C for the one-dimensional problem can be studied using equations V-39 and V-40 for constant surface temperature or adiabatic conditions. For both methods, under these conditions, the discretization error for all nodes including adjacent nodes is given by equation V-34 and the relationship for γ_o, as found from the damping expansions or the discretization error, is an optimum value; however, for the problem with a finite heat-transfer coefficient equation V-34 still applies for the interior points, but apparently the discretization error for the node adjacent to the fluid is of a different form so that γ_o is no longer an optimum value. However, from the complete solution we know that the actual error $v_{m,n}$ for all points is of the order of $1/S^2$ and of $\Delta\tau$; this fact together with a consideration of equation V-19 allows us to conclude that the discretization error for the adjacent nodes in methods G and C is of the order of $1/S^2$ and $\Delta\tau$. One of the most interesting and important future studies suggested by this work is that of finding an expansion for the discretization error for nodes adjacent to a fluid with finite heat-transfer conductances.

2. Time Discretization Error

A series expansion for the time discretization error can be derived by developing a series expansion for $(T_{i,n+1}-T_{i,n})/\Delta\tau$ about both n and (n+1). Then, using the weighting factor γ on the expansion about (n+1) and $(1-\gamma)$ on that about n, there results:

$$\varphi_{i,n}=\frac{T_{i,n+1}-T_{i,n}}{\Delta\tau}-\left[\gamma\frac{\partial T_{i,n+1}}{\partial\tau}+(1-\gamma)\frac{\partial T_{i,n}}{\partial\tau}\right]=\frac{\Delta\tau}{2}\left[(1-\gamma)\frac{\partial^2 T_{i,n}}{\partial\tau^2}-\gamma\frac{\partial^2 T_{i,n+1}}{\partial\tau^2}\right]$$

$$+\frac{(\Delta\tau)^2}{6}\left[(1-\gamma)\frac{\partial^3 T_{i,n}}{\partial\tau^3}+\gamma\frac{\partial^3 T_{i,n+1}}{\partial\tau^3}\right]$$

$$+\frac{(\Delta\tau)^3}{24}\left[(1-\gamma)\frac{\partial^4 T_{i,n}}{\partial\tau^4}-\gamma\frac{\partial^4 T_{i,n+1}}{\partial\tau^4}\right]+\cdots \qquad \text{(V-43)}$$

Thus, the contribution from the time discretization is about proportional to $\Delta\tau$. The effect of the weighting can be seen by considering that the second time derivatives evaluated at n and (n+1) are about equal which would mean that the coefficient of the time increment term is:

$$\varphi_{i,n}\simeq\frac{\Delta\tau}{2}(1-2\gamma)\frac{\partial^2 T_{i,n}}{\partial\tau^2}+\cdots \qquad \text{(V-44)}$$

where the coefficients of the succeeding derivatives at (n+1) would be changed slightly. Although this expression indicates that the optimum γ is $\frac{1}{2}$, it must be remembered that $\varphi_{i,n}$ is only part of the

contribution to the total discretization error $d_{i,n}$. When this is considered for regular networks the best weighting γ is usually less than $\frac{1}{2}$ and the γ of $\frac{1}{2}$ gives a difference solution very close to that of the analog solution. The comparison of approximate solutions for the transient problem in Chapter III indicates that much the same conclusions about the effect of the time increment and γ on accuracy made for the one-dimensional regular network problems in Chapter IV also apply to the two-dimensional problem with an asymmetric network. This would be expected, as the time discretization error appears in the same form for both the one- and two-dimensional problems. These results are discussed in section F in this chapter.

(E) SPACE DISCRETIZATION ERROR

Since the total discretization error is important, and since the contribution from the time discretization error appears to be satisfactory, the success or failure of an asymmetric network depends on the size of the contribution from the space discretization error, $-[\sigma_{i,n}(1-\gamma) + \sigma_{i,n+1}\gamma]$. The main problem with using an asymmetric location of the points is shown to be in the quantities, σ_i's. The following develops a series expansion for the space discretization error at time n, for points within and on the surface of a uniform solid but which have no conductances to a fluid. This is then followed by a summary of expansions for $\upsilon_{i,n}$ for several types of nodes, and a discussion of consistency and practical implications of the space discretization error is included.

1. Expansion for Space Discretization Error, Interior Node.

Consider a general asymmetric node for a two-dimensional problem within a solid with constant thermal properties having a thermometric conductivity of K_0, with continuous temperature distribution in the vicinity of the node. Such a node is shown in Figure V-4. The neighbors for the ith node are assumed to have subscripts numbered counter-clockwise. Moreover, a dimensionless Cartesian coordinate system is superimposed on the system. In order to formulate the series

expansion in the dimensionless coordinates, the dimensionless lengths λ_{ij} and ρ_{ij} are defined according to:

$$\lambda_{ij} = \frac{\ell_{ij}}{L} \tag{V-45}$$

and

$$\rho_{ij} = \frac{r_{ij}}{L} \tag{V-46}$$

where λ_{ij} = dimensionless distance between nodes i and j

ρ_{ij} = dimensionless length of perpendicular bisector of line connecting nodes i and j

The node is assumed to have a total of N neighbors, although 5 are shown in the figure. The specification of the λ_{ij} and ρ_{ij} serves to completely define the geometry of the nodes. However, for convenience in making the expansion the angle β_{ij} between the positive ξ axis and the leg λ_{ij} is introduced. The λ_{ij}'s and β_{ij}'s are equivalent to radial coordinates of neighbor nodes. The angle parameter β_{ij} is not an independent variable and from the geometry the following relationship between ρ_{ij}, λ_{ij}, and β_{ij} can be derived.

$$\rho_{ij} = \frac{\lambda_{i,j+1}\sin(\beta_{i,j+1}-\beta_{ij}) - \lambda_{ij}\sin(\beta_{i,j+1}-\beta_{i,j-1}) + \lambda_{i,j-1}\sin(\beta_{ij}-\beta_{i,j-1})}{[\sin(\beta_{i,j+1}-\beta_{ij})][\sin(\beta_{ij}-\beta_{i,j-1})]} \tag{V-47}$$

where j has a geometrical meaning because of the counter-clockwise numbering. When j is 1, the subscript (j-1) means N and similarly when j is N, the subscript (j+1) is 1. With these substitutions equation V-39 for $\sigma_{i,n}$ is:

$$\sigma_{i,n} = \left[\frac{4}{\sum_{j=1}^{N} \rho_{ij}\,\lambda_{ij}}\right]\left[\sum_{j-1}^{N}\left(\frac{\rho_{ij}}{\lambda_{ij}}\right)\,(T_{j,n}-T_{i,n})\right]-\left[\frac{\partial^2 T_{i,n}}{\partial\xi^2}+\frac{\partial^2 T_{i,n}}{\partial\eta^2}\right] \qquad (V\text{-}48)$$

where N = number of neighboring nodes

i = subscripts of nodes with no conductance to fluid temperatures

j = refers to subscripts of neighbor nodes with non-zero conductance to the ith node, in these equations

$$\rho_{ij}/\lambda_{ij} = y_{ij} = r_{ij}/\ell_{ij}$$

$$A_i = \sum_{j=1}^{N} \rho_{ij}\,\lambda_{ij} = \sum_{j=1}^{N} r_{ij}\,\ell_{ij}/L^2$$

The first step in developing an expansion for $\sigma_{i,n}$ is to expand the quantity (T_j-T_i) using the Taylor series. This is, symbolically,

$$T_{j,n}-T_{i,n} = \sum_{p=1}^{\infty} \frac{1}{p!}\left[\xi_j\,\frac{\partial}{\partial\xi} + \eta_j\,\frac{\partial}{\partial\eta}\right]^p T_{i,n} \qquad (V\text{-}49)$$

where ξ_j and η_j are the coordinates of the jth neighbor based on the coordinate system with the origin at node i. These coordinates are, from the geometry,

$$\xi_j = \lambda_{ij} \cos\beta_{ij} \tag{V-50}$$

$$\eta_j = \lambda_{ij} \sin\beta_{ij} \tag{V-51}$$

and the expansion becomes, after being weighted by the conductance,

$$y_{ij}(T_{j,n}-T_{i,n}) = \frac{\rho_{ij}}{\lambda_{ij}} \sum_{p=1}^{\infty} \frac{1}{p!} \left[\lambda_{ij}(\cos\beta_{ij} \frac{\partial}{\partial\xi} + \sin\beta_{ij} \frac{\partial}{\partial\eta})\right]^p T_{i,n} \tag{V-52}$$

The first several terms of this series are in Table V-8.

The summation of these terms over all neighboring points must now be found.

$$\sum_{j=1}^{N} y_{ij}(T_{j,n}-T_{i,n}) = \sum_{j=1}^{N} \frac{\rho_{ij}}{\lambda_{ij}} \sum_{p=1}^{\infty} \frac{\left[\lambda_{ij}(\cos\beta_{ij} \frac{\partial}{\partial\xi} + \sin\beta_{ij} \frac{\partial}{\partial\eta})\right]^p T_{i,n}}{p!} \tag{V-53}$$

This is best studied by interchanging the summation signs and considering the finite summations on j for the first several p's. MacNeal has shown that for p of 1 the summation on j is zero.

$$\sum_{j=1}^{N} \rho_{ij}\left[\cos\beta_{ij}\ \frac{\partial T_{j,n}}{\partial \xi} + \sin\beta_{ij}\ \frac{\partial T_{i,n}}{\partial \eta}\right]$$

$$= \frac{\partial T_{i,n}}{\partial \xi} \sum_{j=1}^{N} \rho_{ij}\cos\beta_{ij} + \frac{\partial T_{i,n}}{\partial \eta} \sum_{j=1}^{N} \rho_{ij}\sin\beta_{ij}$$

$$= 0 \qquad \text{(V-54)}$$

The argument is that the summations which are the coefficient of the first derivatives in equation V-54 are zero because the ρ_{ij} described a closed polygon. For a p of 2, noting that the derivatives can be removed from the summation, and after a simple algebraic manipulation and rearrangement, the summation on j gives:

$$\sum_{j=1\ p=2}^{N} \equiv$$

$$\frac{1}{2}\sum_{j=1}^{N} \rho_{ij}\lambda_{ij}\left[(\cos\beta_{ij})^2\ \frac{\partial^2 T_{i,n}}{\partial \xi^2} + 2\cos\beta_{ij}\sin\beta_{ij}\ \frac{\partial^2 T_{i,n}}{\partial \xi\,\partial \eta} + (\sin\beta_{ij})^2\ \frac{\partial^2 T_{i,n}}{\partial \eta^2}\right]$$

$$= \left[\frac{\partial^2 T_{i,n}}{\partial \xi^2} + \frac{\partial^2 T_{i,n}}{\partial \eta^2}\right]\left[\frac{\sum_{j=1}^{N} \rho_{ij}\lambda_{ij}\left\{(\cos\beta_{ij})^2 + (\sin\beta_{ij})^2\right\}}{4}\right]$$

$$+ \left[\frac{\partial^2 T_{i,n}}{\partial \xi^2} - \frac{\partial^2 T_{i,n}}{\partial \eta^2}\right]\left[\frac{\sum_{j=1}^{N} \rho_{ij}\lambda_{ij}\left\{(\cos\beta_{ij})^2 - (\sin\beta_{ij})^2\right\}}{4}\right]$$

$$+\left[\frac{\partial^2 T_{i,n}}{\partial\xi\,\partial\eta}\right]\left[\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}\cos\beta_{ij}\sin\beta_{ij}\right] \qquad (V\text{-}55)$$

Now using trigonometric identities the summation for p of 2 can be written as

$$\sum_{j=1}^{N}\sum_{p=2} = \left[\frac{\partial^2 T_{i,n}}{\partial\xi^2} + \frac{\partial^2 T_{i,n}}{\partial\eta}\right]\left[\frac{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}{4}\right]$$

$$+\left[\frac{\partial^2 T_{i,n}}{\partial\xi^2} - \frac{\partial^2 T_{i,n}}{\partial\eta^2}\right]\left[\frac{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}\cos 2\beta_{ij}}{4}\right] \qquad (V\text{-}56)$$

$$+\left[\frac{\partial^2 T_{i,n}}{\partial\xi\,\partial\eta}\right]\left[\frac{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}\sin 2\beta_{ij}}{2}\right]$$

And in the same way the summations for p's of 3 and greater can be obtained. The expansion of the space discretization error $\sigma_{i,n}$ is then found by dividing each of the sums of each of these summations on j by the dimensionless heat capacity A_i or $\sum_{j=1}^{N}(\rho_{ij}\lambda_{ij}/4)$ and subtracting the Laplacian operator evaluated at point i and n, $[(\partial^2 T_{i,n}/\partial\xi^2) + (\partial^2 T_{i,n}/\partial\eta^2)]$. The first terms of the expansion for $\sigma_{i,n}$ are:

$$\sigma_{i,n}=\left[\frac{\partial^2 T_{i,n}}{\partial\xi^2}-\frac{\partial^2 T_{i,n}}{\partial\eta^2}\right]\left[\frac{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}\cos 2\beta_{ij}}{\sum_{j-1}^{N}\rho_{ij}\lambda_{ij}}\right]+$$

$$+\left[\frac{\partial^2 T_{i,n}}{\partial\xi\,\partial\eta}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}\sin 2\beta_{ij}}{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right]+$$

$$+\left[\frac{\partial^3 T_{i,n}}{\partial\xi^3}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^2(\cos\beta_{ij})^3}{3\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right]+$$

$$+\left[\frac{\partial^3 T_{i,n}}{\partial\xi^2\partial\eta}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^2(\cos\beta_{ij})^2(\sin\beta_{ij})}{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right]+$$

$$+\left[\frac{\partial^3 T_{i,n}}{\partial\xi\,\partial\eta^2}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^2(\cos\beta_{ij})(\sin\beta_{ij})^2}{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right]+$$

$$+\left[\frac{\partial^3 T_{i,n}}{\partial\eta^3}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^2(\sin\beta_{ij})^3}{3\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right] +$$

$$+\left[\frac{\partial^4 T_{i,n}}{\partial\xi^4}\right]\left[\frac{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^3(\cos\beta_{ij})^4}{6\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right] +$$

$$+\left[\frac{\partial^4 T_{i,n}}{\partial\xi^3\partial\eta}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^3(\cos\beta_{ij})^3(\sin\beta_{ij})}{3\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right] +$$

$$+\left[\frac{\partial^4 T_{i,n}}{\partial\xi^2\partial\eta^2}\right]\left[\frac{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^3(\cos\beta_{ij})^2(\sin\beta_{ij})^2}{\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}}\right] +$$

$$+\left[\frac{\partial^4 T_{i,n}}{\partial\xi\partial\eta^3}\right]\left[\frac{2\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^3(\cos\beta_{ij})(\sin\beta_{ij})^3}{3\sum_{j=1}^{N}\rho_{ij}\lambda_{ij}^3}\right] +$$

$$+\left[\frac{\partial^4 T_{i,n}}{\partial \eta^4}\right]\left[\frac{\sum_{j=1}^{N} \rho_{ij}\lambda_{ij}^3(\sin\beta_{ij})^4}{6\sum_{j=1}^{N} \rho_{ij}\lambda_{ij}}\right] + \cdots \qquad \text{(V-57)}$$

In the above equation no attempt has been made to use trigonometric identities or other manipulations to simplify the terms corresponding to p's of 3 and 4. Geometric relationships such as equation V-47 could be used to eliminate one of the quantities λ, ρ, or β, from the above equation but such a substitution is too cumbersome to be useful. Although the above equation was based upon a coordinate system with its origin at the center of the node, this coordinate system serves mainly to describe the geometric parameters λ_{ij} and β_{ij}, and the expansion above is independent of the origin of the coordinate system. The derivatives and their coefficients in the expansion are a function of the relative direction of the ξ or η axis, although the space discretization error as defined by equation V-39 obviously has no such dependence.

Despite the complexity of the expansion for the space discretization error equation V-57 is useful for making several observations about the nature of the space error and for studying the space discretization error at a node as a function of the location of its neighboring nodes by actually finding the coefficients of the derivatives.

A summary of several nodes and the expansions for the corresponding space discretization errors is in Table V-9. In this summary a sketch is used to show the node shape, coordinate system, and usually only the first non-zero terms are shown in the expansion. The geometric parameters are shown in the most convenient form for the particular geometry rather than any one consistent method. The selection of nodes for the summary has been made to contrast regular nodes and asymmetric nodes and is not meant to be a complete catalogue of all nodes. The errors for the nodes are based on equation V-57 and hence they apply when the following assumptions are valid: (1) the capacity of the node and the conductances to the neighboring nodes are based on MacNeal's rules; (2) the location of the nodes follows MacNeal's rules; (3) the ith node is an internal node, that is all of its neighboring nodes are within an area which has constant thermal properties of volumetric heat capacity, C_p, and thermal conductivity, k. This means that none of the conductances cross an interface into a second solid or to a fluid with a finite heat-transfer coefficient.

2. General Comments--Space Discretization Error.

A study of the expansion for the space discretization error (equation V-57) shows that each term is the product of the second- or higher-order derivatives and a coefficient which is a ratio of finite sums. The derivative part of the product depends on the location of the node within the solid, the time, the solid boundary, and how the

forcing functions change with both time and space variables. Thus, these derivatives cannot be controlled, as they are functions of the individual problem. Since these derivatives are of the second- or higher-order, the space discretization error is zero for problems or times where the temperature gradient is independent of position.

The coefficients of the derivatives depend upon the geometry of the node and its neighbors; hence, some control over the size of the discretization error can be exercised by arranging the nodes so that these coefficients are small, or at least so that the coefficients of derivatives that are likely to be large are small. These coefficients determine the consistency of the approximation, in the Richtmyer sense. Richtmyer has stated (page 43) that, "...for problems with constant coefficients consistency can be determined by simply examining the truncation error term of the difference formula." Further, in the discussion of the propagation of the discretization error, consistency is important if more, smaller nodes are to replace large nodes to improve accuracy. Thus, most of the following discussion is on consistency and how these coefficients affect the space discretization error.

3. Consistency.

The consistency of a general asymmetric node can be studied by studying the above expansion (equation V-57) and finding how the space discretization error changes as the node is made smaller. The

time discretization has already been shown to be consistent as the contribution of the time discretization error ($\varphi_{i,n}$) to $d_{i,n}$ goes to zero as the time increment $\Delta\tau$ goes to zero. Before proceeding to the study of the expansion, it should be noted that, in Richtmyer's definition of consistency, actually both the time increments and the mesh increments for the whole network go to zero together; consequently, the consistency at a node does not have meaning in the precise Richtmyer sense. However, in view of Richtmyer's statement quoted above, a consistent node can be defined as one where the space discretization error goes to zero as its area or a characteristic linear dimension goes to zero. A network of such nodes would follow Richtmyer's definition of consistency, providing that the network could be refined in such a way that each succeeding mesh contained only consistent nodes. A further discussion of these concepts is best delayed until after a discussion of the coefficients in the expansion.

A study of the coefficients in the expansion indicates that the coefficients for the second derivatives, the hyperbolic term, $(\partial^2 T_{i,n}/\partial\xi^2 - \partial^2 T_{i,n}/\partial\eta^2)$, and the mixed partial derivative $(\partial^2 T_{i,n}/\partial\xi\,\partial\eta)$, depend only on the geometry of the node and its neighbors, and not upon the distance between nodes, or the area of the ith node. In order to show this the longest distance between the ith node and any of its neighbors is designated λ. For a node where the distances to each of its neighbors remain in proportion, and the angles

remain constant, reducing this length gives a sequence of geometrically similar nodes, but with smaller areas. The coefficient of the hyperbolic term is

$$\frac{\sum_{j-1}^{N} \left(\frac{\rho_{ij}}{\lambda}\right)\left(\frac{\lambda_{ij}}{\lambda}\right)(\cos\beta_{ij})}{\sum_{j=1}^{N} \left(\frac{\rho_{ij}}{\lambda}\right)\left(\frac{\lambda_{ij}}{\lambda}\right)}$$

and that for the mixed derivative is:

$$\frac{\sum_{j=1}^{N} \left(\frac{\rho_{ij}}{\lambda}\right)\left(\frac{\lambda_{ij}}{\lambda}\right)(\sin 2\beta_{ij})}{\sum_{j=1}^{N} \left(\frac{\rho_{ij}}{\lambda}\right)\left(\frac{\lambda_{ij}}{\lambda}\right)}$$

Since λ_{ij}/λ and the angles β_{ij} are constant the ratios ρ_{ij}/λ are also constant, and therefore these coefficients are independent of the λ or area of the nodes. Therefore, the space discretization errors for such nodes, where the coefficients of the hyperbolic term and the mixed partial derivative are identical and not zero, have a zero-order error term in their series expansions. Using the same arguments, we can say that the coefficients of the third derivatives depend upon the geometry, and, for a sequence of geometrically similar nodes, these coefficients are proportional to λ. In general, the pth order derivatives have coefficients, for geometrically similar nodes, that are proportional to a linear distance (λ) in the node raised to the (p-2)

power, or λ^{p-2}. In the following a term in the expansion proportional to λ^{p-2} and the p space derivative is called a (p-2) order term, i.e., when p is 2, the term is a zero-order term.

A study of the series expansions for the asymmetric nodes in Table V-9 shows that many of the asymmetric nodes do indeed have zero-order error terms, and, hence, if they are refined in a sequence of geometrically similar nodes, they are inconsistent nodes. Further, a comparison of the expansions for the discretization error of the trapezoidal nodes, F, G, H, and I among themselves, and with the expansions for the rectangular nodes C and D, shows that the coefficients for the zero-order error terms are proportional to the geometrical parameters that are associated with the irregularity, or difference from the rectangular node. That is, the coefficient of these terms for node I, the most regular of the trapezoids, is proportional to the square of the sine of the angle by which two of the legs differ from being a rectangle. And, as the lengths and angles are further distorted in nodes H, G, and F, terms proportional to the differences in lengths of the legs and to the sines and cosines of the angles are added to the coefficients of these non-zero terms. Consequently, the conclusion must be made that irregularity or asymmetry in the geometry of the nodes produces the zero-order error terms. And further, if such an asymmetric network, containing asymmetric nodes with zero-order error terms, is refined to zero so that each succeeding network contains

only nodes which are geometrically similar to those in the original asymmetric network, the discretization does not go to zero, and Richtmyer's consistency condition is not satisfied.

In order to verify that, in the expansion for $\sigma_{i,n}$, no unnoticed simplification actually makes the zero-order error terms zero, a numerical study of the space-discretization error was made for nodes J and L from Table V-9. In this study, a node of the shape and proportions of node J or L was assumed at a position of $\xi = 0$ and $\eta = 1$. in a solid bounded at $\xi = \pm \pi/2$, $\eta = 0$, and $\eta = \infty$, with a steady-state temperature distribution and the following boundary conditions:

$$T(\xi, 0) = 1 \qquad -\frac{\pi}{2} < \xi < \frac{\pi}{2} \qquad \text{(V-58)}$$

$$T(\eta, \tfrac{\pi}{2}) = 0 \qquad 0 < \eta \leq \infty \qquad \text{(V-59)}$$

$$T(\eta, -\tfrac{\pi}{2}) = 0 \qquad 0 < \eta \leq \infty \qquad \text{(V-60)}$$

The solution to this problem is given in Churchill (28), page 194, and is

$$T = \frac{2}{\pi} \Psi \qquad \text{(V-61)}$$

where Ψ is given by:

$$\tan \Psi = \frac{\cos\xi}{\sinh\eta} \qquad 0 \leq \Psi \leq \frac{\pi}{2}$$

The temperature at the node and its neighbors can be computed from this expression, and the space discretization error can be found directly from its definition. Also, the second- and third-order derivatives can be found, and the constant and linear terms as functions of

the length λ can be calculated for the series expansions. Thus, the space discretization error can be checked in two ways, directly from the definition,

$$\sigma_{i,SS} = \frac{1}{A_i} \sum_{j=1}^{N} y_{ij}(T_j - T_i)_{SS} \qquad (V\text{-}62)$$

and approximately from the constant (second derivatives) and linear (third derivatives) terms.

The results of these calculations for nodes J and L are shown in Figures V-5 and V-6, respectively. The expansion for node J in Table V-9 shows it to be a consistent node, and the calculation of the space discretization error from the exact defining relationship shows this to be true. Further, the discretization, as found from equation V-62, follows fairly closely the behavior that would be expected merely on the basis of the linear term of the expansion. The expansion for node L indicates that a zero-order term should occur. The calculation, based on the exact, defining equation for σ_i, shows a σ_i of about -0.1257 at a λ of zero, or just about the value predicted by the expansion. Further, the discretization error estimated by the expansion follows closely that found by the defining equation. Great care was required in calculating the space discretization error from the defining equation, because, as the node becomes smaller, $(T_j - T_i)$ becomes very small, and a great many significant figures must be calculated in both T_j and T_i to insure that the quantity $(T_j - T_i)$ contains any

significant figures. As a result, these calculations were made using 10-place tables for the cosine (34), hyperbolic sine (34), and the arc tangent (35), and the temperatures were calculated to 10 decimal places. The scattering of the points for both nodes for the smallest λ's is due to round-off error, and more accurate calculations of T_j and T_i would be required to remove this scatter. However, the calculations are sufficiently accurate to show the validity of the expansions, and that the zero-order error term is a reality for node L.

The regular trapezoid, node I, shown in Table V-9, is another asymmetric node which has a zero-order error term. This node is like the regular rectangular node D, except that the two neighboring nodes, which are located on the ξ-axis for the rectangle, lie above the ξ-axis, so that the legs from node I to these nodes make an angle α with the ξ-axis. The area associated with the node is then trapezoidal in shape. The expansion for the space discretization error is:

$$\sigma_{I,n} = -(\sin\alpha)^2 \left[\frac{\partial^2 T_{I,n}}{\partial \xi^2} - \frac{\partial^2 T_{I,n}}{\partial \eta^2} \right] + \lambda_1 \sin\alpha(\cos\alpha)^2 \frac{\partial^3 T_{I,n}}{\partial \xi^2 \partial \eta} \qquad \text{(V-63)}$$

$$+ \left[\frac{\sin\alpha}{3} \right] \left[\lambda_1 (\sin\alpha)^2 - \frac{\lambda_2^2}{\lambda_1} \right] \frac{\partial^3 T_{I,n}}{\partial \eta^3} + \cdots$$

As previously noted, the cause of the zero-order term is the asymmetry introduced by the inclination of two of the legs at an angle of α to the ξ-axis. And this node must give a non-zero space discretization error at zero area, if the refinement is made using a sequence of geometrically similar nodes.

$$\lim_{\lambda \to 0} \sigma_{I,n} = -(\sin\alpha)^2 \left[\frac{\partial^2 T_{I,n}}{\partial \xi^2} - \frac{\partial^2 T_{I,n}}{\partial \eta^2} \right] \qquad \text{(V-64)}$$

$\frac{\lambda_1}{\lambda}$, $\frac{\lambda_2}{\lambda}$, and α are constant

This non-zero error would be expected to be small if the angle α is small.

However, a network of regular trapezoidal nodes can be used as an approximation to a two-dimensional problem in cylindrical coordinates, with gradients in the radial and angular directions, if MacNeal's rules are followed in setting up the network. Such a network is in Figure V-7. The main difference between this approximation and that usually used for cylindrical coordinates is that the perpendicular bisectors of the radial conductors are used for the conductances, and to bound the node, rather than the $\rho \Delta \theta$ arc lengths. Because of the geometry of the network, α is related to the angle increment, and the length λ_2 is equal to the radial increment:

$$\alpha = \frac{\Delta\theta}{2} \tag{V-65}$$

$$\lambda_2 = \Delta\rho \tag{V-66}$$

Making these substitutions in the expansion (equation V-57) gives:

$$\sigma_{I,n} = \tfrac{1}{2}(\cos\Delta\theta - 1)\left[\frac{\partial^2 T_{I,n}}{\partial\xi^2} - \frac{\partial^2 T_{I,n}}{\partial\eta^2}\right] + \lambda_1\left[\sin\frac{\Delta\theta}{2}\left(\cos\frac{\Delta\theta}{2}\right)^2\right]\frac{\partial^3 T_{I,n}}{\partial\xi^2\partial\eta} +$$

$$+\left[\frac{\sin\frac{\Delta\theta}{2}}{3}\right]\left[\lambda_1\left(\sin\frac{\Delta\theta}{2}\right)^2 - \frac{(\Delta\rho)^2}{\lambda_1}\right]\left[\frac{\partial^3 T_{I,n}}{\partial\eta^3}\right] + \cdots \tag{V-67}$$

Now, if the network and the trapezoidal nodes are considered an approximation to the Laplacian in cylindrical coordinates, and the network and nodes are refined such that the angle increment, $\Delta\theta$, and radial increment, $\Delta\rho$, go to zero, the space discretization error goes to zero:

$$\lim_{\substack{\Delta\theta\to 0\\ \Delta\rho\to 0}} \sigma_{I,n} \to 0 \tag{V-68}$$

and the network and the nodes satisfy Richtmyer's condition for consistency.

Thus, the important point can be made that the manner in which a network is refined determines whether or not a network of nodes satisfies Richtmyer's consistency condition. Likewise, the manner in which an area is made smaller determines whether or not the space discretization error at a node is reduced as its area is made smaller.

Practical Aspects of Consistency. The meaning of the consistency condition, for a network containing asymmetric nodes, can now be interpreted as follows: that the shape of the asymmetric node in a network, or the geometry of its neighbors, must be changed as more nodes are added, if the accuracy of the approximation is to be improved. Thus, in general, the additional nodes should be added so that the coefficients of the hyperbolic and mixed second-order derivatives go to zero, which requires that

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \rho_{ij}\lambda_{ij}\cos 2\beta_{ij} \rightarrow 0 \qquad i = 1, \ldots \qquad \text{(V-69)}$$

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \rho_{ij}\lambda_{ij}\sin 2\beta_{ij} \rightarrow 0 \qquad i = 1, \ldots \qquad \text{(V-70)}$$

Since, for any practical asymmetric network, one is able to add nodes so that the network becomes at least an irregular rectangular network, containing nodes of type C, Table V-9, the asymmetric network can be considered to be a consistent approximation. In refining the mesh all the node shapes need not go to rectangles, but to any of the consistent node shapes shown in the table, or any which are possible. In particular, it would be more logical to consider that the refinements of an asymmetric triangular network could approach the network of equilateral triangles with hexagonal-shaped nodes, type N, Table V-9, rather than a rectangular network.

In most cases of a general asymmetric network, one is not able to refine the network to zero mesh spacing in a manner that allows all, or even some, of the nodes to be geometrically similar to the nodes in the coarse network. Thus, the expansion for $\sigma_{i,n}$ (equation V-57) does not, when interpreted correctly, show that the network of asymmetric nodes is inconsistent. Further, in most practical situations the network is not going to be refined, and in those situations the expansion allows an estimate of the space discretization error if sufficient information is known, or can be approximated, about the second- and higher-order derivatives.

Because Richtmyer's consistency, stability, and convergence definitions are based on successive refinements to zero of both space and time increments, the main practical consequences apply mainly to space networks containing a relatively large number of points (S probably much greater than thirty). When such a fine space mesh is to be used, both Richtmyer's theory and practical considerations dictate the use of a rectangular network or a network of equilateral triangles for the interior points. This is true for the following reasons: (1) The space discretization error is assured to be small only for small, consistent nodes, such as nodes D and N. Further, it should be noted that even radial nodes such as node I become rectangular as $\alpha \rightarrow 0$, according to the refinement just discussed. (2) If a large number of points is used in a regular network, curved or irregular boundaries

can be followed closely, and such boundaries can be followed exactly by the addition of only a few irregular nodes along the boundary.

(3) Since no partial difference equation or recurrence relationship is available for asymmetric nodes, their conductances and capacities must be calculated specifically for each node. Setting up the difference equations can require an amount of calculation which would be prohibitive if done on a desk calculator, and bothersome to program on a digital computer.

Possible Ways of Eliminating Zero-order Error Terms. Before leaving the topic of consistency, it should be mentioned that consistent approximations can be found for irregular nodes of many shapes by solving a system of equations which make the coefficients of the first and second derivatives in the expansion for $\sigma_{i,n}$ zero (10). This can be done by replacing the elements in the row for such a node in the Y/A matrix, $(1/A_i)(\rho_{ij}/\lambda_{ij})$, with a weighting μ_{ij}. Thus equation V-54 becomes:

$$\sigma_{i,n} = \sum_{j=1}^{N} \mu_{ij}(T_{j,n}-T_{i,n}) - \frac{\partial^2 T_{i,n}}{\partial \xi^2} - \frac{\partial^2 T_{i,n}}{\partial \eta^2} \qquad (V-71)$$

Now, using the expansions of $(T_{j,n}-T_{i,n})$, and setting the coefficients of $\partial T_{i,n}/\partial \xi$, $\partial T_{i,n}/\partial \eta$, $\partial^2 T_{i,n}/\partial \xi \partial \eta$ equal to zero, and of $\partial^2 T_{i,n}/\partial \xi^2$ and $\partial^2 T_{i,n}/\partial \eta^2$ equal to one, the following system of equations is obtained:

$$\sum_{j=1}^{N} \lambda_{ij}(\cos\beta_{ij})\mu_{ij} = 0 \qquad (V-72)$$

$$\sum_{j=1}^{N} \lambda_{ij}(\sin\beta_{ij})\mu_{ij} = 0 \qquad (V-73)$$

$$\tfrac{1}{2}\sum_{j=1}^{N} (\lambda_{ij})^2(\sin\beta_{ij})(\cos\beta_{ij})\mu_{ij} = 0 \qquad (V-74)$$

$$\tfrac{1}{2}\sum_{j=1}^{N} (\lambda_{ij})^2(\cos\beta_{ij})^2\mu_{ij} = 1 \qquad (V-75)$$

$$\tfrac{1}{2}\sum_{j=1}^{N} (\lambda_{ij})^2(\sin\beta_{ij})^2\mu_{ij} = 1 \qquad (V-76)$$

The μ_{ij} that satisfy these equations would then be the off-diagonal elements in the ith row, j^{th} position, in the equation for the ith node; the diagonal element is, for the ith row, $\sum_{j=1}^{N} \mu_{ij}$. The existence or uniqueness of the μ_{ij} that satisfy the above equations is not guaranteed, but depends upon the number of neighbors N and their geometrical relationships. Indeed, Forsythe and Wasow (10) state that there are geometrical situations for which no solution exists.

A second, and essentially equivalent, system of equations which has an unknown that can be given a geometrical interpretation can be derived in the same way. This is done by determining a length ρ'_{ij} to replace the length of the perpendicular bisector ρ_{ij} in both the

conductances and heat capacity of a node. An analysis like that above shows that the ρ'_{ij} must satisfy the following system of homogeneous equations:

$$\sum_{j=1}^{N} \cos\beta_{ij}\rho'_{ij} = 0 \qquad \text{(V-77)}$$

$$\sum_{j=1}^{N} \sin\beta_{ij}\rho'_{ij} = 0 \qquad \text{(V-78)}$$

$$\sum_{j=1}^{N} \lambda_{ij}(\sin 2\beta_{ij})\,\rho'_{ij} = 0 \qquad \text{(V-79)}$$

$$\sum_{j=1}^{N} \lambda_{ij}(\cos 2\beta_{ij})\,\rho'_{ij} = 0 \qquad \text{(V-80)}$$

The first two equations are satisfied if the ρ'_{ij} form a closed polygon; the second two equations have the same form as those for the conditions under which a MacNeal node is consistent (equations V-69 and V-70). In principle, the zero-order error terms can be eliminated from the discretization error of all nodes in some asymmetric networks. This could be done by solving a system of equations such as equations V-72 to V-76 for μ_{ij} or V-77 to V-80 for ρ'_{ij} for each node in the network. From a solution for a node one row in the Y/A matrix would be generated. Because of the time-consuming nature of solving such systems of equations, this is probably not practical. However, such a

technique could be useful if solutions were required only for a few irregular nodes which are used to follow an irregular boundary, the rest of the network containing regular, consistent nodes. Unfortunately, because the above analysis applies only to interior nodes, the system of equations does not apply to adjacent nodes with heat-transfer coefficient conductances to a fluid temperature, thus further restricting their utility.

However, before the solution of either system of equations to eliminate zero-order error terms became a general technique, the following studies would have to be carried out: first, the geometrical conditions for which a solution to the equation exists should be found. Second, if several solutions to the equations exist, a criterion for selecting one of them should be found. Third, a determination should be made of the effect of using the solution of these equations on the coefficients of the third and fourth derivatives in the expansion, to be sure that these coefficients are not greatly increased. Fourth, a study would have to be made of the properties of the Y/A matrix which contains elements that are a solution of the above system, to see if modifications in the stability and non-oscillatory criteria are required.

(F) ACCURACY OF NUMERICAL RESULTS BASED ON AN ASYMMETRIC NETWORK

1. Problems Studied and Network Used.

To further investigate the effect of asymmetry on the accuracy of the approximation solutions, the numerical results for the transient problem used in Chapter III, section D, and for the steady-state problem mentioned in section C of this chapter can be studied and compared with their corresponding continuous solutions. Both of these approximate calculations are based on the asymmetric network in Figure III-1.

The transient problem is described in Chapter III, section D. Briefly, a solid with a unit square cross-section is initially at a unit temperature, and at time zero the surfaces are held at zero temperature. This problem is important as the eigenvalues and eigenfunctions for this problem are the same for any transient problem in the unit square with a different initial temperature distribution and with surface temperature a function of space and time; hence, the accuracy with which the asymmetric approximation solves the above transient problem is an indication of the accuracy of the transient part of the approximate solution. In this problem only the triangular one-eighth of the cross-section bounded by $\xi = 0$, $\xi = -\eta$, and the surface, $\eta = -\frac{1}{2}$,

as shown in Figure III-1, need be used, because the temperature distribution is always symmetric about both the hypotenuse ($\xi = -\eta$) and the vertical boundary ($\xi = 0$). Thus, the problem can be considered to be just the triangle with the surfaces at $\xi = 0$ and $\xi = -\eta$ adiabatic, and the surface at $\eta = -\frac{1}{2}$ at zero temperature. The details of the approximate transient calculation are given in Chapter III, and the continuous solution is given by equations III-34 and III-35.

The steady-state problem has been described in section C of this chapter. It is to find the steady-state temperature distribution in a unit square with its horizontal surfaces at unit temperature and its vertical surfaces at zero. The steady-state solution is the remaining part of the complete solution for the problem of transient temperature distribution; the complete continuous or approximate solution actually is often found by finding the transient and steady-state parts separately and then adding them (see Chapter II and reference (1)). Thus, the results of this study can be used as an indication of the accuracy of the steady-state part of the approximate solution. Further, the errors in this solution are a direct result only of the space discretization error; and hence, this study gives an understanding of the practical effect of zero-order error terms in the expansion for the space discretization error. The continuous solution of this problem is given in equation V-32. The temperature distribution for this problem is symmetric about the vertical ($\xi = 0$) axis and anti-symmetric about the

hypotenuse ($\xi = -\eta$). Therefore, again, only the 1/8-cross-section used in Figure III-1 need be considered, as a problem with the surface at $\xi = 0$ adiabatic, the hypotenuse ($\xi = -\eta$) at a temperature of $\frac{1}{2}$, and the surface at $\eta = -\frac{1}{2}$ at unit temperature.

The arrangement of points used for both problems is shown in Figure III-1. This arrangement was deliberately made very irregular so that the effects of asymmetry could be studied and it is far from the optimum, which would probably be a network of squares. Further, the network contains only 8 or 9 points at which the temperature is to be found; this is an extremely coarse network equivalent to about 3 points for a one-dimensional problem.

The surfaces of the triangle are described by the two types of nodes; the first, located on the $\eta = -0.5$ boundary, follows MacNeal's rules by being on the boundary. However, the nodes along the vertical boundary and the hypotenuse, except for node 9, are located inside the solid so that part of the boundary of the nodes follows the boundary of the solid, rather than the node being located on the surface boundary. The location, calculation of the conductances and capacitances, and the advantage of this type of node are discussed in the next section.

Although the arrangement of the points for the two problems is the same, the Y matrices are different because of the different boundary conditions along the hypotenuse. In the transient problem the heat-transfer coefficient is zero, but it is infinite in the steady-state problem.

Thus, the Y matrix for the transient problem is 9 x 9, but because the temperature at node 9 is specified in the steady-state problem it is fixed at $\frac{1}{2}$ and the Y matrix for the steady-state problem is 8 x 8. However, the approximation is in reality the same for each problem. This can be seen by considering the square which is $\frac{1}{4}$ of the original square rather than the triangular 1/8. Now all that has been changed is boundary temperatures, which do affect the coefficients of the derivatives in the expansion for the discretization error.

2. Transient Problem.

The approximate solution to the transient problem was stepped out for each of several time increments for each of the most common values of the weighting factor of γ, of 0, $\frac{1}{2}$, and 1. Table V-10 is an abstract of errors for each of the nine points at 6 times for several different combinations of the time differencing parameters. The results at temperature points 1 and 6 were studied in more detail for only four combinations of time differencing parameters. These studies are shown in semi-logarithmic graphs for both the continuous and approximate solutions for the temperature versus dimensionless time. The graphs for point 1, for short, intermediate, and long times, are in Figures V-9, V-10, and V-11, respectively; and for point 6 for short and long times in Figures V-12 and V-13, respectively. After determining the effect of the time differencing parameters the effect of the space discretization errors can be studied.

Time Differencing Parameters. The solution labeled analog solution in the table and on the graphs actually is the fourth-order Runge-Kutta method of integration for a system of ordinary differential equations. Three such calculations were made with time increments of 0.01, 0.001, and 0.0001. The calculations with the last two time increments agreed to within at least the fourth significant digit for most points even at very early times. Because of this and because of the reported high accuracy (18) of the Runge-Kutta methods, this calculation was assumed to be essentially the analog solution.

Several observations can be made by studying Table V-10. First, for all points and all times, with the exception of point 1 at short times, the analog solution has approximately the same error as do the difference solutions with relatively small time increments ($\Delta\tau \leq 0.0005$) for all values of weighting factors, and difference solutions with a γ of $\frac{1}{2}$ for time increments up to and including 0.0015. This would be expected from the conclusions about the time differencing parameters for the one-dimensional problem. In Chapter IV, section E-4, and from Figure IV-15, it was concluded that the damping factors for approximate solutions with a small time increment, for a weighting γ of $\frac{1}{2}$, are very close to those of the analog solution. Thus, this dependence on time differencing parameters for an approximate solution in two dimensions would be expected, because the form of the expansion for the time discretization error is the same for both one- and two-dimensional approximations.

Second, the approximate solutions for the largest time increment of 0.01 for weightings γ both of $\frac{1}{2}$ and 1 show relatively large errors, although both solutions are stable. The large error in the approximate solution for a weighting of $\frac{1}{2}$ is caused by excessive oscillation (see Figure III-3). The large inaccuracies for the run with the weighting of one agrees with the conclusion in Chapter IV, section E-2, where it was shown that this weighting gives the least accurate damping factors for a fixed time increment and a fixed number of points. Note also that, of the approximate solutions that are like the analog solution, the solution with a γ of 1 and time increment 0.0005 shows the poorest agreement with the analog and also is the least accurate of the approximations.

Third, the solutions for node 6, and its neighboring nodes 4, 7, and 8, have the largest errors. Further, these errors for node 6 do not change much over a wide change in time differencing parameters. However, the errors for nodes 1, 2, and 3 are much smaller, and these errors do depend upon the time differencing parameters. This effect can be explained by the space discretization error and is discussed in the next section.

The graphs of temperature versus dimensionless time for the continuous solution and for each of several approximations are shown in Figures V-8, 9, and 10 for node 1 and in V-11 and V-12 for node 6. In the graphs the analog solution is shown and considered to be

representative of the solutions with small time increments ($\Delta\tau \leq 0.005$) or with weighting γ of $\frac{1}{2}$ up to time increments of 0.0015. Also shown are the two backward difference solutions with time increments of 0.0005 and 0.01 and the forward difference solution with a time increment of 0.001.

As shown in Figures V-8 and V-9 the continuous temperature at node 1 decreases very rapidly with time because it is close to the surface. This means that the smaller damping factors have a relatively large weighting in the solution for this node and give a significant contribution to the solution at short times. The approximate solutions for node 1 are the most sensitive to oscillations because of this relatively large weighting for the smaller damping factors. Moreover, the significant contribution of these smaller damping factors causes the disagreement between the approximate difference solutions and the analog solution at short times. Probably even smaller time increments would be required to make these smaller damping factors agree with those of the analog solution. In Figure V-10 the temperature at node 1 becomes linear on the semi-logarithmic paper at long times, as expected. The behavior of the continuous and approximate solutions at nodes 2, 3, and 4 is like that at node 1 except that the weightings of the smaller damping factors are not as great at these nodes.

In contrast to the behavior at node 1, the continuous temperature at node 6 remains about constant for a short time after the surface

temperature is changed, and then decreases almost linearly with time as shown on the semi-logarithmic graph, Figure V-11. This indicates that at most only one or two damping factor terms give a significant contribution to the solution at node 6. The straight line behavior actually begins when node 6 has reached a temperature of 0.65 compared to node 1, which does not show this behavior until a temperature of 0.007 is reached. This behavior of both the continuous and approximate solutions at node 6 is like that at nodes 5, 7, 8, and 9.

From a study of the graphs for long times the following important observations about the straight-line portion can be made for both nodes. The approximate solution for the backward difference calculation ($\gamma = 1$) with a large time increment, 0.01, decreases more slowly than the central difference calculation ($\gamma = \frac{1}{2}$) as represented by the analog solution. The central difference calculation ($\gamma = \frac{1}{2}$), in turn, decreases more slowly than the explicit calculation ($\gamma = 0$) with a time increment of 0.001. This means that the slowest-decaying damping factor, q_{MAX}, for the backward difference calculation is greater than that for the central difference calculation which is greater than that for the explicit calculation. From the behavior of the approximate solutions, this ranking of damping factors probably also applies for the smaller damping factors. This is the same dependency of the damping factors on the weighting γ which was shown in Chapter IV, section E-4, for the one-dimensional problem. Such a similarity in the behavior of the

damping factors and the weighting γ between the one-dimensional regular network and the two-dimensional asymmetric network would be expected because the time discretization error has the same form in both cases. Thus the conclusion can be made that the approximate transient solution for a weighting γ of 1 decays much too slowly, compared to the continuous solution, and this is the reason for its inaccuracy.

However, a discrepancy does occur between the regular one-dimensional approximation and the two-dimensional asymmetric approximation. This is that for the one-dimensional approximation the solution with a weighting of $\frac{1}{2}$ gives damping factors larger than the corresponding damping factors of the continuous solution; for a γ of zero damping factors less than the corresponding damping factors of the continuous solution are obtained; and a γ of γ_o gives damping factors very nearly equal to those of the continuous solution. However, for the asymmetric approximation even the explicit solution with a γ of zero and a time increment of 0.001 decays more slowly than the continuous solution, indicating damping factors larger than those of the continuous solution. A possible reason for this discrepancy is that the space discretization error for the network used was too large to obtain the necessary compensation with the time discretization error. The large discretization could be caused either by the asymmetry or the coarseness of the network, but most probably by the asymmetry.

Other observations based on the graphs are:

(1) The approximate solution for the explicit calculation with the relatively large time increment of 0.001 oscillates too wildly at node 1 at short times to be useful. However, for node 1 at long times and node 6 it is one of the most accurate solutions.

(2) The implicit calculation with γ of 1 and a large time increment of 0.01 is inaccurate almost all of the time.

(3) The approximate solutions that give results equivalent to the analog solution give reasonably accurate solutions despite the coarseness and asymmetry of the network. The important calculation in that group of approximations is the one with γ of $\frac{1}{2}$, which does not require too small a time increment to give an approximation close to the analog.

Selection of Time Differencing Parameters for an Asymmetric Network. The numerical results obtained above and the form of the time discretization error are very much alike for both the one-dimensional regular mesh and the two-dimensional asymmetric network. Consequently, the rules for selecting the time differencing parameters are also similar. However, since there does not appear to be a specific optimum weighting, γ_o, for the asymmetric network, the accuracy restriction is somewhat arbitrary, and it is suggested that the weighting γ not be larger than $\frac{1}{2}$. It should be pointed out that, for the asymmetric network above, using any weighting other than zero increases

the errors, and the weighting factors that give the most accurate q_{MAX} are negative. Thus, using a large weighting factor, approaching one, for an asymmetric network can introduce significant errors.

The oscillatory restriction that assures that the maximum oscillatory contribution is less than V at the initial time of interest is found by writing equation IV-295 for an arbitrary minimum matrix norm, M:

$$V >> \bar{\Phi}(\underline{\nu})^{\tau_0/\Delta\tau} = \bar{\Phi}\left[\frac{M(1-\gamma)\Delta\tau - 1}{M\Delta\tau - 1}\right]^{\tau_0/\Delta\tau} \qquad (V-81)$$

Thus, using this equation, the allowable time increments for a fixed network of points with γ's of 0 and $\frac{1}{2}$ can be found; also, the weighting γ can be found which will allow the use of the minimum practical time increment.

In order that an implicit solution involve less time-consuming calculation than the explicit solution, the ratio of the time increment for the implicit solution to that for the explicit must satisfy the following relationship,

$$\frac{\Delta\tau_{Im}}{\Delta\tau_{Ex}} > \frac{5U + 2S + \sum_{i=1}^{S} u_i^2}{2U + S} \qquad (V-82)$$

where U = total number of off-diagonal elements above diagonal in

$$Y/A \text{ matrix} = \sum_{i=1}^{S} u_i$$

and u_i = number of off-diagonal elements above diagonal in i^{th} row in Y/A matrix

This is based on the number of non-zero multiplications required to advance one time increment where the implicit calculation is carried out according to the elimination method discussed in Chapter III, section E. If u_i does not change much from node to node, the summation of the u_i^2 terms can be approximated by

$$\sum_{i=1}^{S} u_i^2 \simeq \bar{u}\, U \qquad \text{(V-83)}$$

where $\bar{u} = \left(\sum_{i=1}^{S} u_i \right) / S$ = one-half the average number of nodes that surround each node in the network.

If a rectangular network is used, a node has four neighbors, u_i is two, and the relationship becomes:

$$\frac{\Delta\tau_{Im}}{\Delta\tau_{Ex}} > \frac{16}{5} = 3.2 \qquad \text{(V-84)}$$

If the network contains equilateral triangles, each node has 6 neighbors, and u_i is three; the ratio becomes:

$$\frac{\Delta\tau_{Im}}{\Delta\tau_{Ex}} > \frac{26}{7} = 3.72 \qquad \text{(V-85}$$

For most asymmetric networks the average $\bar{u}$ is probably somewhere between two and three and thus $\Delta\tau_{Im}$ must be somewhat larger than about $3.5\Delta\tau_{Ex}$ for the implicit calculation to be more efficient. This is the same ratio found for one-dimensional problems with a regular mesh where vector-matrix multiplication is not used (equation IV-305).

Based on these results, it is probable that for most asymmetric networks the weighting γ of $\frac{1}{2}$ should be used with a time increment about the maximum given by equation V-81. Although this might not give the most accurate solution possible with the specific network, in all probability it should give reasonable accuracy with no significant oscillations, and should require about the minimum calculations.

Space Differencing Parameters. One of the effects on the approximate transient solution, which is probably attributable to the asymmetric location of the temperature points, is that the eigenvalues of the Y/A matrix are sufficiently different from the eigenvalues υ_j^2 of the continuous problem, so that the damping factors of the approximate solution are greater than those of the continuous solution for any reasonable combination of weighting and time increment.

The other inaccuracy in the transient solution caused by the space discretization error, as indicated in Chapter II, section D, and Chapter IV, section E-2, is the error in the intercepts of the transient

terms, when plotted versus time on semi-logarithmic paper. This difference is $(a_j b_{ij} - g_j c_{ij})$. On the long-time plot for nodes 1 and 6, Figures V-10 and V-12, the straight-line portion is extrapolated to zero time to find the intercept corresponding to the maximum damping factor for both the continuous and approximate solutions. This shows that the error in this quantity is much larger for node 6, about 0.09, or about 10 per cent error, compared to 0.01, or about a 3 per cent error, for node 1. Further, since in both of the approximate intercepts the quantity g_1 is the same for both nodes, the larger error for node 6 is caused by the fact that its component c_{61} in the first eigenvector is more in error than the corresponding component c_{11} for node 1. Moreover, none of the approximate solutions for node 6 give the very slow changes in temperature for very short times ($\tau < 0.005$) in comparison with the analog solution for node 1 which is relatively close to the continuous solution. This indicates that the intercepts for node 6, $g_j c_{6j}$, for the remaining terms in the approximate solution are quite inaccurate, but that the corresponding intercepts for node 1, $g_j c_{1j}$, are accurate. In both cases, much of the difference in accuracy must be due to the fact that the elements in row 6 of the eigenvector matrix do not agree with the eigenfunctions evaluated at node 6; but row 1 in the eigenvector matrix agrees with the first several eigenfunctions evaluated at point 1. Secondary reasons for the higher accuracy for node 1 are that its approximate solution might contain higher weightings

of the approximate transient terms which are more accurate, and that errors in the several terms which are significant at short times might compensate.

The relative accuracy and behavior of the transient terms for nodes 1 and 6, and for the other nodes, can be explained qualitatively in terms of the error propagation equation and the space discretization error expansions with only a superficial knowledge of the derivatives. First, the expansions for the space discretization for nodes 1 and 6 are shown in Table V-9 as nodes O and P, and these expansions have non-zero coefficients of the zero-order error terms, which are the hyperbolic derivative term $\left(\frac{\partial^2 T}{\partial \xi^2} - \frac{\partial^2 T}{\partial \eta^2} \right)$ and the mixed derivative $(\partial^2 T/\partial \xi \partial \eta)$. Examination of the network in Figure III-1 indicates that these terms would be expected for all but node 3 which is an irregular rectangular node with the first term in the expansion for its discretization error proportional to a linear distance in the node (see node C, Table V-9). A consideration of the symmetry of the temperature distribution or the adiabatic condition for the vertical boundary ($\xi = 0$) and the hypotenuse boundary ($\xi = -\eta$) allows the following conclusions about the derivatives in the zero-order error terms.

(1) The hyperbolic derivative term $\left(\frac{\partial^2 T}{\partial \xi^2} - \frac{\partial^2 T}{\partial \eta^2} \right)$ is small for nodes 1, 2, 3, 5, and 7, located close to the hypotenuse, because $\partial^2 T/\partial \xi^2$ and $\partial^2 T/\partial \eta^2$ are the same size and have the same sign.

(These second derivatives are equal on the hypotenuse.) The mixed partial derivative in this region is probably intermediate in size.

(2) The hyperbolic-derivative terms for node 4, located very close to the horizontal surface and relatively close to the adiabatic boundary ($\xi = 0$), and node 6, located very close to the adiabatic boundary ($\xi = 0$), are large as the derivative $\partial^2 T/\partial \eta^2$ has a large absolute value, but $\partial^2 T/\partial \xi^2$ is small. The ξ derivatives are small on the vertical boundary. The mixed derivative is very small for these nodes.

(3) The space discretization error for node 9 is small because of the low rates of heat transfer from this node, and because of symmetry.

(4) Both the space and time derivatives at each node are zero at zero time and increase to a maximum value when the temperature gradients become large. This maximum value is attained at different times for different nodes, occurring at short times for nodes 1, 2, 3, and 4, but at long times for nodes 6, 7, and 8. After this maximum is reached the derivatives go to zero as the steady-state condition of a uniform zero temperature is approached.

Based on these considerations, both the discretization error $d_{i,n}$ and the averaged discretization error $\bar{d}_{i,n}$ would be expected to be relatively large for nodes 4 and 6, and relatively small for nodes 1, 2, 3, 5, 7, and 9. The larger discretization errors for node 4 should occur at short times; the larger discretization errors for node 6 occur at somewhat longer times. As discussed previously, and as seen from

equation V-22, the discretization error associated with each node is now multiplied by the node's area to find the area-weighted averaged discretization error vector elements $A_i \bar{d}_{i,n}$. The elements in this vector, although less in absolute value than those in $\bar{d}_{i,n}$, are increased for the large nodes 4, 5, 6, 7, and 8 relative to the size of errors at the small nodes 1, 2, and 3. This is particularly true for the weighting at node 6 because it has the largest area of any of the nodes.

The quantities which are the relative weighting of the area-weighted discretization error in the error at each point are the elements in the inverse conductance matrix Y^{-1}. This matrix is in Table V-1. It shows that the large discretization error for node 4 at short times should contribute more to the error at nodes 4, 5, 6, 7, 8, and 9, than at nodes 2 and 3, and it should contribute almost insignificantly to the error at node 1. The errors in Table V-10 indicate qualitatively that this is true. The errors for dimensionless times of 0.003 and 0.006 are largest for nodes 4 and 6; the next larger errors at nodes 2 and 3 are probably caused by discretization errors at those nodes in addition to the contribution from node 4. Further, the error at node 4 is almost independent of the time differencing (except for the oscillatory run, $\gamma = 0$, $\Delta\tau = 0.001$), indicating that its discretization error is mostly caused by the space differencing parameters. On the other hand, node 1, despite its asymmetry, has a very accurate solution, because it is not affected by nodes other than 2 and 3, which probably have a low discretization error.

At intermediate and long times the discretization errors at nodes 1, 2, 3, and 4 become very small and most of the approximate solutions become accurate at these nodes. However, the space discretization error at node 6 is now large and because of its large area this discretization error contributes to the large error in its own approximate solution, and to the errors of nodes 5, 7, 8, and 9. In particular, node 9 is known to have a very small space discretization error, and yet its solution is not accurate, because of the dominant effect of the large area-weighted discretization error at node 6, $A_6 \bar{d}_{6,n}$.

3. Steady-State Problem.

In the steady-state problem only the space discretization errors need be considered and only for one time. The analysis of the effect on the errors in the approximate steady-state solution follows much the same pattern as above. In the transient problem just discussed, the temperatures approach a constant value throughout the solid and the discretization error, and hence the error in the solution approaches zero with time; however, in the steady-state problem, the boundary conditions are such that the temperature distribution is very non-uniform. This problem was selected particularly because of this non-uniformity, as a test of the space discretization error.

As pointed out previously, despite the radically different boundary condition along the hypotenuse, ($\xi = -\eta$), the coefficients of the derivatives in the expansion for the space discretization error are the same

as for the transient problem; however, the values of the derivatives at the several nodes are radically changed. As this is a steady-state problem the continuous temperature distribution satisfies Laplace's equation and the hyperbolic derivative becomes

$$\frac{\partial^2 T}{\partial \xi^2} - \frac{\partial^2 T}{\partial \eta^2} = 2\frac{\partial^2 T}{\partial \xi^2} = -2\frac{\partial^2 T}{\partial \eta^2} \tag{V-86}$$

Consequently, the hyperbolic derivative is very large in regions of high net energy flow in the ξ or η direction. From a consideration of the boundary conditions, one such area is the corner where nodes 1, 2, and 3 are located. In this region the mixed second partial derivative is large also. Thus, nodes 1 and 2, which contain these terms in their expansions, would be expected to have very large space discretization errors. The expansion for node 3 is given as node C in Table V-9, and although the zero-order error terms are not present, the third derivatives are also high, as can be seen by writing the expansion for the steady-state problem at node 3 as

$$\sigma_{3,SS} = 0.118\frac{\partial^2 T_{3,SS}}{\partial \xi'^2} + 0.0115\frac{\partial^3 T_{3,SS}}{\partial \eta'^3} + \ldots \tag{V-87}$$

where ξ' is perpendicular to the hypotenuse and η' is parallel to the hypotenuse. Although the third derivative along η' is small, the one along ξ' perpendicular to the hypotenuse is large and causes the discretization error at node 3 to be large. This is true even though the

expansion for node 3 contains no zero-order error terms.

The discretization error would be expected to be decreased as one moves along the hypotenuse toward the origin; thus, the discretization errors for nodes 5, 7, and 8 should be significantly smaller than those for nodes 1, 2, and 3. Node 5 would be expected to have a larger discretization error than nodes 7 and 8.

On the other hand, at node 6, not close to either boundary of specified temperature, but located close to the adiabatic vertical boundary ($\xi = 0$), the second partial derivative $\partial^2 T/\partial \xi^2$ is much smaller than in the corner region. The mixed second partial derivative is also smaller, and it would be expected that the higher-order derivatives are also small. Consequently, the space discretization error at node 6 should be small.

The space discretization error for each of nodes 1 to 8 has been calculated directly from the definition of this error, equation V-4, using the continuous solution temperatures, as computed from equation V-32. These results are summarized in Table V-7 (1) for the case when the boundary temperature vector is taken as the continuous boundary temperature vector. A study of the space discretization error $d_{i,SS}$ does indeed confirm the qualitative expectations mentioned above. That is, nodes 1, 2, and 3 have very large discretization errors compared to node 6, and larger errors than nodes 5 and 7. Node 2 has a significantly smaller discretization error than 1 and 3, as it is located somewhat farther from the hypotenuse. Node 4 has a large error as

it is located close to a region where $\partial^2 T/\partial \xi^2$ is high, due to the closeness of the boundary with a specified temperature ($\xi = -0.5$). Consequently, we conclude that the series expansions for the space discretization errors are valid, and, more importantly, that, with only a very superficial knowledge of the expected temperature distribution or its derivatives, the relative size of the space discretization error can be predicted.

The discretization errors shown in Table V-7 for the nodes in Figure III-1, with the temperature distribution of the steady-state problem, are much larger than the errors shown for nodes J and L in Figures V-5 and V-6, for a steady state which is similar to that under consideration here. There are two reasons for this. First, nodes J and L were located on the semi-infinite solid on the vertical axis of symmetry and fairly well into the interior, which would correspond roughly, for the problem considered here, to a point located at $\xi = 0$ with the same η coordinate as node 6. Thus, these nodes are positioned at a point where the derivatives are small rather than at the corner where the derivatives must be large. Second, if the two problems had their space dimension scaled in the same way, the largest dimension λ for nodes J and L would correspond to about 0.03 instead of 0.10 as shown. Thus the largest nodes J and L have a dimensionless area of 0.0009 compared to the areas of the nodes in Figure III-1, which range from 0.0034 to 0.0242, and the calculations

for J and L are for nodes in a much finer network.

The area-weighted discretization error $A_i d_{i,SS}$ and the errors in the approximate solutions $v_{i,SS}$ are also shown in Table V-7. The area weighting of the discretization error for the small nodes 1, 2, and 3 reduces the large discretization error relative to the area weighting of the discretization errors at the larger nodes. This area-weighted discretization error is largest for nodes 1 and 4; and this error at nodes 2, 3, and 5 is significantly larger than at 6, 7, and 8. The error in the approximate solution for the ith node is found by multiplying each $A_j d_{j,SS}$ by the jth element in the ith row of the inverse conductance matrix. This matrix is in Table V-2. In general, those nodes in a region with high discretization errors should have the largest errors in their approximate solution; nodes with smaller discretization errors should have more accurate solutions, if they do not have large conductances to a large area with large discretization errors.

These results are confirmed in Table V-7, which shows larger errors at nodes 1, 2, 3, and 4, than at 5, 6, 7, and 8. The approximate solution for each point, if rounded to two significant digits, would agree, within a difference of one in the last digit, with the continuous solution, similarly rounded; this is true despite the large discretization errors caused by the asymmetry and the coarseness of the network. This surprisingly good agreement is partially accounted for by the

compensation caused by the discretization errors having different signs, and by the Y^{-1} matrix containing only negative elements.

4. Summary of Conclusions--Numerical Studies.

The conclusions that can be made based on the numerical results for the transient and steady-state problems are:

(1) In spite of the coarseness and the asymmetry of the network, both the approximate transient solution and the approximate steady-state solution are reasonably accurate. This means that the eigenvectors, eigenvalues, and the resulting initial vectors for an asymmetric distribution of points can be of reasonable accuracy.

(2) The relative accuracy of these solutions at different points and at different times is at least qualitatively predictable on the basis of the series expansions for the discretization error and a consideration of the quantities in the equation for the propagation of this error. This can be done by considering the derivatives for the expected temperature distribution, and the relative areas and conductances between the points.

(3) The error in the approximate solution for each point in a fixed network is largely dependent on the second- and higher-order space derivatives for the specific temperature distribution being approximated. That is, a network which gives an accurate approximate solution for nodes in a certain region for one problem might give a

rather inaccurate solution for another problem, where the temperature distributions are different.

(4) The time differencing parameters, time increment and weighting factor γ, influence the approximate solution based on an asymmetric network about as they do for a one-dimensional problem with a regular network. In selecting these parameters for a fixed network of points, the oscillatory criteria probably limit the size of time increment that can be used. The weighting factor γ should probably not be much greater than $\frac{1}{2}$, if accuracy is important, but the solution for a γ of $\frac{1}{2}$ should give an adequately accurate solution.

(G) RULES FOR LOCATING A LIMITED NUMBER OF POINTS

Because Richtmyer's consistency condition is based on how the discretization error changes as the network is refined, it is not directly useful in many practical situations where the network contains only a limited number of points. Although networks of 100 to 1000 points in each space coordinate direction have been suggested by Forsythe and Wasow (10), to obtain moderate or fine detail, in most practical cases approximate solutions of adequate engineering accuracy can be obtained with a coarse mesh of less than thirty points. Furthermore, computers that can handle such a large number of points are not always available. In simulation problems, where the transient temperature distribution calculation is only a part of the total calculation, only part of the memory capacity of the computer is available for the transient calculation, and the time allowed for the calculation might also be limited. Consequently, even with the large high-speed computers currently in use, approximate methods which do not require a large number of points are important practical tools to obtain solutions that are of sufficient accuracy for engineering purposes.

In these cases, where a relatively coarse network of less than thirty points is to be used, the accuracy of the approximation is affected by the discretization errors of the network to be used, and the way in which the discretization changes as more points are added is not

of practical significance (unless these points are actually added). Consequently, the expansion for the discretization error (equation V-57) is important in this situation because it allows an estimate of the discretization error with only a minimal knowledge of the temperature distribution. Further, for a relatively coarse network, the first- and second-order error terms can be as large as the zero-order term; and, since the network is not to be refined, the zero-order term has no special significance. The primary concern is the space discretization error, which is the sum of all terms. Moreover, the numerical results just discussed show that solutions of reasonable accuracy can be obtained with a coarse asymmetric network that is probably very different from the best network for the problems approximated. In the light of the above remarks, and also the numerical results, a properly selected asymmetric network used to describe a solid with irregular boundaries should not be rejected merely because series expansions based on Cartesian coordinates for some nodes in the network contain a zero-order error term.

The following is a summary of several general rules on locating nodes in a solid with irregular boundaries and, possibly composed of materials of different thermal properties. The rules assume that MacNeal's rules are to be used in locating the points and in calculating the conductances; however, on the basis of the results of Chapter IV, a modification is proposed to the location of the points that follow the

curved boundary. In most instances the rules are based on physical considerations or mathematical analogies rather than rigorous mathematical proofs. The discussion is divided into two parts; one on the location of the adjacent nodes that describe the boundary; and one on the location of the interior points within the solid.

It should also be pointed out that because of the general nature of the problems that can occur, and the usual lack of all but a superficial knowledge of derivatives, these rules cannot predict quantitatively the error for a fixed network, or how many points are necessary to obtain a certain desired accuracy. However, a qualitative estimate of the expected relative accuracy can be obtained, based on an analysis like that used in the discussion of the numerical results. A quantitative study would require a series expansion for the discretization error and quantitative estimates of the derivatives at each node. Thus, in the following discussion, the maximum number of points that can be used is assumed to be relatively small and known. The average area for each node in a uniform solid can then be estimated by dividing the total area of the solid by the number of nodes. The average distance between nodes can be estimated by taking the square root of this number.

1. Location of Adjacent Nodes.

MacNeal's original idea of the asymmetric network was to follow a curved or irregular boundary by actually locating the points on the

boundary. The argument for this is that any errors introduced by the resulting asymmetry are expected to be less than the errors caused by making the solid boundary conform to a regular network or other approximation. Locating the points on the boundary is a generalization of method G based on mesh $\Delta\xi$ for the one-dimensional problem; hence, this technique is subject to the same disadvantages that are associated with method G. A method of following the boundary exactly, but locating the nodes a short distance inside the boundary, is suggested. This location of the adjacent nodes is a generalization of the one-dimensional method C based on mesh $\Delta\xi/2$, and it is shown to preserve the advantages of method C.

Two types of boundary conditions are discussed separately. The first is the heat-transfer coefficient condition given by equation V-42. This equation applies to the external surfaces of the solid which are exposed to the fluid. Since the heat-transfer coefficient is allowed to vary from zero to infinity inclusive, this condition includes the adiabatic boundary condition, and the specified temperature boundary condition. The second type of boundary equation considered is the interface boundary condition between two solids of different uniform thermal properties in a composite solid. This condition assumes perfect thermal contact:

$$k_I \frac{\partial T_I}{\partial \eta} (\xi_i, \eta_i, \tau) = k_{II} \frac{\partial T_{II}}{\partial \eta} (\xi_i, \eta_i, \tau) \qquad \text{(V-88)}$$

and

$$T_I(\xi_i, \eta_i, \tau) = T_{II}(\xi_i, \eta_i, \tau) \qquad \text{(V-89)}$$

where $T_I(\xi_i, \eta_i, \tau)$ = temperature of solid I with thermal conductivity I evaluated at the interface with solid II,

and $T_{II}(\xi_i, \eta_i, \tau)$ = corresponding function for solid II.

Fluid Boundary Condition, $0 \leq h \leq \infty$. The disadvantage of locating the points directly on a boundary surface in contact with a fluid, as shown in Figure V-1, are:

(1) Heat capacity is associated with the boundary surface temperature in the approximate method, but no heat capacity is associated with the surface in the continuous formulation; consequently, the approximation is not a direct discretization of the continuous boundary condition, equation V-42.

(2) For infinite heat-transfer coefficient, some heat capacity of the solid is neglected.

(3) For very large heat-transfer coefficient, heat capacity of the surface node must be neglected, or the stability and non-oscillatory criteria become very restrictive because the minimum norm M of the Y/A matrix is directly proportional to the heat-transfer coefficient.

All of the above objections can be avoided and the curved boundary followed exactly if the adjacent nodes are located a short distance inside the surface rather than on the surface. This gives a generalization of method C based on mesh $\Delta\xi/2$.

A portion of such a network is in Figure V-13. This can be constructed easily by first drawing a curve that is approximately parallel to the boundary and inside the boundary by about half the average distance between nodes. The adjacent nodes are then located on this curve. From each of these adjacent nodes a line normal to the boundary is drawn. The interior nodes can be located, and the network and perpendicular bisectors constructed, in the usual manner.

The equation for the adjacent nodes is based on an energy balance on the area bounded by the perpendicular bisectors and a portion of the boundary of the solid. This area is cross-hatched for the ith node in Figure V-13, and it is the area used for the heat capacity. (The feature that distinguishes this type of adjacent node from interior nodes is that a boundary of the adjacent node is also the boundary of the uniform solid.) The energy into the node from the fluid at temperature $t_{f_i,n}$ can be found by using the concept of thermal resistances. The resistance between the fluid and the surface of the solid is $1/A_{f_i}h$, and between the surface and the node is $\ell_{i,f_i}/A_{f_i}k$. The conductance between the node and the fluid is then the reciprocal of the sum of these two series resistances, divided by k_0 to make the expression dimensionless.

$$y_{i,f_i} = \frac{A_{f_i}}{k_0} \left[\frac{1}{\frac{1}{h} + \frac{\ell_{i,f_i}}{k}} \right] \qquad (V\text{-}90)$$

Use of this dimensionless conductance, together with other dimensionless conductances found in the usual way and the dimensionless heat capacity found using the area bounded by the perpendicular bisectors and the solid boundary, gives all the elements in the Y/A and Y_B/A matrices in the row of the ith adjacent node. This energy balance is shown in Figure V-13 in dimensional coordinates. Comparison between the equation above for the dimensionless conductance and the derivation for mesh $\Delta\xi/2$, equation IV-32, shows that this formulation is analogous to method C, where A_{f_i} is one and ℓ_{i,f_i} is $\Delta\xi/2$. The conductance to the fluid temperature approaches a constant value as the heat-transfer coefficient goes to infinity.

$$\lim_{h \to \infty} y_{i,f_i} = \frac{A_{f_i} k}{k_0 \ell_{i,f_i}} \qquad (V\text{-}91)$$

whereas equation V-2 shows that the analogous conductance for MacNeal's boundary nodes goes to infinity as h goes to infinity. Thus, the minimum norm M of the Y/A matrix does not become large, and, when only modified adjacent nodes are used, the minimum norm of the Y/A matrix approaches a constant as h becomes large and does not give

restrictive stability conditions. Moreover, even for an infinite heat-transfer coefficient no heat capacity of the solid is neglected. The conductance in equation V-91 goes to zero when h is zero, and thus, the modified adjacent nodes are suitable for adiabatic boundaries.

If the temperature of the solid boundary opposite the ith adjacent node must be known it can be found from the following equation, which is derived using thermal resistances.

$$t^{*}_{i,n} = \frac{\frac{h \ell_{i,f_i}}{k} t_{f_i,n} + t_{i,n}}{\frac{h \ell_{i,f_i}}{k} + 1} \qquad \text{(V-92)}$$

where t^{*}_{i} = surface temperature opposite the ith adjacent node.

The asterisk means that this surface temperature is not associated with a heat capacity, and it does not add a degree of freedom to the approximate calculation. Rearranging, this equation results in

$$h(t_{f_i,n} - t^{*}_{i}) = k \left[\frac{t^{*}_{i,n} - t_{i,n}}{\ell_{i,f_i}} \right] \qquad \text{(V-93)}$$

which is a direct discretization of the continuous boundary condition, equation V-42.

It appears that this modified type of adjacent node can be made to follow the boundary exactly and the conductances and capacities are

as easily calculated as for the MacNeal adjacent node. In addition, such nodes meet all the objections raised about the nodes located on the boundary. Specifically, the advantages are:

(1) They give an equation which is a direct discretization of the boundary condition equation for a heat-transfer coefficient or for an interface with a second uniform solid.

(2) They can be used for all values of heat-transfer coefficient from zero to infinity, inclusive, directly without neglecting any of the heat capacity of the solid and, of equal importance, without causing restrictive stability conditions.

In addition to these two advantages, and on the basis of the comparison between methods G and C in Chapter IV, we would expect approximations using these modified nodes to be more accurate than those where the nodes are on the boundary. Even without this expectation of accuracy, the less restrictive stability conditions that result when the adjacent nodes are located away from the boundary would justify their use. This technique is one of the simplest for obtaining a good non-oscillatory solution for many problems without using a very small time increment or a value of γ that is so large that it introduces inaccuracies.

In order to show that a more accurate approximation results from using adjacent nodes which follow the boundary, and which are

located away from the boundary, we would require valid series expansions for the discretization error for these nodes. This has not been done.

Interface Boundary Condition. The location of a MacNeal node on the interface between two uniform solids is shown in Figure V-14. From a generalization of his rules, the dimensionless heat capacity of the interfacial node is

$$A_i = \frac{(\text{Area})_{i\,I}\ \sigma_{II} C_{p\,I} + (\text{Area})_{i\,II}\ \sigma_{II} C_{p\,II}}{\sigma_0 C_{p\,0} L^2} \qquad \text{(V-94)}$$

where the $(\text{Area})_{i\,I}$ is an area within the polygon of perpendicular bisectors of solid I with thermal properties subscripted by I and $(\text{Area})_{i\,II}$ is the corresponding area for solid II. The conductance to other nodes on the interface is assumed to be the sum of two parallel conductances, one in each solid:

$$y_{i\,l} = \frac{r_{i,l\,I}\ k_I + r_{i,l\,II}\ k_{II}}{k_0\ \ell_{i,l}} \qquad \text{(V-95)}$$

where $r_{i,l\,I}$ is the length of perpendicular bisector of $\ell_{i,l}$ within solid I, $r_{i,l\,II}$ is the length of this perpendicular bisector in solid II, and $\ell_{i,l}$ is the distance between the nodes, as usual. The conductances to neighboring nodes wholly within either solid are computed in the usual manner, equation II-26.

This type of adjacent node for an interfacial boundary, like the same type for a fluid boundary, associates a heat capacity with the interfacial temperature, which the continuous formulation does not, and does not give a direct discretization of the boundary equation V-88. Another disadvantage of this type of node can be seen from Figure V-15, which relates the temperature along the normal to such an interface versus the normal distance for two solids with widely different thermal conductivities. Solid I, with the high conductivity, shows a relatively slowly changing temperature with distance, but the temperature in solid II, with a low conductivity, changes very rapidly, particularly at the boundary. In such a situation the temperature of a node at the interface cannot accurately represent the average temperature of its surrounding area, because of the discontinuity of the temperature distribution.

Both of the above objections can be eliminated by locating the nodes away from the interfacial surface. Such a location gives a direct discretization of equation V-88, allows an interfacial temperature to be calculated from thermal resistances (no heat capacity), and each node contains material of uniform thermal properties. A portion of a network for an interfacial boundary is in Figure V-16. The adjacent boundary nodes, i and i', for each solid are located on a common normal ii' to the interface. The conductance between them is found by taking two thermal resistances in series:

$$y_{ii'} = \frac{A_{ii'}}{k_0} \left[\frac{1}{\frac{\ell_{i\,I}}{k_I} + \frac{\ell_{i'\,II}}{k_{II}}} \right] \tag{V-96}$$

where $A_{ii'}$ is the length of interfacial arc bounded by the perpendicular bisectors, $\ell_{i\,I}$ is the distance along the common normal from node i in solid I to the interface, and $\ell_{i'\,II}$ is the corresponding distance for node i' in solid I. The heat capacity for each node, i and i', is based on the area bounded by the perpendicular bisectors and the interface. The conduction to the other neighboring nodes is found in the usual manner. The interfacial temperature, as found by linear interpolation using thermal resistance, is

$$t^*_{ii'} = \frac{\frac{k_I\, t_{i\,I}}{\ell_{i\,I}} + \frac{k_{II}\, t_{i'\,II}}{\ell_{i'\,II}}}{\frac{k_I}{\ell_{i\,I}} + \frac{k_{II}}{\ell_{i\,II}}} \tag{V-97}$$

This expression can be rearranged to give an expression which is a direct discretization of equation V-88.

$$k_I \left[\frac{t_{i\,I} - t^*_{ii'}}{\ell_{i\,I}} \right] = k_{II} \left[\frac{t^*_{ii'} - t_{i'\,II}}{\ell_{i'\,II}} \right] \tag{V-98}$$

Thus, each adjacent node contains only material of uniform thermal properties, and a direct discretization of equation V-88 is obtained.

However, a disadvantage does occur for adjacent nodes that are located away from an interface. In order that the perpendicular bisectors of the network legs between adjacent nodes meet at a common point on the interfacial surface (e.g., point C in Figure V-16), the network surrounding the interface must be of isosceles trapezoids, which are the most general quadrilaterals for which this can occur. (Some networks other than of isosceles trapezoids can be made, in principle, but their construction is more difficult.) This network can be constructed by drawing normals to the interfacial boundary at selected intervals. Then, after an i and corresponding i' are selected for one normal, the rest of the nodes can be located by drawing a line from i to the normal next to it on the boundary, so that the angle of intersection with this normal is the same as with the first normal. A line parallel to this is then drawn for node i'. The nodes on the next normal are thus located, and the trapezoidal network can be constructed surrounding the interface. In closing this network, the position of a normal can be changed by trial and error. If the perpendicular bisectors do not intersect at a common point on the interface, then the series resistances could be multiplied by different arc lengths. This would be equivalent to multiplying the left side of equation V-98 by $A_{i\,I}$ and the right side by $A_{i'\,II}$, where $A_{i\,I}$ is the length of arc bounded by the bisectors for solid I, and $A_{i'\,II}$ is the same for solid II. However, this node location now obviously no longer has the advantage of giving

a direct discretization of the boundary conditions; the advantage of having no discontinuities in temperature gradient within a solid remains.

Additional studies are required to show whether the modified adjacent nodes for interfacial boundaries would improve the accuracy of an approximate solution.

2. Location of Interior Nodes.

The location of the points within the interior of a uniform solid obviously should be done in such a manner that the space discretization error is as small as possible for the number of points to be used. for the particular geometry and fluid temperature functions. The following is a discussion and presentation of rules and guide lines, based on the analytic and numerical studies for arranging the interior nodes. However, because of the wide range of conditions possible in problems of transient temperature distribution, exceptions to the rules do occur. Although many of the exceptions are mentioned, it is not possible to anticipate all possible situations. Consequently, these rules should be used as guide lines but not as rigid rules.

Nodes. The series expansions in Table V-9, based on Cartesian coordinates, show that any asymmetry or irregularity in dimension of an interior node probably increases the space discretization error. Asymmetry which consists in having different angles or different lengths to neighboring nodes usually results in an increase in the

coefficient of a low-order term in the series or the addition of a new lower-order term to the series. (Although at times this might produce a compensating effect, it is not assured, and is not probable.) Thus, the interior nodes should be selected from those which have only small coefficients for the second-order terms.

The best shapes are the square node E and the hexagonal node N shown in Table V-9. Both of these nodes are located at the center of the surrounding area and have equal branch lengths separated by equal angles. Nodes which have equal branch lengths in opposite directions, but unequal lengths in the two perpendicular directions, such as the rectangular node D, also are located at the centroid, and are particularly useful in problems which have much larger gradients in one direction. They can then be oriented with their short dimension parallel to the direction of highest gradient. Among other nodes like the rectangle, which have not been studied, but which should have an error expansion like the rectangle's, is a distorted hexagon (like node N) with the network shape of isosceles triangles instead of equilateral triangles. Both the irregular rectangular node C and the trapezoidal node I are not located in the center of the surrounding area; however, their error terms would still be expected to be small. The irregular rectangle is often useful in fitting irregular boundaries; the trapezoidal node is useful for radial geometry. Further, small errors probably also result from nodes that deviate only slightly, in angles or in lengths, from the type

of hexagonal nodes (node N) which form networks of equilateral triangles.

Networks. The networks in which these nodes are to be used are also restricted. Assuming a uniform solid, where the temperature gradients are fairly uniform throughout, and where equal accuracy is required throughout, several characteristics of a good network should be pointed out. The areas of the nodes should be about equal. The conductances throughout also should all be about the same. This then keeps the ratio of conductance to capacitance from being too large at any one point. The diagonal elements in the Y/A matrix are about the same size, which keeps the minimum norm low, and thus a severe stability restriction does not result. The requirement of equal areas and equal branch lengths also dictates that the network be very regular in the interior. That is, the interior nodes should approach or be the rectangular node D or the hexagonal node N.

However, a network that does not meet the requirements for equal area, equal conductance, and equal ratios y_{ij}/A_i is the trapezoidal network of a radial network of nodes bounded by arcs and radii as shown in Figure V-7. These networks are useful for many problems in radial symmetry. In this case MacNeal's trapezoids give an approximation to a direct discretization of the Laplacian operator written in the cylindrical coordinate system. The network is thus distorted based on a relationship with the boundary which is conveniently represented in

the cylindrical coordinate system. Further with this network the nodal heat capacity A_i goes to the differential volume element dV in cylindrical coordinates as the network is refined. Thus, because this makes the several finite summations for the difference solution more like the integrations for the continuous solution, it might be expected that this type of a network would have eigenvectors c_j, eigenvalues λ_j, and initial vector g that are closer to those of the continuous solution for many problems than those obtainable when a rectangular mesh is used in the interior of the circle.

Although for a problem in cylindrical symmetry a convenient orthogonal curvilinear system based on the boundary is known, for problems with a general curved boundary the distortions required for a curvilinear network in the interior would have to be estimated from the boundary geometry. Also, because the isosceles trapezoid network shape is the most general quadrilateral allowed by MacNeal's rules, networks for geometries other than cylindrical would have to be based on a distortion of a network of equilateral triangles. It would be hoped that a series expansion based on the curvilinear coordinates (not Cartesian) would show that such nodes give a high-order approximation to the Laplacian in that coordinate system. This expansion probably could not be carried out.

Further, more than one orthogonal curvilinear coordinate system can be used to describe the boundary. Thus, rather than use an

obvious coordinate system, based only on the geometry of the boundary, one might select a network distorted according to an orthogonal curvilinear system that would follow or be parallel to the streamlines and iso-potential lines. This would give a very fine network in regions of high gradients, as desired. However, in a transient problem these lines obviously can be functions of time, and only an average grid could be used. For problems with sufficient symmetry so that the streamlines are not a function of time it might be possible, in principle, to eliminate a space dimension, as the use of cylindrical coordinates does for some problems; however, it is highly doubtful that one can take advantage of this for most problems with irregular boundaries. Thus, for most problems, if a distorted network is to be used, it is probably simplest to use a distortion based on the curvilinear coordinate system most convenient for the geometry of the boundary.

Consequently, there are two types of networks which use the best nodes and can give good approximate solutions. In both the adjacent nodes follow the boundary exactly but are centered a short distance from the boundary. In the first network, the interior points are located in as regular a manner as possible. All the node shapes approach as closely as possible either rectangles or regular hexagons. The larger the number of points or the more regular the boundary, the closer this approach can be made. In the second type of network the interior nodes are deliberately distorted in a pattern related to the

boundary or an expected temperature distribution. Examples of both types of network applied to a thick C-shaped solid are shown in Figures V-17 and V-18.

In Figure V-17 the adjacent nodes are located on straight lines that are approximately parallel to the curved boundary. This allows square nodes to be used throughout the interior of the solids; only the adjacent nodes have any irregularity, and only they require special calculation for conductances and capacitances. In addition to this advantage, the network is very simple to lay out, and the expected space discretization error is small for the nodes. Also it should be mentioned that this network shows that, by locating the adjacent points away from the boundary, and only approximately parallel to it, one can make the interior network quite regular even with a limited number of points.

The actual network shown contains 30 points. It can be made considerably coarser without destroying the description of the boundary by removing each alternate row of points or possibly only the center row. (Some small adjustments in point location would be required to maintain a good description of the boundary.) This would make the interior points rectangles, which for a problem with a constant initial temperature and a different constant boundary temperature, might give almost as good an approximate solution. This would be so because the higher gradients would occur in the horizontal direction more than in the vertical direction. Thus, the removal of the alternate rows means that

the solid can be described very well with only 18 points and a regular rectangular network of nodes can be used in the interior.

In Figure V-18 a triangular network is used on the same cross-section. This network is an attempt to distort a network of equilateral triangles (node N) for this particular geometry. This network contains 26 points which is about a 13 per cent saving in points, but each point is asymmetric and the Y/A matrix must be computed for each point.

The triangles are constructed so that they should be approximations to curvilinear triangles for an orthogonal curvilinear coordinate system convenient for the particular shape; and it would be hoped that a series expansion for the space discretization errors in these curvilinear coordinates would show that the error should be small. In general the distortion of the hexagonal nodes in the central section is what would be expected for the particular solid. In the region inside of the "prongs" of the thick C-shape, the geometry indicates that the nodes should be very small, much as the nodes become small at the center of a circle. Since the network is to be coarse, these nodes were combined to make all the nodes about equal in size. Further, the specific network used does not appear to describe the "prongs" as satisfactorily as does the rectangular network in Figure V-17.

Consequently, as the series expansions in the curvilinear coordinates cannot be readily found, there is no assurance that this network can give better or even equivalent results to a rectangular

network. Moreover, in addition to requiring a special calculation for each node to find the conductances and capacitances, this network is not as easily constructed as the rectangular network. Such a construction involves several trial and error network constructions until the location looks satisfactory, and the restrictions are satisfied. Thus, practical considerations of laying out the network and individual calculations for each point limit the use of nodes which are distorted in relationship to a curvilinear coordinate system to shapes where the convenient curvilinear coordinates are well known. Any improvement in accuracy using such an approximation could be determined by comparing, for the circular cross-section, a continuous solution with approximations using a direct discretization of the Laplacian operator in cylindrical coordinate systems with nodes bounded by arcs and radii, MacNeal trapezoids, a square network, and a network of equilateral triangles. This has not been done.

Other Rules. In addition to the discussion above about the type of node and the general network, several other points should be made concerning location of the points.

In regions where very accurate results are desired, or where the second- and higher-order derivatives are large, a small rectangular network or a network of equilateral triangles should be used. This is true even in the distorted network, as, when these nodes are refined, they tend to become rectangular. This finer mesh should also extend

away from the region of interest far enough so that any errors in the nodes where the network size is changed, and beyond, should not affect the solution at the particular nodes of interest. Since nodes used to change the network size often have zero-order terms in their expansions, they should be located in a region of low second derivatives if possible. If not too many nodes are required in changing the mesh size, equations V-72 to V-76 might be solved for each of the nodes to improve the accuracy. The space discretization error of any node much larger than the other nodes should be investigated, even though it might be some distance from any region of interest. This is true because the discretization error for this node is weighted so much more, and can cause errors in the solution for nodes with small discretization error located some distance away.

A converse of the above rule is that, in regions where very accurate approximations are not needed, and where the space derivatives are small, a coarse network can be used. Of course, both this region and that where the network size is changed should not be close to nodes where high accuracy is required. An example of such a region, where a coarse network can be safely used, is the region in the center of a circle when the conditions of the problem are such that the derivatives with respect to the angle parameter are small. In this case relatively large square nodes could be safely used without affecting the accuracy of the solution at the surface.

A corollary of the above remarks is that, for problems where the temperature differences are larger in one direction than in the perpendicular direction, a rectangular node oriented with its short dimension in the direction of the larger gradient is appropriate. This gives equivalently a small discretization error for each node without using a square node with a short dimension in both directions. Thus, the rule about having equidistant nodes can be modified under the above conditions to mean that the temperature difference between each neighboring node and the ith node should be kept constant. This rule is primarily useful in a solid with one dimension longer than the other, with constant initial temperature, and constant surrounding temperatures.

As mentioned previously, the nodes should be located so that the ratio of conductances to areas does not vary widely. Although this rule is secondary to the above comments about different-sized networks, it is of practical importance not to have one very small node with large conductances. The row or column sum associated with such a node very probably determines the minimum norm as a large number, which gives more restrictive stability and non-oscillatory criteria. Moreover, by keeping the elements of the Y/A matrix relatively in scale the effects of round-off error are somewhat reduced.

A practical application of this rule of having approximately equal ratios of conductance to capacitance is that it indicates for a composite

solid that relatively larger nodes should be used in parts of the solid where the thermal conductivity is very high, or the volumetric heat capacity low. In the extreme it helps one recognize when nodes should be combined. Further, if even the use of a large node still results in high conductance to capacitance ratio, the heat capacity of that region can be neglected. These approximations are justified by physical reasons in the Longwell report (9); however, the above rules tell us when these approximations are satisfactory.

It should be pointed out that if the transients are important in regions of high thermal conductivity, the large conductance to capacitance ratios should be used. Then a small time increment must be used not only because of stability considerations, but primarily because the temperature in such a region changes so rapidly that unless a small time increment is used the transients in that region cannot be observed. This indicates that there is no justification for including the heat capacity for such nodes, and then using the inaccurate weighting γ of 1 to allow use of a large time increment.

3. Summary--Location of Nodes.

The main rule for the location of nodes where a relatively coarse network is to be used, and where the limiting number of points is known, is, for adjacent nodes:

The adjacent nodes should not be placed on the boundary, but should be located on a line approximately parallel to the surface about

half a branch length inside the surface. The solid surface forms one of the boundaries for the node . This applies both to external boundaries in contact with a fluid with a heat-transfer coefficient of 0 to infinity, inclusive, and to interfacial boundaries of a composite solid. The advantages of this type of adjacent node used to describe the boundary, over nodes located on the surface, are:

(1) This type of adjacent node can be used for all values of heat-transfer coefficients, without neglecting any heat capacity of the solid and, more importantly, without giving a severe stability restriction.

(2) An equation can be derived that shows that this type of adjacent node gives a direct discretization of the continuous boundary conditions.

(3) This somewhat flexible location of the adjacent nodes allows the internal network to be more regular, and thus it can approach more closely or be a rectangular network or a network of equilateral triangles.

The main rule for the interior nodes is:

The interior nodes should be located so that the interior network is as regular as possible. That is, each node should be as close to either the rectangular node D or the hexagonal node N, as allowed by the number of points and the boundary. The only exception to the above rule is that, if the boundary can be described in a well known curvilinear coordinate system, the nodes can be distorted as dictated by the

Laplacian operator written in those coordinates. The advantages of a regular internal network are:

(1) There are no zero-order error terms and the total discretization error should be small.

(2) The network can be relatively easily constructed, and the conductances and capacities easily found.

Two other subsidiary rules for locating the nodes are, first, a finer network of more regular nodes should be used in regions of high gradients or in regions of special interest. The corollary to this rule is that a smaller branch length should be used in directions with the largest temperature changes. Second, the ratio of conductance to capacitance should be kept about constant; nodes which have high ratios can be combined, and, if necessary, their heat capacity can be neglected.

The actual location of the adjacent nodes also follows the same general rules as indicated above for the interior points. Indeed, in actually constructing a network, obviously both the interior and adjacent nodes must be laid out on the basis of the above rules, and taking into consideration the interrelationship of their positions.

(H) SUMMARY AND CONCLUSIONS

A study has been made of the accuracy of approximate solutions for transient temperature distributions in irregular solids based on an asymmetric network of points. In this investigation the relationship between the error in the approximate solution and the discretization error, caused by replacing the continuous operators with differences, was derived as the solution of the difference equation for the propagation of the error. The discretization error caused by both the time and space differencing was studied using series expansions. The validity and usefulness of these relationships were shown by explaining the errors in numerical results for both a transient and a steady-state problem, with only a superficial knowledge of the derivatives. The important conclusions about the effects of asymmetry on the discretization error and the rules for locating the nodes and selecting the time differencing parameters are summarized below.

The series expansion based on Cartesian coordinates for the discretization error associated with the space differencing for a general asymmetric node is different from those for regular rectangular or equilateral hexagonal nodes, in that terms of lower order occur in the series for the asymmetric nodes. That is, the expansion for an asymmetric node usually contains a zero-order error term that depends only on the shape of the node and the second space derivatives, and

does not depend upon the size of the node. This is in contrast to the expansion for regular nodes mentioned above, which has as its first term in the series a second-order term that is proportional to the area of the node and the fourth space derivatives. Consequently, an approximation based on a network containing asymmetric nodes is not consistent in the Richtmyer sense, and the approximate solution cannot be made arbitrarily accurate by the addition of more points and using smaller time increments, if each succeeding network contains asymmetric nodes geometrically similar to those in the original network. However, the consistency condition can be satisfied as the network of nodes can be made regular by changing the shape of the asymmetric nodes during the refinement. Thus, the consistency condition has practical significance only for very fine networks, and it means that when a great many points can be used, they should be located in a regular network to take advantage of the small discretization error of small regular nodes. This is also dictated by practical considerations, first, that the boundaries of even an irregular solid can be followed closely with a fine regular network, and second, that the conductances and capacities are very easily determined for such networks.

However, for a relatively coarse network the size of the discretization error is important, and not how the error changes with the size of the node. Thus, the zero-order terms in the expansion have no special significance. In these cases the guide line summarized in a previous section (G-3) should be used.

Briefly, these rules are: (1)The adjacent nodes should be located not on the boundary but on a line approximately parallel to the boundary, about a half branch length from the boundary. One boundary of the nodal area coincides with the solid surface. The reasons for this are that a less severe stability restriction results, that the interior nodes can be made more regular with the same number of points, and that higher accuracy is expected because there is a better approximation to the continuous boundary equation. The spacing and location of the adjacent nodes also should follow the next several rules for the interior nodes and should allow the interior nodes to follow them.

(2) The interior network should be made as regular as possible with the nodes having as small a deviation as possible from regular rectangles or regular hexagons; any distortions should probably have a relationship to the boundary or to the expected streamlines for heat flow. All nodes should have about the same area associated with them and about equal conductances to their neighbors, with the following exceptions, listed as (3) and (4).

(3) If possible, the average branch length should be shorter in regions where a highly accurate solution is required or where the temperature gradients are large. A smaller branch length should be used in the direction of the largest gradients. For regions where the accuracy of the solution is not important and the gradients are small, a coarser network is satisfactory.

(4) The ratio of conductances to capacitances should be kept about constant to avoid having a severe stability restriction. This rule indicates, in the absence of other considerations, when nodes should be combined or when the heat capacity of a node should be neglected, and is particularly useful for composite solids of widely different thermal properties.

The selection of the time differencing parameters $\Delta\tau$ and γ for a fixed network of points is based on the stability or non-oscillatory criteria. The accuracy requirements for most problems are probably satisfied if the weighting factor γ is equal to or less than $\frac{1}{2}$. In section F-2 a procedure was shown to allow the selection of differencing parameters that would keep the calculations close to a minimum and that would give negligible oscillations. However, the numerical transient solution indicates that the continuous damping factors cannot always be bracketed by changing the time increment and the weighting factor γ, but that the inaccuracy caused by using high weighting factors is more severe for an asymmetric network than for regular networks.

Consequently, for many solids with curved or irregular boundaries, approximate solutions of adequate engineering accuracy can be obtained using relatively coarse networks of less than 30 points. This can be done by locating the points according to the above rules by following the boundary exactly and tolerating, but minimizing, any required asymmetry in the interior.

(I) SUGGESTIONS FOR FUTURE WORK

Of the many possible further investigations of approximate solutions to the diffusion equation, the following two projects show promise of giving results of important practical significance.

First, a valid expression or expansion should be developed for the discretization error for adjacent nodes with a heat-transfer coefficient conductance to a fluid. This would allow a firm conclusion to be made about any improvement in accuracy that occurs when these adjacent nodes are located away from the boundaries. Preliminary attempts to use a Taylor series to find such an expansion have not been successful even for one-dimensional problems, probably because the difference equation was not compared to a proper combination of the boundary-condition equation and differential equation. However, future studies based on the Taylor expansion or other method should give a useful expression for this error.

Second, a relation should be developed between the error propagation analysis, used in this chapter for the asymmetric network, and the error analysis based on the complete difference solution, used in Chapter IV for regular one-dimensional networks. This would involve setting the error as given in equations V-13, V-19, or V-22 equal to the error given in equation IV-209, or other appropriate equation in Chapter IV. Such a study should allow conclusions to be made about

how the discretization error affects the error in each part of the difference solution. That is, one might be able to make precise conclusions about how the discretization error affects the error in the particular solution, damping factors, eigenvectors, and initial vector of the analytic expression of the difference solution. This would then allow the conclusions from the series expansions for the discretization error to be applied quantitatively to the error in these parts of the solutions. Since the discretization error expansions are more easily obtained than are the complete difference and continuous solutions for very general problems, this would allow a precise quantitative error analysis for very complicated problems based only on the discretization error.

Three other suggestions that are of somewhat less importance are:

(1) The series expansions in Cartesian coordinates for the space discretization errors for several other node shapes should be derived. Among the node shapes that should be studied are (a) a node associated with an equilateral pentagonal area. The node is surrounded by five neighboring nodes, each equidistant from the node and from each other. (b) A node associated with an equilateral heptagonal area, surrounded by seven neighboring nodes, each equidistant from the node and from each other. Both of these nodes should have sufficient symmetry to give a good approximation of the Laplacian operator. Although

a network cannot be constructed merely from one of these two types of nodes, they might be useful in other triangular networks.

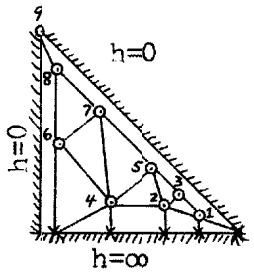
Another type of node that should be further studied is the hexagonal node. A sequence of irregular hexagons can be studied, like the sequence of irregular quadrilateral nodes shown on Table V-9, nodes C to I, which can be arranged to progress from a square to trapezoidal nodes.

(2) An investigation should be made of the elimination of the zero-order error terms in the space discretization error expansions by solving a system of equations. Equations V-72 to V-76, when solved for the ith node, give a corresponding row in the Y/A matrix. Equations V-77 to V-80, when solved, would give the set of lengths of p'_{ij} for the ith node that could be used to compute the elements in the ith row of the Y/A matrix. The specific studies required to make this into a useful technique for the solution of the diffusion equation are mentioned in section E-3. The results of such a study might give weightings to use to eliminate the zero-order error term from the expansion for irregular nodes used to change the network branch length. Further, it would be hoped that a set of geometric conditions of a node and its neighbors would result under which satisfactory solutions could be found to these equations, so that the zero-order terms could be eliminated from irregular interior nodes close to the boundaries.

(3) A study should be made of the distortions of both the rectangular network and the network of equilateral triangles according to a curvilinear coordinate system based on either the boundary or an expected temperature distribution. The first part of such a study should be to determine if any improvement in accuracy occurs when such a curvilinear, or approximation to a curvilinear, network is used. This could be done by comparing the numerical results for the several possible networks for problems in circular symmetry with the continuous solution. If an analytic solution can be found for the one-dimensional partial difference equation of diffusion in cylindrical radial coordinates with z-transforms some analytic comparisons would be possible. If the cylindrical coordinate network or MacNeal's approximation to this network does give significantly more accurate approximations when using the same number of points, then the rules for distorting both rectangular networks and networks of equilateral triangles for a general curved boundary or expected temperature distribution must be found so that such a distorted network would give a good approximation for a particular geometric boundary. This study could then result in finding a set of rules for locating temperature points in an asymmetric fashion throughout the solid in such a way that a regular mesh might require more points to give the same accuracy.

Table V-1

CONDUCTANCE MATRIX (Y) AND INVERSE CONDUCTANCE MATRIX (Y^{-1})
First Row: Elements in Y^{-1} Matrix; Second Row: Elements in Y Matrix
Transient Problem, Chapter III

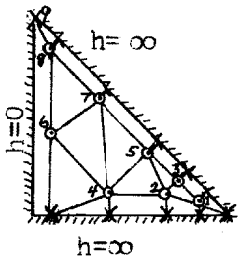


	1	2	3	4	5	6	7	8	9
1	-0.24804	-0.05029	-0.08799	-0.01096	-0.03693	-0.01435	-0.02073	-0.01798	-0.01800
	-4.332	0.1703	0.750	0	0	0	0	0	0
2	-0.05029	-0.29282	-0.22402	-0.04860	-0.14142	-0.05726	-0.08127	-0.07098	-0.07098
	0.1703	-5.653	2.500	0.3333	0.500	0	0	0	0
3	-0.08799	-0.22402	-0.45738	-0.05229	-0.18117	-0.06988	-0.10126	-0.08782	-0.08776
	0.75	2.50	-3.755	0	0.5046	0	0	0	0
4	-0.01096	-0.04860	-0.05229	-0.22837	-0.13203	-0.11752	-0.12830	-0.12368	-0.12368
	0	0.3333	0	-5.230	0.8748	0.40	0.1219	0	0
5	-0.03693	-0.14142	-0.18117	-0.13203	-0.59256	-0.21491	-0.32005	-0.27497	-0.27497
	0	0.500	0.5046	0.8748	-2.480	0	0.6001	0	0
6	-0.01435	-0.05126	-0.06988	-0.11752	-0.21483	-0.83279	-0.61037	-0.70563	-0.70566
	0	0	0	0.4000	0	-2.3981	1.123	0.375	0
7	-0.02073	-0.08127	-0.10126	-0.12830	-0.32005	-0.61037	-0.98275	-0.82312	-0.82312
	0	0	0	0.1219	0.6001	1.123	-2.345	0.4999	0
8	-0.01798	-0.07098	-0.08782	-0.12368	-0.27497	-0.70563	-0.82312	-1.91569	-1.91569
	0	0	0	0	0	0.375	0.4999	-1.2909	0.416
9	-0.01800	-0.07098	-0.08782	-0.12368	-0.27497	-0.70566	-0.82312	-1.91569	-4.31943
	0	0	0	0	0	0	0	0.416	-0.416
Column Sums	-0.50527	-1.03164	-1.34963	-0.96543	-2.16898	-3.32837	-3.89097	-5.93556	-8.33929

Column Sums = Row Sums (Y^{-1} Matrix)

Table V-2

CONDUCTANCE MATRIX (Y) AND INVERSE CONDUCTANCE MATRIX (Y^{-1})
First Row: Elements in Y^{-1} Matrix; Second Row: Elements in Y Matrix
Steady-State Problem, Chapter V

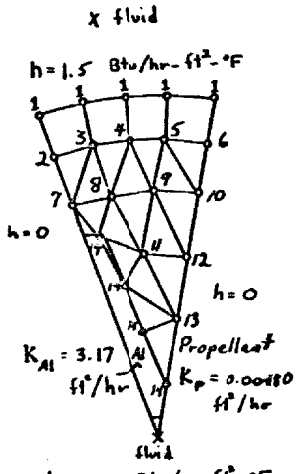


	1	2	3	4	5	6	7	8
1	-0.16802 -6.092	-0.01767 0.170	-0.02752 0.750	-0.00181 0	-0.00379 0	-0.00051 0	-0.00042 0	-0.00008 0
2	-0.01767 0.170	-0.22212 -5.654	-0.09311 2.500	-0.01917 0.333	-0.02729 2.500	-0.00487 0	-0.00334 0	-0.00068 0
3	-0.02752 0.75	-0.09311 2.500	-0.20238 -6.255	-0.01034 0	-0.02459 0.505	-0.00306 0	-0.00268 0	-0.00048 0
4	-0.00181 0	-0.01917 0.333	-0.01034 0	-0.20090 -5.230	-0.03053 0.875	-0.03986 0.400	-0.01220 0.122	-0.00410 0
5	-0.00379 0	-0.02729 0.500	-0.02459 0.505	-0.03053 0.875	-0.16396 -6.480	-0.01301 0	-0.01606 0.600	-0.00251 0
6	-0.00051 0	-0.00487 0	-0.00306 0	-0.03986 0.400	-0.01301 0	-0.46560 -2.398	-0.07573 1.123	-0.04140 0.375
7	-0.00042 0	-0.00334 0	-0.00268 0	-0.01220 0.122	-0.01606 0.600	-0.07573 1.123	-0.15062 -7.345	-0.02021 0.500
8	-0.00008 0	-0.00068 0	-0.00048 0	-0.00410 0	-0.00251 0	-0.04140 0.375	-0.02021 0.500	-0.19985 -5.132
Column Sums	-0.21982	-0.38825	-0.36416	-0.30891	-0.28174	-0.64403	-0.28126	-0.26931

Column Sums = Row Sums (Y^{-1} Matrix)

Table V-3

CONDUCTANCE MATRIX (Y) AND INVERSE CONDUCTANCE MATRIX (Y^{-1})
First Row: Elements in Y^{-1} Matrix; Second Row: Elements in Y Matrix
Shape I, see Longwell (9)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	-0.25311 -6.8524	-0.17448 0.5103	-0.17282 1.0206	-0.17846 1.0206	-0.18624 1.0206	-0.18963 0.5103	-0.06678 0	-0.04174 0	-0.09572 0	-0.11006 0	-0.03033 0	-0.03928 0	-0.00857 0	-0.00358 0
2	-0.17448 0.5103	-0.74874 -2.1333	-0.34304 1.270	-0.20667 0	-0.16757 0	-0.15966 0	-0.20563 0.353	-0.07171 0	-0.10480 0	-0.10976 0	-0.03493 0	-0.04317 0	-0.01143 0	-0.00625 0
3	-0.17282 1.0206	-0.34303 1.270	-0.46857 -4.2196	-0.24943 1.212	-0.18506 0	-0.17153 0	-0.13747 0.124	-0.07721 0.593	-0.11691 0	-0.12101 0	-0.03787 0	-0.04702 0	-0.01148 0	-0.00567 0
4	-0.17846 1.0206	-0.20667 0	-0.24943 1.212	-0.45642 -4.1626	-0.27251 1.212	-0.23295 0	-0.09364 0	-0.07263 0.359	-0.16478 0.359	-0.16938 0	-0.05059 0	-0.06382 0	-0.01394 0	-0.00579 0
5	-0.18624 1.0206	-0.16757 0	-0.18506 0	-0.27251 1.212	-0.51389 -4.2196	-0.39035 1.270	-0.07767 0	-0.06419 0	-0.20647 0.593	-0.24093 0.124	-0.06381 0	-0.08382 0	-0.01707 0	-0.00624 0
6	-0.18963 0.5103	-0.15966 0	-0.17153 0	-0.23295 0	-0.39035 1.270	-0.79705 -2.1333	-0.07361 0	-0.06077 0	-0.20424 0	-0.30547 0.353	-0.06785 0	-0.09499 0	-0.01865 0	-0.00640 0
7	-0.06678 0	-0.20563 0.353	-0.13747 0.124	-0.09364 0	-0.07767 0	-0.07361 0	-0.65158 -1.807	-0.09522 0.862	-0.07431 0	-0.06887 0	-0.03101 0	-0.03493 0	-0.01541 0	-0.01218 0.468
8	-0.04174 0	-0.07171 0	-0.07721 0.593	-0.07261 0.359	-0.06419 0	-0.06077 0	-0.09522 0.862	-0.15565 -3.813	-0.09041 1.145	-0.07600 0	-0.03256 0.097	-0.03680 0	-0.01074 0	-0.00640 0.757
9	-0.09572 0	-0.10480 0	-0.11691 0	-0.16478 0.359	-0.20647 0.593	-0.20424 0	-0.07431 0	-0.09041 1.145	-0.45638 -3.941	-0.35310 1.236	-0.12454 0.556	-0.14983 0.052	-0.03014 0	-0.01039 0
10	-0.11006 0	-0.10976 0	-0.12101 0	-0.16938 0	-0.24093 0.124	-0.30547 0.353	-0.06887 0	-0.07600 0	-0.35310 1.236	-0.82014 -1.993	-0.13665 0	-0.21569 0.280	-0.03893 0	-0.01107 0
11	-0.03033 0	-0.03493 0	-0.03787 0	-0.05059 0	-0.06381 0	-0.06785 0	-0.03101 0	-0.03256 0.097	-0.12454 0.556	-0.13665 0	-0.38229 -4.266	-0.30909 1.650	-0.06718 0.058	-0.02338 1.905
12	-0.03928 0	-0.04317 0	-0.04702 0	-0.06382 0	-0.08382 0	-0.09499 0	-0.03493 0	-0.03680 0	-0.14983 0.052	-0.21569 0.280	-0.30909 1.650	-0.71702 -2.242	-0.11297 0.260	-0.02208 0
13	-0.00857 0	-0.01143 0	-0.01148 0	-0.01394 0	-0.01707 0	-0.01865 0	-0.01541 0	-0.01074 0	-0.03014 0	-0.03893 0	-0.06718 0.058	-0.11297 0.260	-0.49987 -2.171	-0.02804 1.853
14	-0.00358 0	-0.00625 0	-0.00567 0	-0.00579 0	-0.00624 0	-0.00640 0	-0.01218 0.468	-0.00640 0.757	-0.01039 0	-0.01107 0	-0.02338 1.905	-0.02208 0	-0.02804 1.853	-0.02902 -38.150
Column Sums	-1.55000	-2.38783	-2.14509	-2.23103	-2.47581	-2.77314	-1.63831	-0.89204	-2.18203	-2.77704	-1.39207	-1.97052	-0.88443	-0.17648

Column Sums = Row Sums (Y^{-1} Matrix)

Table V-4

CONDUCTANCE MATRIX (Y) AND INVERSE CONDUCTANCE MATRIX (Y^{-1})
First Row: Elements in Y^{-1} Matrix; Second Row: Elements in Y Matrix
Shape II, see Longwell (9)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-0.2834	-0.2455	-0.2440	-0.2389	-0.2249	-0.1867	-0.1960	-0.1937	-0.1871	-0.1792	-0.1233	-0.1219	-0.1186	-0.1157	-0.0120	-0.0016
	-6.8554	0.5103	1.0206	1.0206	1.0206	0.5103	0	0	0	0	0	0	0	0	0	0
2	-0.2455	-0.8930	-0.4896	-0.3379	-0.2691	-0.2118	-0.4475	-0.3649	-0.2932	-0.2632	-0.2350	-0.2232	-0.2050	-0.1948	-0.0222	-0.0027
	0.5103	-2.1333	1.270	0	0	0	0.353	0	0	0	0	0	0	0	0	0
3	-0.2440	-0.4896	-0.6166	-0.3812	-0.2860	-0.2211	-0.3878	-0.3727	-0.3056	-0.2739	-0.2271	-0.2213	-0.2065	-0.1972	-0.0218	-0.0028
	1.0206	1.270	-4.2196	1.212	0	0	0.124	0.593	0	0	0	0	0	0	0	0
4	-0.2389	-0.3379	-0.3812	-0.5713	-0.3548	-0.2568	-0.3252	-0.3419	-0.3326	-0.3018	-0.2129	-0.2122	-0.2068	-0.2006	-0.0209	-0.0028
	1.0206	0	1.212	-4.1626	1.212	0	0	0.359	0.359	0	0	0	0	0	0	0
5	-0.2249	-0.2691	-0.2860	-0.3548	-0.5575	-0.3558	-0.2721	-0.2899	-0.3365	-0.3338	-0.1924	-0.1945	-0.1983	-0.1974	-0.0192	-0.0026
	1.0206	0	0	1.212	-4.2196	1.270	0	0	0.593	0.124	0	0	0	0	0	0
6	-0.1867	-0.2118	-0.2211	-0.2568	-0.3558	-0.6287	-0.2148	-0.2285	-0.2712	-0.3251	-0.1563	-0.1588	-0.1647	-0.1694	-0.0157	-0.0022
	0.5103	0	0	0	1.270	-2.6436	0	0	0	0.353	0	0	0	0	0	0
7	-0.1960	-0.4475	-0.3878	-0.3252	-0.2721	-0.2148	-1.0260	-0.5843	-0.4018	-0.3462	-0.4250	-0.3762	-0.3245	-0.3004	-0.0379	-0.0044
	0	0.353	0.124	0	0	0	-1.997	1.235	0	0	0.285	0	0	0	0	0
8	-0.1937	-0.3649	-0.3727	-0.3419	-0.2899	-0.2285	-0.5843	-0.7162	-0.4604	-0.3879	-0.3768	-0.3746	-0.3355	-0.3139	-0.0366	-0.0045
	0	0	0.593	0.359	0	0	1.235	-3.947	1.19	0	0	0.57	0	0	0	0
9	-0.1871	-0.2932	-0.3056	-0.3326	-0.3365	-0.2712	-0.4018	-0.4604	-0.6969	-0.5499	-0.3245	-0.3353	-0.3588	-0.3511	-0.0330	-0.0047
	0	0	0	0.359	0.593	0	0	1.19	-3.947	1.235	0	0	0.57	0	0	0
10	-0.1792	-0.2632	-0.2739	-0.3018	-0.3338	-0.3251	-0.3462	-0.3879	-0.5499	-0.9748	-0.2998	-0.3129	-0.3485	-0.3909	-0.0309	-0.0048
	0	0	0	0	0.124	0.353	0	0	1.235	-1.997	0	0	0	0.285	0	0
11	-0.1233	-0.2350	-0.2271	-0.2129	-0.1924	-0.1563	-0.4250	-0.3768	-0.3245	-0.2998	-0.9553	-0.6405	-0.4765	-0.4174	-0.0703	-0.0066
	0	0	0	0	0	0	0.285	0	0	0	-2.694	2.25	0	0	0.159	0
12	-0.1219	-0.2232	-0.2213	-0.2122	-0.1945	-0.1588	-0.3762	-0.3746	-0.3353	-0.3129	-0.6405	-0.7146	-0.5257	-0.4585	-0.0661	-0.0070
	0	0	0	0	0	0	0	0.57	0	0	2.25	-5.401	2.25	0	0.331	0
13	-0.1186	-0.2050	-0.2065	-0.2068	-0.1983	-0.1647	-0.3245	-0.3355	-0.3588	-0.3485	-0.4765	-0.5257	-0.6928	-0.5968	-0.0523	-0.0085
	0	0	0	0	0	0	0	0	0.57	0	0	2.25	-5.194	2.02	0.05	0.304
14	-0.1157	-0.1948	-0.1972	-0.2006	-0.1974	-0.1694	-0.3004	-0.3139	-0.3511	-0.3909	-0.4174	-0.4585	-0.5968	-0.9224	-0.0459	-0.0092
	0	0	0	0	0	0	0	0	0	0.285	0	0	2.02	-2.514	0	0.209
15	-0.0120	-0.0222	-0.0218	-0.0209	-0.0192	-0.0157	-0.0379	-0.0366	-0.0330	-0.0309	-0.0703	-0.0663	-0.0523	-0.0459	-0.1855	-0.0047
	0	0	0	0	0	0	0	0	0	0	0.159	0.331	0.05	0	-5.6088	1.0219
16	-0.0016	-0.0027	-0.0028	-0.0028	-0.0026	-0.0022	-0.0044	-0.0045	-0.0047	-0.0048	-0.0066	-0.0070	-0.0085	-0.0092	-0.0047	-0.0220
	0	0	0	0	0	0	0	0	0	0	0	0	0.304	0.209	1.0219	-45.9019
Column Sums	-2.6725	-4.6986	-4.4552	-4.2986	-4.0848	-3.5676	-5.6701	-5.3823	-5.2426	-5.3236	-5.1397	-4.9435	-4.8198	-4.8816	-0.6752	-0.0911

Column Sums = Row Sums (Y^{-1} Matrix)

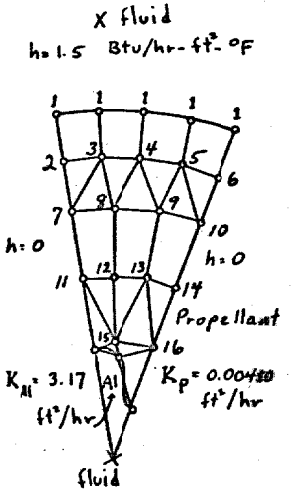
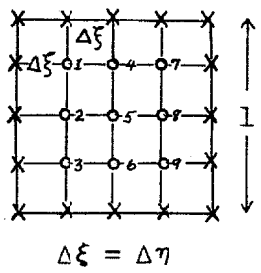


Table V-5

CONDUCTANCE MATRIX (Y) AND INVERSE CONDUCTANCE MATRIX (Y^{-1})
First Row: Elements in Y^{-1} Matrix; Second Row: Elements in Y Matrix
Square Grid - Specified Surface Temperature

	1	2	3	4	5	6	7	8	9
1	-0.29911	-0.09821	-0.03125	-0.09821	-0.06250	-0.02679	-0.03125	-0.02679	-0.01339
	-4.0	1.0	0	1.0	0	0	0	0	0
2	-0.09821	-0.33036	-0.09821	-0.06250	-0.12500	-0.06250	-0.02679	-0.04464	-0.02679
	1.0	-4.0	1.0	0	1.0	0	0	0	0
3	-0.03125	-0.09821	-0.29911	-0.02679	-0.06250	-0.09821	-0.01339	-0.02679	-0.03125
	0	1.0	-4.0	0	0	1.0	0	0	0
4	-0.09821	-0.06250	-0.02679	-0.33036	-0.12500	-0.04464	-0.09821	-0.06250	-0.02679
	1.0	0	0	-4.0	1.0	0	1.0	0	0
5	-0.06250	-0.12500	-0.06250	-0.12500	-0.37500	-0.12500	-0.06250	-0.12500	-0.06250
	0	1.0	0	1.0	-4.0	1.0	0	1.0	0
6	-0.02679	-0.06250	-0.09821	-0.04464	-0.12500	-0.33036	-0.02679	-0.06250	-0.09821
	0	0	1.0	0	1.0	-4.0	0	0	1.0
7	-0.03125	-0.02679	-0.01339	-0.09821	-0.06250	-0.02679	-0.29911	-0.09821	-0.03125
	0	0	0	1.0	0	0	-4.0	1.0	0
8	-0.02679	-0.04464	-0.02679	-0.06250	-0.12500	-0.06250	-0.09821	-0.33036	-0.09821
	0	0	0	0	1.0	0	1.0	-4.0	1.0
9	-0.01339	-0.02679	-0.03125	-0.02679	-0.06250	-0.09821	-0.03125	-0.09821	-0.29911
	0	0	0	0	0	1.0	0	1.0	-4.0
Column Sums	-0.68750	-0.87500	-0.68750	-0.87500	-1.12500	-0.87500	-0.68750	-0.87500	-0.68750

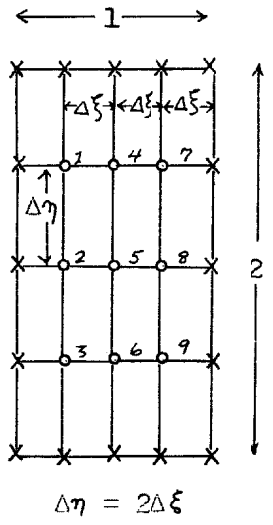
Column Sums = Row Sums (Y^{-1} Matrix)



$\Delta\xi = \Delta\eta$

Table V-5

CONDUCTANCE MATRIX (Y) AND INVERSE CONDUCTANCE MATRIX (Y^{-1})
First Row: Elements in Y^{-1} Matrix; Second Row: Elements in Y Matrix
Regular Rectangular Grid - Specified Surface Temperature



$\Delta\eta = 2\Delta\xi$

	1	2	3	4	5	6	7	8	9
1	-0.25504	-0.04191	-0.00798	-0.12712	-0.03902	-0.00947	-0.05300	-0.02150	-0.00594
	-5.0	0.50	0	2.0	0	0	0	0	0
2	-0.04191	-0.26302	-0.04191	-0.03902	-0.13659	-0.03902	-0.02150	-0.05893	-0.02150
	0.50	-5.0	0.50	0	2.0	0	0	0	0
3	-0.00798	-0.04191	-0.25504	-0.00947	-0.03902	-0.12712	-0.00594	-0.02150	-0.05300
	0	0.50	-5.0	0	0	2.0	0	0	0
4	-0.12712	-0.03902	-0.00947	-0.30803	-0.06341	-0.01392	-0.12712	-0.03902	-0.00947
	2.0	0	0	-5.0	0.50	0	2.0	0	0
5	-0.03902	-0.13659	-0.03902	-0.06341	-0.32195	-0.06341	-0.03902	-0.13659	-0.03902
	0	2.0	0	0.50	-5.0	0.50	0	2.0	0
6	-0.00947	-0.03902	-0.12712	-0.01392	-0.06341	-0.30803	-0.00947	-0.03902	-0.12712
	0	0	2.0	0	0.50	-5.0	0	0	2.0
7	-0.05300	-0.02150	-0.00594	-0.12712	-0.03902	-0.00947	-0.25504	-0.04191	-0.00798
	0	0	0	2.0	0	0	-5.0	0.50	0
8	-0.02150	-0.05893	-0.02150	-0.03902	-0.13659	-0.03902	-0.04191	-0.26302	-0.04191
	0	0	0	0	2.0	0	0.50	-5.0	0.50
9	-0.00594	-0.02150	-0.05300	-0.00947	-0.03902	-0.12712	-0.00798	-0.04191	-0.25504
	0	0	0	0	0	2.0	0	0.50	-5.0
Column Sums	-0.56098	-0.66341	-0.56098	-0.73659	-0.87805	-0.73659	-0.56098	-0.66341	-0.56098

Column Sums = Row Sums (Y^{-1} Matrix)

Table V-7

STEADY-STATE STUDY, SQUARE SHAPE

Continuous, Approximate Solutions, Errors, and Discretization Error

	Node i	1	2	3	4	5	6	7	8
(1)	Continuous Solution, $T_{i,SS}$	0.7953	0.7947	0.6557	0.8909	0.5838	0.6900	0.5451	0.5164
	Approximate Solution, $t_{i,SS}$	0.7886	0.8015	0.6626	0.8850	0.5920	0.6922	0.5445	0.5184
	Error, $v_{i,SS} = T_{i,SS} - t_{i,SS}$	0.0067	-0.0068	-0.0069	0.0059	-0.0032	-0.0022	0.0006	-0.0020
	Area-Weighted Discretization Error, $A_i d_{i,SS}$	-0.0468	0.0229	0.0293	-0.0359	0.0198	0.0075	-0.0084	0.0095
	Discretization Error, $d_{i,SS}$	-13.69	3.952	6.350	-2.132	1.364	0.3102	-0.4373	0.6426
(2)	Approximate Solution	0.7988	0.8026	0.6642	0.8851	0.5923	0.6922	0.5446	0.5184
	Error	-0.0035	-0.0079	-0.0085	0.0058	-0.0035	-0.0022	0.0005	-0.0020
	Area-Weighted Discretization Error	0.0136	As	In	(1)				
(3)	Approximate Solution	0.8089	0.8037	0.6659	0.8852	0.5925	0.6923	0.5446	0.5184
	Error	-0.0136	-0.0090	-0.0102	0.0057	-0.0037	0.0023	0.0005	0.0020
	Area-Weighted Discretization Error	0.0740	As	In	(1)				

Approximate Calculation (1) Uses $t_B = T_B$

(2) Uses $t_B = T_B$ except for node at A which is used as 3/4 instead of 1/2

(3) Uses $t_B = T_B$ except for node A which is taken as 1 instead of 1/2

Table V-8

TAYLOR SERIES EXPANSION

$(T_j - T_i)$

$$\frac{\rho_{ij}}{\lambda_{ij}}(T_j - T_i) = \frac{\rho_{ij}}{\lambda_{ij}} \sum_{p=1}^{\infty} \frac{1}{p!}\left[\lambda_{ij}\left\{\cos\beta_{ij}\frac{\partial}{\partial \xi} + \sin\beta_{ij}\frac{\partial}{\partial \eta}\right\}\right]^p T_{i,n}$$

p	Terms
1	$\rho_{ij}\left[\cos\beta_{ij}\frac{\partial T_{i,n}}{\partial \xi} + \sin\beta_{ij}\frac{\partial T_{i,n}}{\partial \eta}\right]$
2	$\frac{\rho_{ij}\lambda_{ij}}{2}\left[(\cos\beta_{ij})^2\frac{^2T_{i,n}}{\xi^2} + 2\cos\beta_{ij}\sin\beta_{ij}\frac{^2T_{i,n}}{\xi\ \eta} + (\sin\beta_{ij})^2\frac{^2T_{i,n}}{\eta^2}\right]$
3	$\frac{\rho_{ij}(\lambda_{ij})^2}{6}\left[(\cos\beta_{ij})^3\frac{\partial^3 T_{i,n}}{\partial \xi^3} + 3(\cos\beta_{ij})^2(\sin\beta_{ij})\frac{\partial^3 T_{i,n}}{\partial \xi^2 \partial \eta^2}\right.$ $\left. + 3(\cos\beta_{ij})(\sin\beta_{ij})^2\frac{\partial^3 T_{i,n}}{\partial \xi \partial \eta^2} + (\sin\beta_{ij})^3\frac{\partial^3 T_{i,n}}{\partial \eta^3}\right]$
4	$\frac{\rho_{ij}(\lambda_{ij})^3}{24}\left[(\cos\beta_{ij})^4\frac{\partial^4 T_{i,n}}{\partial \xi^4} + 4(\cos\beta_{ij})^3(\sin\beta_{ij})\frac{\partial^4 T_{i,n}}{\partial \xi^3 \partial \eta} + \right.$ $+ 6(\cos\beta_{ij})^2(\sin\beta_{ij})^2\frac{\partial^4 T_{i,n}}{\partial \xi^2 \partial \eta^2} + 4(\cos\beta_{ij})(\sin\beta_{ij})^3\frac{\partial^4 T_{i,n}}{\partial \xi \partial \eta^3} +$ $\left. + (\sin\beta_{ij})^4\frac{\partial^4 T_{i,n}}{\partial \eta^4}\right]$
.	.
.	.
.	.

Table V-9

SPACE DISCRETIZATION ERROR

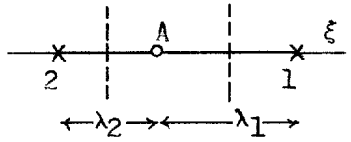
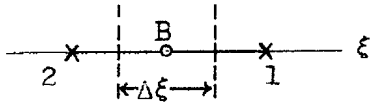
Notes Pertaining to This Table:

(1) If the last term in the expansion contains a derivative of order p, the next term not shown contains a derivative of order (p+1); all terms containing p are shown.

(2) ∘ Location of node × Location of neighboring node

(3) Letter at origin designates node

(4) —— Conductance, length of network leg – – – – Perpendicular Bisector, designates node shape and area

Node i	Expansion for $\sigma_{i,n} = \frac{1}{A_i} \sum_{j=1}^{N} y_{ij}(T_{j,n}-T_{i,n}) - \nabla'^2 T_{i,n}$
A 	One-Dimensional Problem in Cartesian Coordinates* $\left[\frac{\lambda_1-\lambda_2}{3}\right]\left[\frac{\partial^3 T_{A,n}}{\partial \xi^3}\right] + \left[\frac{\lambda_1^2-\lambda_1\lambda_2+\lambda_2^2}{12}\right]\left[\frac{\partial^4 T_{A,n}}{\partial \xi^4}\right] + \cdots$
B 	One-Dimensional Problem in Cartesian Coordinates, Regular Network $\lambda_1=\lambda_2$ * $\left[\frac{(\Delta\xi)^2}{12}\right]\left[\frac{\partial^4 T_{B,n}}{\partial \xi^4}\right] + \cdots$

* Based on Laplacian operator, one-dimensional problem

Table V-9 (Cont.)

SPACE DISCRETIZATION ERROR

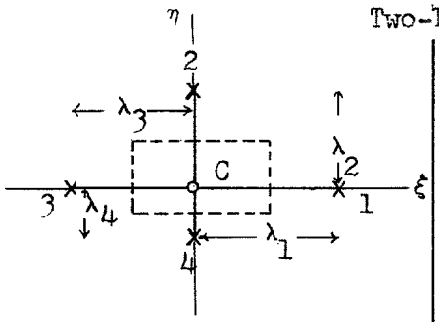
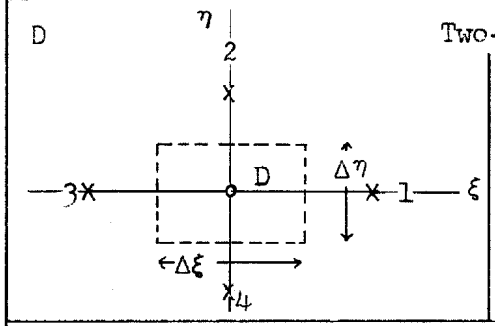
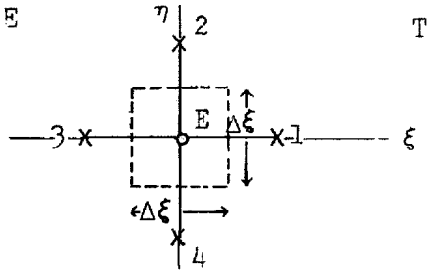
Node i	Expansion for $\sigma_{i,n}$
C 	Two-Dimensional Irregular Rectangular Node $\frac{1}{3}\left[(\lambda_1-\lambda_3)\frac{\partial^3 T_{C,n}}{\partial\xi^3}+(\lambda_2-\lambda_4)\frac{\partial^3 T_{C,n}}{\partial\eta^3}\right]$ $+\frac{1}{12}\left[(\lambda_1^2-\lambda_1\lambda_3+\lambda_3^2)\frac{\partial^4 T_{C,n}}{\partial\xi^4}+(\lambda_2^2-\lambda_2\lambda_4+\lambda_4^2)\frac{\partial^4 T_{C,n}}{\partial\eta^4}\right]+\cdots$
D 	Two-Dimensional Regular Rectangular Node $\lambda_1=\lambda_3=\Delta\xi$ $\lambda_2=\lambda_4=\Delta\eta$ $\frac{1}{12}\left[(\Delta\xi)^2\frac{\partial^4 T_{D,n}}{\partial\xi^4}+(\Delta\eta)^2\frac{\partial^4 T_{D,n}}{\partial\eta^4}\right]+\cdots$
E 	Two-Dimensional Square Node $\lambda_1=\lambda_2=\lambda_3=\lambda_4=\Delta\xi=\Delta\eta$ $\frac{(\Delta\xi)^2}{12}\left[\frac{\partial^4 T_{E,n}}{\partial\xi^4}+\frac{\partial^4 T_{E,n}}{\partial\eta^4}\right]+\cdots$

Table V-9 (Cont.)

SPACE DISCRETIZATION ERROR

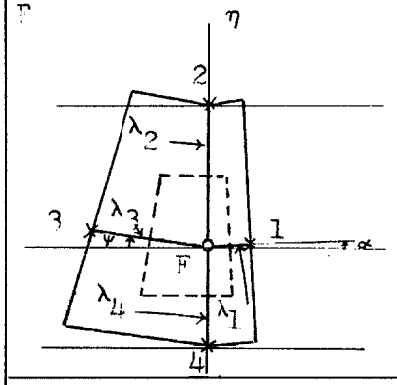
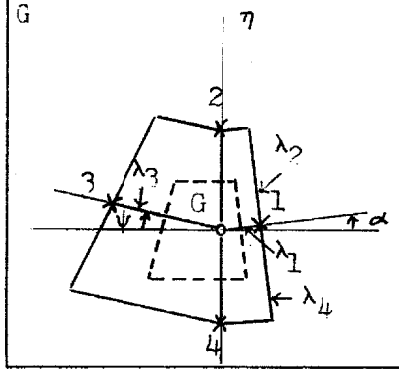
Node i	Expansion for $\sigma_{i,n}$
F 	Irregular Trapezoidal Node $\frac{(\lambda_2-\lambda_4)(\sin\alpha)\left[\sin(\alpha+\psi)\right]-2\lambda_1\left[(\cos\psi)(\sin\alpha)^3+(\cos\alpha)(\sin\psi)^3\right]}{\left[\sin(\alpha+\psi)\right]\left[2\lambda_1+(\lambda_4-\lambda_2)\sin\alpha\right]}\left[\frac{\partial^2 T_{F,n}}{\partial\xi^2}-\frac{\partial^2 T_{F,n}}{\partial\eta^2}\right]$ $+\left[\frac{2\lambda_1\cos\alpha\cos\psi\,\sin(\alpha-\psi)}{2\lambda_1+(\lambda_4-\lambda_2)(\sin\alpha)}\right]\frac{\partial^2 T_{F,n}}{\partial\xi\,\partial\eta}+\ldots$
G 	Irregular Trapezoidal Node $\lambda_2=\lambda_4$ $\lambda_1\neq\lambda_3$ $\alpha\neq\psi$ $-\left[\frac{(\sin\alpha)^3(\cos\psi)+(\sin\psi)^3(\cos\alpha)}{\sin(\alpha+\psi)}\right]\left[\frac{\partial^2 T_{G,n}}{\partial\xi^2}-\frac{\partial^2 T_{G,n}}{\partial\eta^2}\right]+$ $+\left[(\cos\alpha)(\cos\psi)\sin(\alpha-\psi)\right]\left[\frac{\partial^2 T_{G,n}}{\partial\xi\,\partial\eta}\right]+\ldots$

Table V-9 (Cont.)

SPACE DISCRETIZATION ERROR

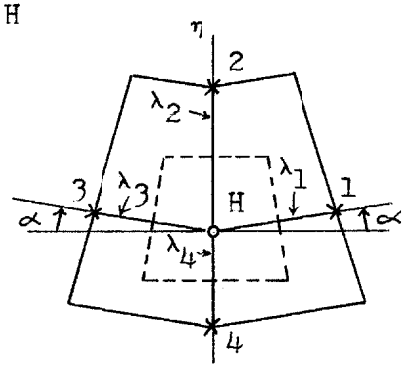
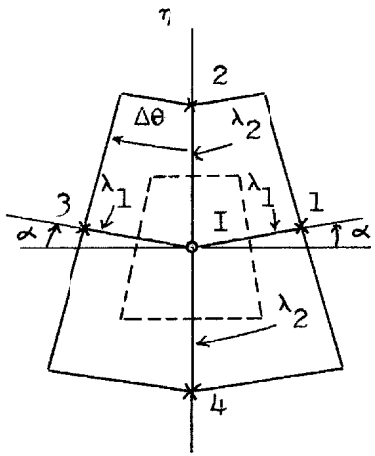
Node i	Expansion for $\sigma_{i,n}$
H 	Irregular Trapezoidal Node $\lambda_2 \neq \lambda_4$ $\lambda_1 = \lambda_3$ $\alpha = \psi$ $\left\{ \frac{(\sin\alpha)\left[(\lambda_2-\lambda_4) - 2\lambda_1\sin\alpha\right]}{(\lambda_4-\lambda_2)(\sin\alpha) + 2\lambda_1} \right\} \left\{ \frac{\partial^2 T_{H,n}}{\partial\xi^2} - \frac{\partial^2 T_{H,n}}{\partial\eta^2} \right\} + \ldots$
I 	Regular Trapezoidal Node $\lambda_2 = \lambda_4 = \Delta\rho$ $\lambda_1 = \lambda_3$ $\alpha = \psi = \Delta\theta/2$ $-(\sin\alpha)^2\left[\frac{\partial^2 T_{I,n}}{\partial\xi^2} - \frac{\partial^2 T_{I,n}}{\partial\eta^2}\right] + \lambda_1\sin\alpha(\cos\alpha)^2 \frac{\partial^3 T_{I,n}}{\partial\xi^2\partial\eta}$ $+ \left[\frac{\sin\alpha}{3}\right]\left[\lambda_1(\sin\alpha)^2 - \frac{\lambda_2^2}{\lambda_1}\right] \frac{\partial^3 T_{I,n}}{\partial\eta^3} + \ldots$ or $\frac{1}{2}(\cos\Delta\theta - 1)\left[\frac{\partial^2 T_{I,n}}{\partial\xi^2} - \frac{\partial^2 T_{I,n}}{\partial\eta^2}\right] + \lambda_1(\sin\frac{\Delta\theta}{2})(\cos\frac{\Delta\theta}{2})^2 \frac{\partial^3 T_{I,n}}{\partial\xi^2\partial\eta}$ $+ \frac{1}{3}(\sin\frac{\Delta\theta}{2})\left[\lambda_1(\sin\frac{\Delta\theta}{2})^2 - \frac{(\Delta\rho)^2}{\lambda_1}\right] \frac{\partial^3 T_{I,n}}{\partial\eta^3} + \ldots$

Table V-9 (Cont.)

SPACE DISCRETIZATION ERROR

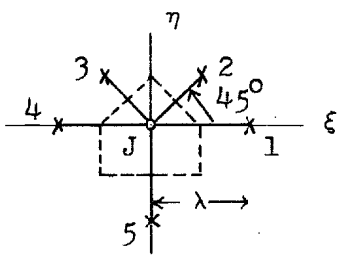
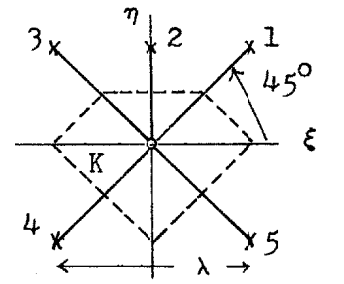
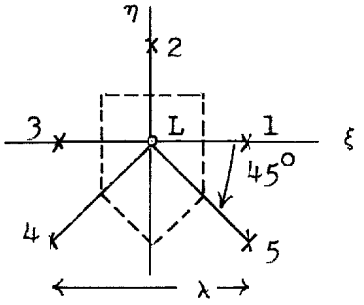
Node i	Expansion for $\sigma_{i,n}$
J 	Node for Changing Network Size: MacNeal (8), Fig. 6A $-\frac{\lambda}{6}\left[\frac{\partial^3 T_{J,n}}{\partial \eta^3} - \frac{\partial^3 T_{J,n}}{\partial \xi^2 \partial \eta}\right] + \ldots$
K 	Node for Changing Network Size: MacNeal (8), Fig. 6A $-\frac{1}{7}\left[\frac{\partial^2 T_{K,n}}{\partial \xi^2} - \frac{\partial^2 T_{K,n}}{\partial \eta^2}\right] - \frac{\lambda}{7}\left[\frac{\partial^3 T_{K,n}}{\partial \xi^2 \partial \eta}\right] + \ldots$
L 	Node for Changing Network Size: MacNeal (8), Fig. 6A $\frac{1}{5}\left[\frac{\partial^2 T_{L,n}}{\partial \xi^2} - \frac{\partial^2 T_{L,n}}{\partial \eta^2}\right] - \frac{\lambda}{5}\left[\frac{\partial^3 T_{L,n}}{\partial \xi^2 \partial \eta}\right] + \ldots$

Table V-9 (Cont.)

SPACE DISCRETIZATION ERROR

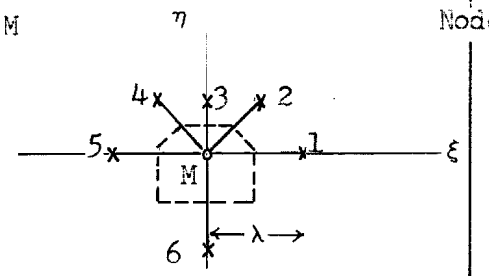
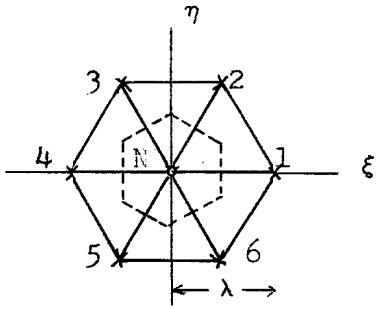
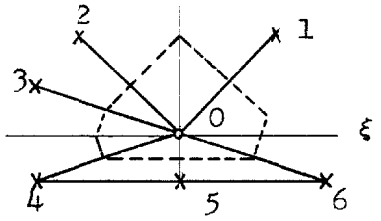
Node i	Expansion for $\sigma_{i,n}$
M 	Node for Changing Network Size: MacNeal (8), Fig. 6B $-\frac{1}{11}\left[\frac{\partial^2 T_{M,n}}{\partial \xi^2} - \frac{\partial^2 T_{M,n}}{\partial \eta^2}\right] + \frac{\lambda}{11}\left[\frac{\partial^3 T_{M,n}}{\partial \xi^2 \partial \eta} - 2\frac{\partial^3 T_{M,n}}{\partial \eta^3}\right] + \cdots$
N 	Network of Equilateral Triangles: Hexagonal Node $\frac{\lambda^2}{16}\left[\frac{\partial^4 T_{N,n}}{\partial \xi^4} + 2\frac{\partial^4 T_{N,n}}{\partial \xi^2 \partial \eta^2} + \frac{\partial^4 T_{N,n}}{\partial \eta^4}\right] + \cdots$ or $\frac{\lambda^2}{16}\left[\nabla'^4 T_{N,n}\right] + \cdots$
O 	Node 1 in Figure III-1: Adiabatic or Symmetric Boundary about Hypotenuse $0.070\left[\frac{\partial^2 T_{O,n}}{\partial \xi^2} - \frac{\partial^2 T_{O,n}}{\partial \eta^2}\right] - 0.019\left[\frac{\partial^2 T_{O,n}}{\partial \xi\, \partial \eta}\right] + \cdots$

Table V-9 (Cont.)

SPACE DISCRETIZATION ERROR

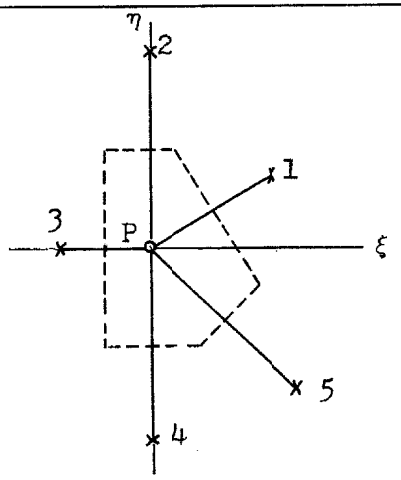
Node i	Expansion for $\sigma_{i,n}$
	Node 6 in Figure III-1 $$- 0.039\left[\frac{\partial^2 T_{P,n}}{\partial \xi^2} - \frac{\partial^2 T_{P,n}}{\partial \eta^2}\right] + 0.064\left[\frac{\partial^2 T_{P,n}}{\partial \xi \, \partial \eta}\right] + \ldots$$

Table V- 10

ERRORS - TRANSIENT PROBLEM - CHAPTER III

γ	$\Delta\tau$	τ	T_1-t_1	T_2-t_2	T_3-t_3	T_4-t_4	T_5-t_5	T_6-t_6	T_7-t_7	T_8-t_8	T_9-t_9
Analog	0	0.003	0.0046	-0.0276	-0.0203	-0.1010	0.0179	0.0549	0.0111	0.0025	0.0006
0	0.0001		0.0094	-0.0235	-0.0155	-0.0975	0.0171	0.0549	0.0107	0.0025	0.0006
	0.0005		0.0249	-0.0060	0.0065	-0.0824	0.0132	0.0551	0.0093	0.0021	0.0004
	0.001		0.0668	0.0377	0.0162	-0.0612	0.0071	0.0551	0.0073	0.0016	0.0002
0.5	0.0005		0.0066	-0.0269	-0.0199	-0.1006	0.0177	0.0549	0.0111	0.0026	0.0007
	0.0015		0.0192	-0.0231	-0.0140	-0.0971	0.0154	0.0548	0.0110	0.0026	0.0007
1.0	0.0005		-0.0222	-0.0472	-0.0406	-0.1174	0.0125	0.0549	0.0128	0.0031	0.0010
Analog	0	0.006	0.0018	-0.0142	-0.0099	-0.0377	-0.0076	0.0600	0.0284	0.0108	0.0045
0	0.0001		0.0030	-0.0113	-0.0062	-0.0347	-0.0071	0.0602	0.0279	0.0106	0.0044
	0.0005		0.0073	0.0002	0.0088	-0.0224	-0.0051	0.0609	0.0261	0.0098	0.0038
	0.001		0.0104	0.0131	0.0299	-0.0061	-0.0020	0.0619	0.0236	0.0090	0.0029
0.5	0.0005		0.0020	-0.0137	-0.0094	-0.0374	-0.0078	0.0600	0.0283	0.0108	0.0046
	0.0015		0.0035	-0.0106	-0.0052	-0.0346	-0.0088	0.0600	0.0280	0.0108	0.0046
1.0	0.0005		-0.0047	-0.0280	-0.0279	-0.0521	-0.0099	0.0592	0.0304	0.0118	0.0065

Table V-10(Cont.)

ERRORS - TRANSIENT PROBLEM - CHAPTER III

γ	$\Delta\tau$	τ	T_1-t_1	T_2-t_2	T_3-t_3	T_4-t_4	T_5-t_5	T_6-t_6	T_7-t_7	T_8-t_8	T_9-t_9
Analog	0	0.01	0.0004	-0.0057	-0.0035	0.0019	-0.0146	0.0456	0.0322	0.0256	0.0162
0	0.0001		0.0009	-0.0043	-0.0016	0.0036	-0.0131	0.0459	0.0319	0.0253	0.0159
	0.0005		0.0028	0.0313	0.0056	0.0106	-0.0070	0.0471	0.0310	0.0241	0.0146
	0.001		0.0048	0.0078	0.0141	0.0191	0.0014	0.0486	0.0298	0.0225	0.0130
0.5	0.0005		0.0005	-0.0056	-0.0032	0.0021	-0.0146	0.0456	0.0321	0.0256	0.0162
	0.0015*		0.0034	0.0335	0.0089	0.0130	0.0033	0.0554	0.0407	0.0297	0.0184
	0.01		0.5778	0.2404	0.0892	0.1563	-0.0706	0.0439	0.0163	0.0238	0.0169
1.0	0.0005		-0.0021	-0.0131	-0.0129	-0.0069	-0.0215	0.0443	0.0332	0.0270	0.0177
	0.01		-0.0973	-0.1493	-0.1733	-0.1481	-0.0961	0.0241	0.0412	0.0448	0.0373
Analog	0	0.05	0.0002	0.0003	0.0018	0.0079	0.0058	0.0018	0.0009	0.0086	0.0116
0	0.0001		0.0003	0.0010	0.0019	0.0080	0.0061	0.0021	0.0013	0.0089	0.0119
	0.0005		-0.0001	0.0013	0.0022	0.0084	0.0071	0.0035	0.0029	0.0103	0.0132
	0.001		0.0004	0.0016	0.0027	0.0089	0.0083	0.0053	0.0050	0.0120	0.0148
0.5	0.0005		0.0002	0.0009	0.0018	0.0079	0.0058	0.0017	0.0009	0.0086	0.0116
	0.0015*		0.0002	0.0004	0.0012	0.0072	0.0042	-0.0017	-0.0033	0.0040	0.0065
	0.01		0.1509	0.0292	-0.0497	0.0087	0.0080	0.0033	0.0031	0.0081	0.0106
1.0	0.0005		0.0001	0.0005	0.0013	0.0074	0.0045	0.0001	-0.0012	0.0069	0.0101
	0.01		-0.0025	-0.0090	-0.0108	-0.0042	-0.0248	-0.0301	-0.0384	-0.0190	-0.0135

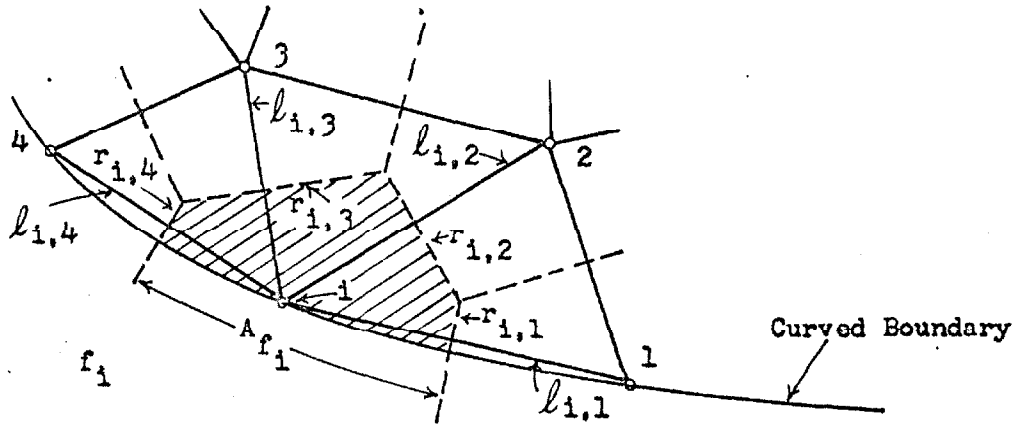
*Data for $\gamma=0.5$, $\Delta\tau=0.0015$ taken at increment closest to τ.

Table V-10(Cont.)

ERRORS - TRANSIENT PROBLEM - CHAPTER III

γ	$\Delta\tau$	τ	T_1-t_1	T_2-t_2	T_3-t_3	T_4-t_4	T_5-t_5	T_6-t_6	T_7-t_7	T_8-t_8	T_9-t_9
Analog	0	0.10	-0.0001	-0.0004	-0.0003	0.0016	-0.0005	-0.0059	-0.0072	-0.0107	-0.0108
0	0.0001		-0.0001	-0.0004	-0.0002	0.0016	-0.0004	-0.0058	-0.0069	-0.0104	-0.0104
	0.0005		-0.0001	-0.0003	-0.0001	0.0019	0.00004	-0.0045	-0.0058	-0.0088	-0.0088
	0.001		-0.00003	-0.0001	-0.0001	0.0021	0.0006	-0.0036	-0.0045	-0.0069	-0.0068
0.5	0.0005		-0.0001	-0.0004	-0.0003	0.0016	-0.0005	-0.0060	-0.0072	-0.0107	-0.0108
	0.0015*		0.0030	-0.0002	-0.0001	0.0019	0.0001	-0.0043	-0.0060	-0.0085	-0.0089
	0.01		0.0004	-0.0085	0.0144	0.0019	-0.0004	-0.0052	-0.0063	-0.0096	-0.0096
1.0	0.0005		-0.0002	-0.0005	-0.0004	0.0014	-0.0011	-0.0072	-0.0085	-0.0126	-0.0128
	0.01		-0.0009	-0.0034	-0.0040	-0.0031	-0.0114	-0.0293	-0.0333	-0.0465	-0.0483
Analog	0	0.20	-0.0001	-0.0004	-0.0004	-0.0001	-0.0011	-0.0037	-0.0035	-0.0052	-0.0055
0	0.0001		-0.0001	-0.0003	-0.0004	-0.0002	-0.0011	-0.0031	-0.0035	-0.0052	-0.0054
	0.0005		-0.0001	-0.0003	-0.0003	-0.0002	-0.0009	-0.0028	-0.0032	-0.0047	-0.0049
	0.001		-0.00002	-0.0002	-0.0005	-0.0002	-0.0015	-0.0028	-0.0033	-0.0016	-0.0048
0.5	0.0005		-0.00008	-0.00031	-0.00038	-0.00023	-0.00111	-0.0025	-0.0010	-0.0054	-0.0059
	0.0015*		-0.0001	-0.0003	-0.0003	-0.0001	-0.0010	-0.0027	-0.0031	-0.0046	-0.0048
	0.01		-0.0017	-0.0009	0.0006	-0.0001	-0.0010	-0.0029	-0.0033	-0.0049	-0.0051
1.0	0.0005		-0.0001	-0.0004	-0.0004	-0.0003	-0.0013	-0.0035	-0.0040	-0.0059	-0.0061
	0.01		-0.0003	-0.0013	-0.0015	-0.0017	-0.0045	-0.0108	-0.0122	-0.0178	-0.0188

*Data for $\gamma=0.5$, $\Delta\tau=0.0015$ taken at increment closest to τ.



——— Connectors

– – – – Perpendicular Bisectors

Energy Balance for i^{th} Surface Node:

$$h(t_{f_i}-t_i)_\theta \; A_{f_i} + k\left[\sum_{j=1}^{4} \frac{r_{ij}}{\ell_{ij}}(t_j-t_i)_\theta\right] = \text{Net Energy In at Time } \theta$$

$$C_p \sigma (\text{Area})_i \left[\frac{dt_i}{d\theta}\right] = \text{Accumulation at Time } \theta$$

$(\text{Area})_i$ = Cross-Hatched Area

Figure V-1. MacNeal's approximation for surface node on curved boundary with finite non-zero heat-transfer coefficient.

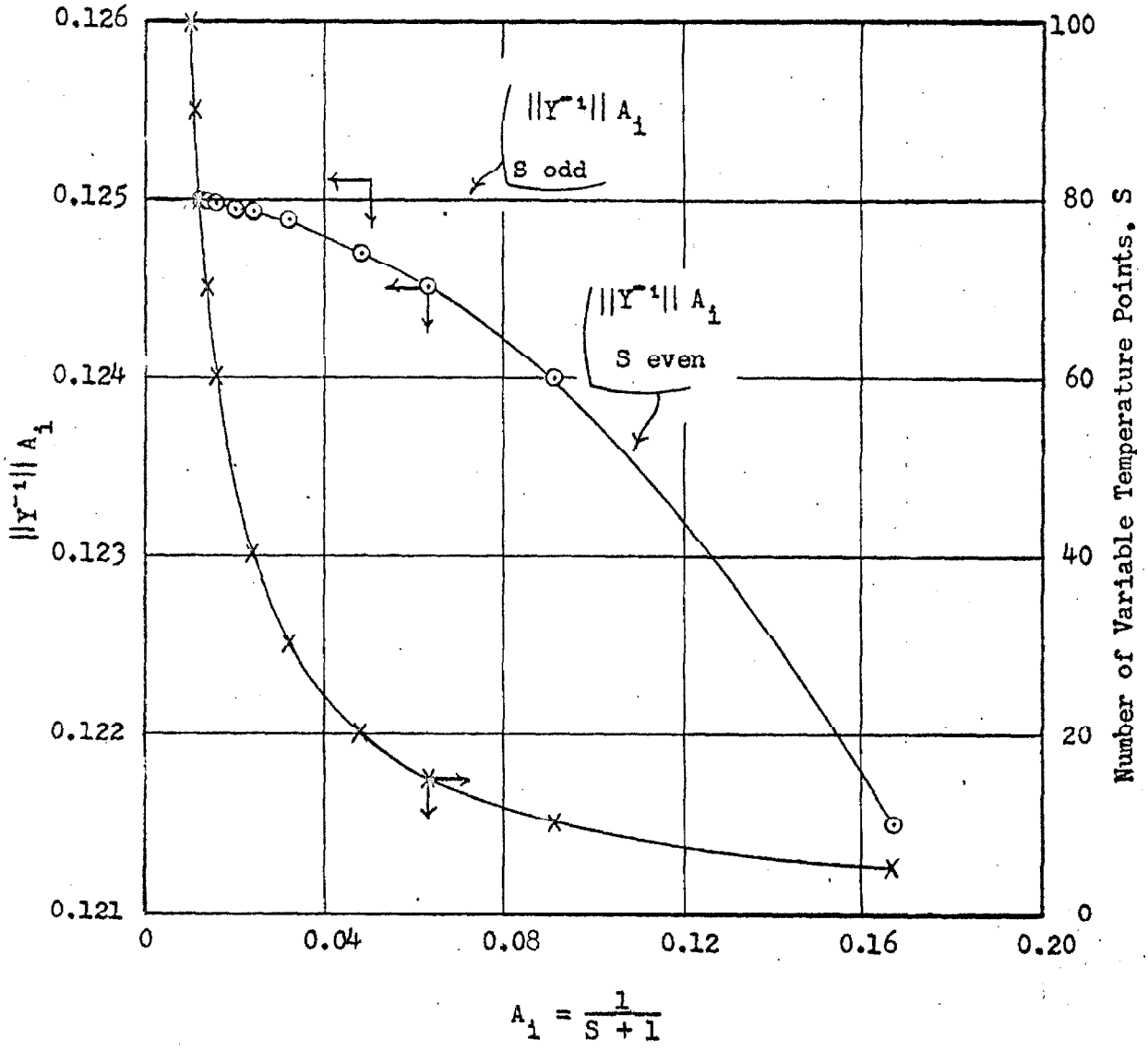


Figure V-2. Relationship between nodal area and the product of the norm of the Y^{-1} matrix and the nodal area, one-dimensional problem, regular network: infinite h.

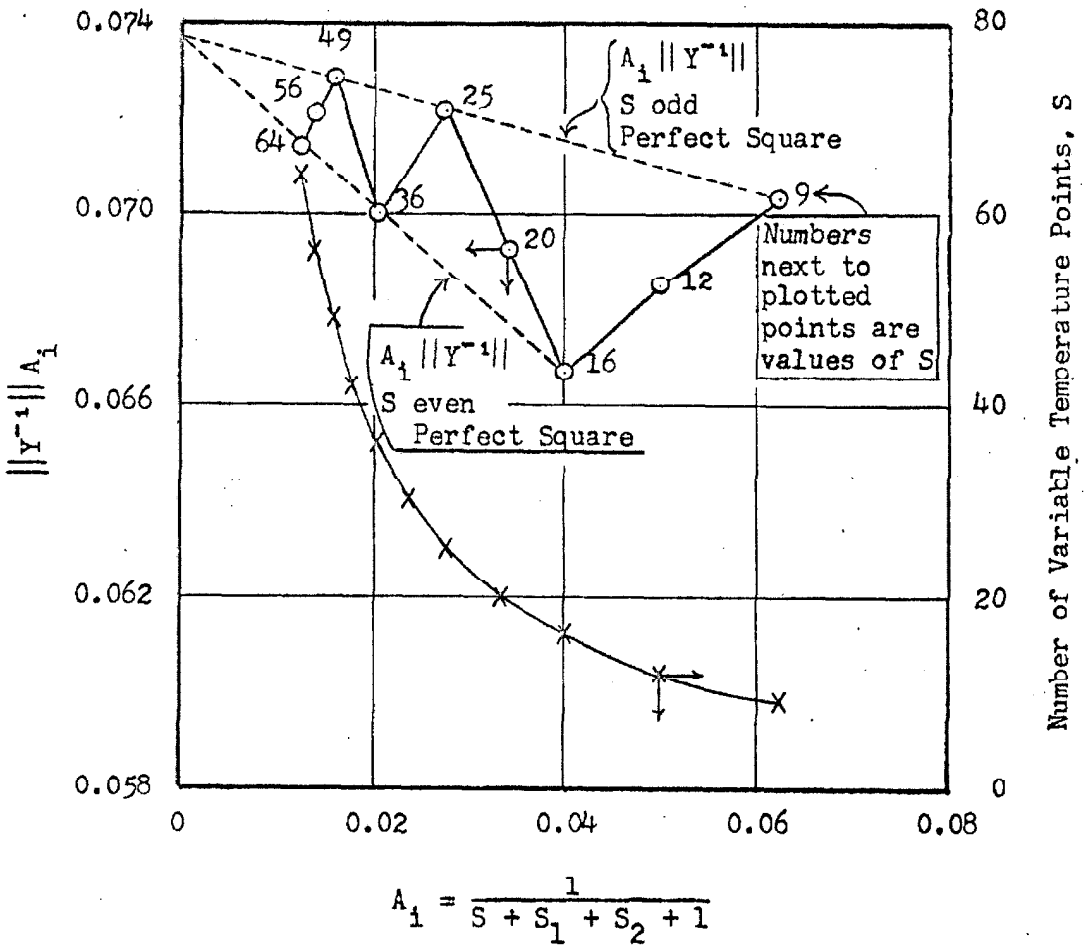
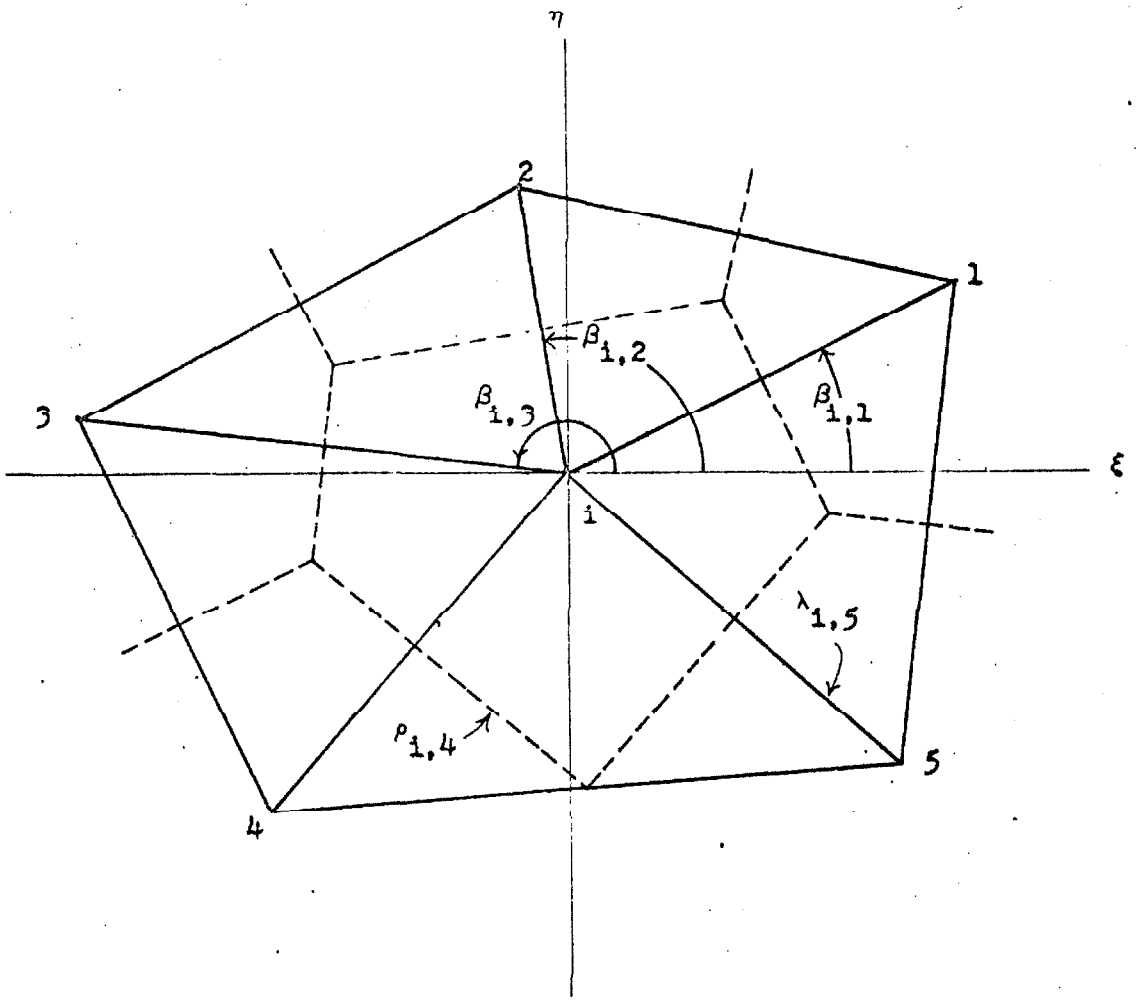


Figure V-3. Relationship between nodal area and the product of the norm of the Y^{-1} matrix and the nodal area, two-dimensional problem, regular rectangular network; infinite h.



——— Lines connecting nodes, length designated by λ_{ij} (see $\lambda_{i,5}$)

– – – – Perpendicular bisectors of lines connecting nodes, length designated by ρ_{ij} (see $\rho_{i,4}$)

Figure V-4 . General asymmetric node, two-dimensional problem showing geometry in dimensionless coordinates.

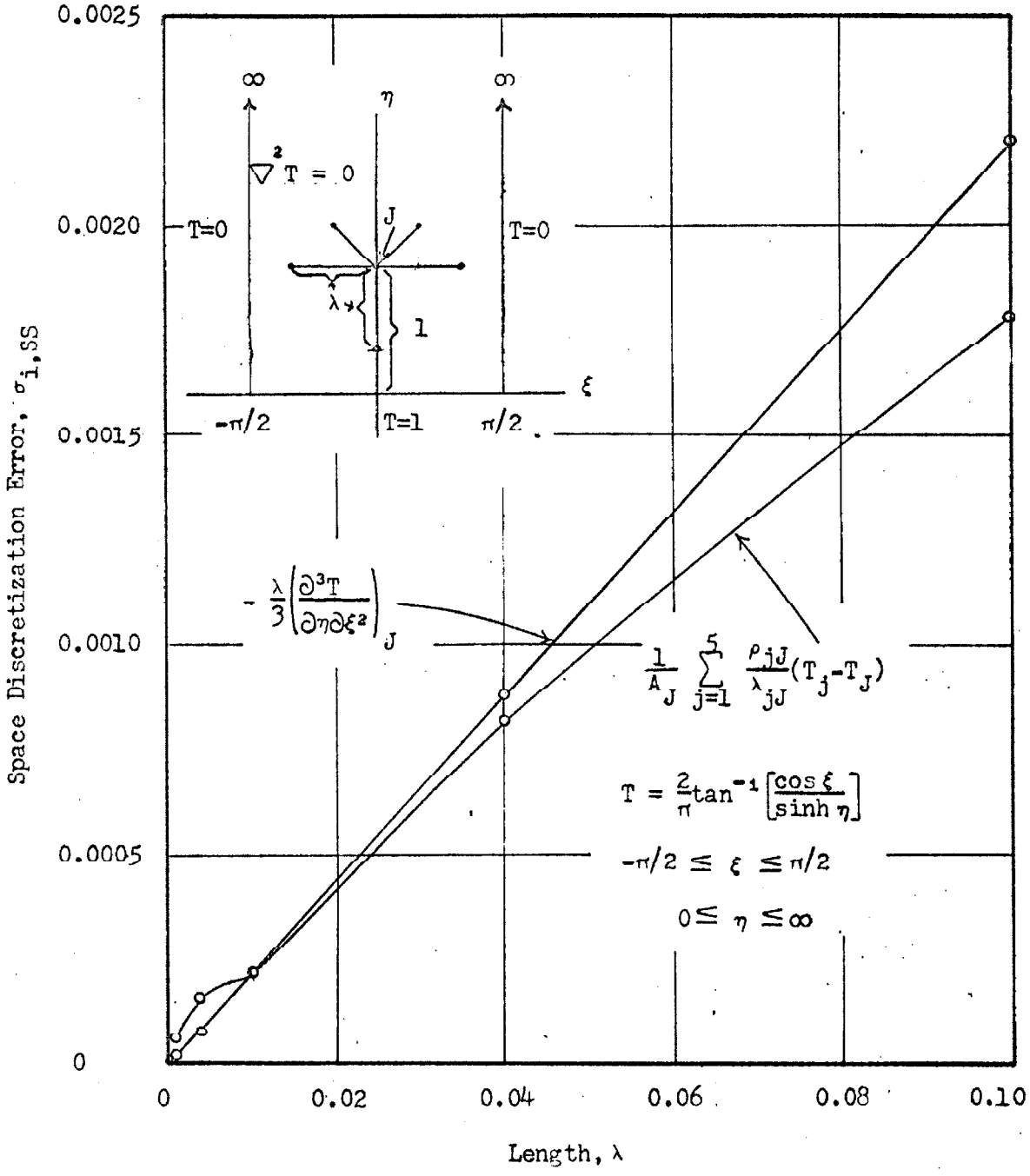


Figure V-5 . Space discretization error, node J (see Table V- 9), sequence of geometrically similar nodes.

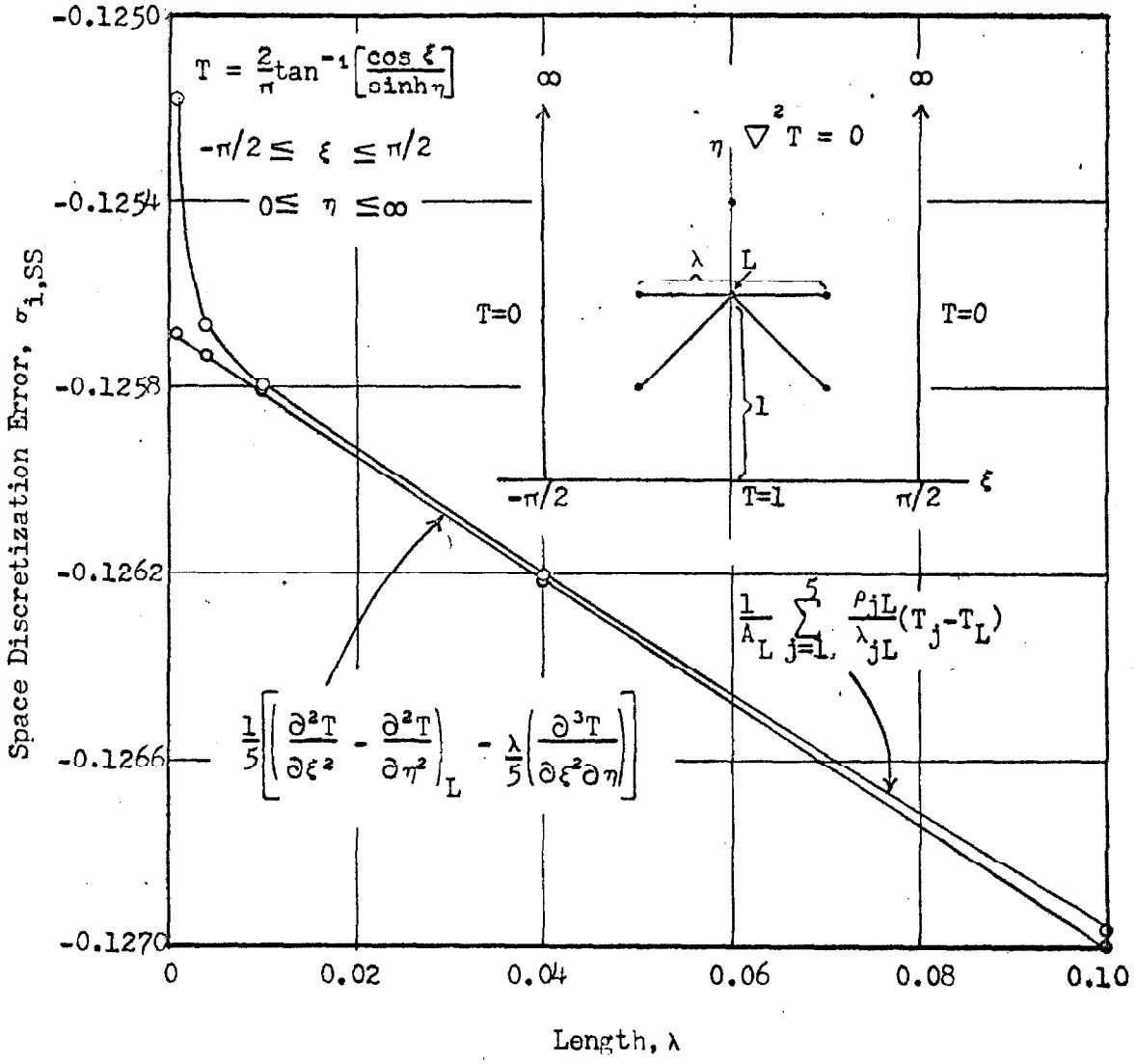
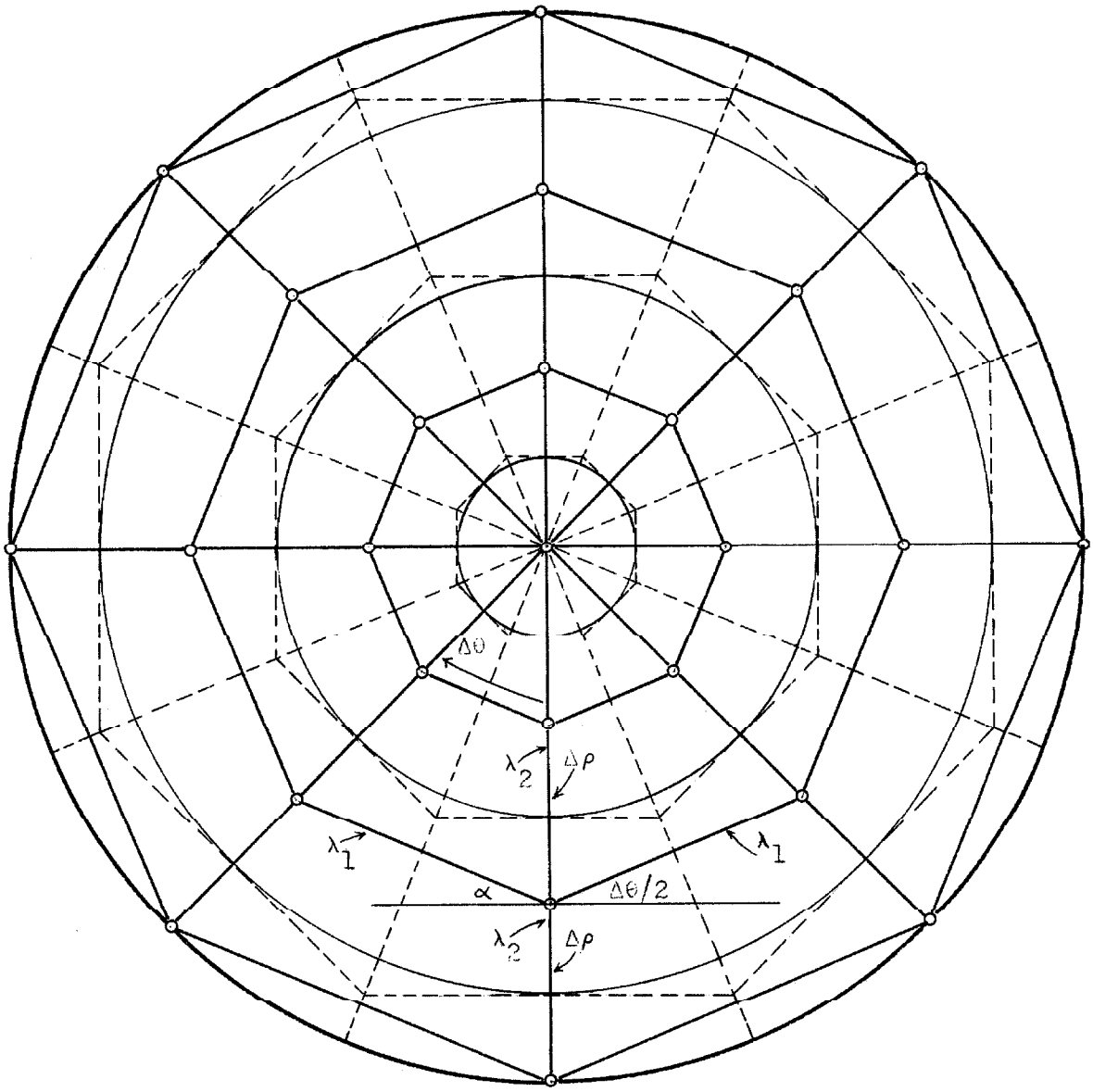


Figure V-6 . Space discretization error, node L (see Table V- 9), sequence of geometrically similar nodes.



——— Connectors

– – – – Perpendicular Bisectors

——— Boundary of Nodes for Usual Cylindrical Coordinates

Figure V-7. Comparison between annular network and network of trapezoidal nodes for cylindrical coordinates based on MacNeal's rules; annular network is usual approximation.

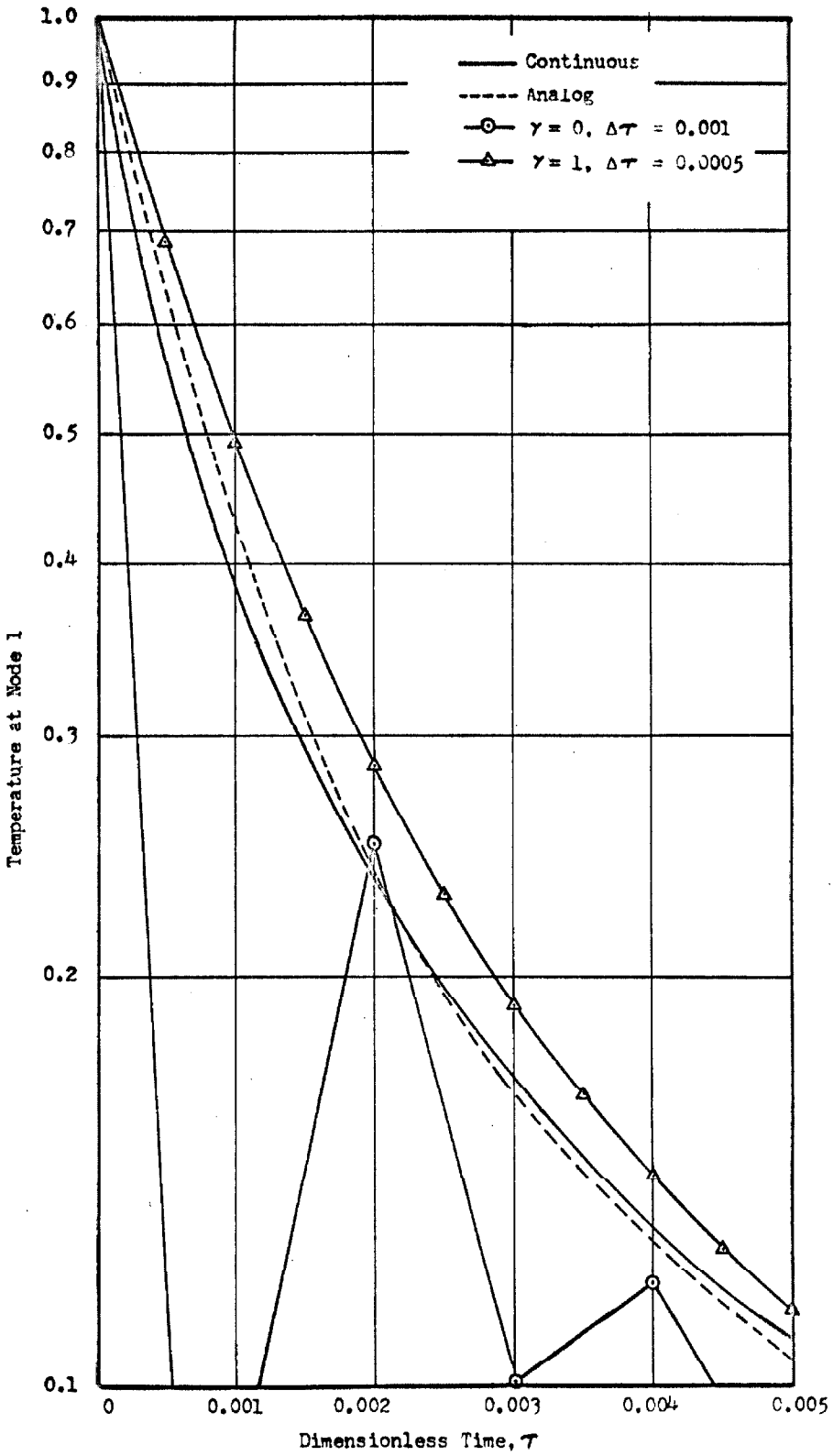


Figure V-2. Temperature at node 1 (Fig. III-1) versus dimensionless time; short time.

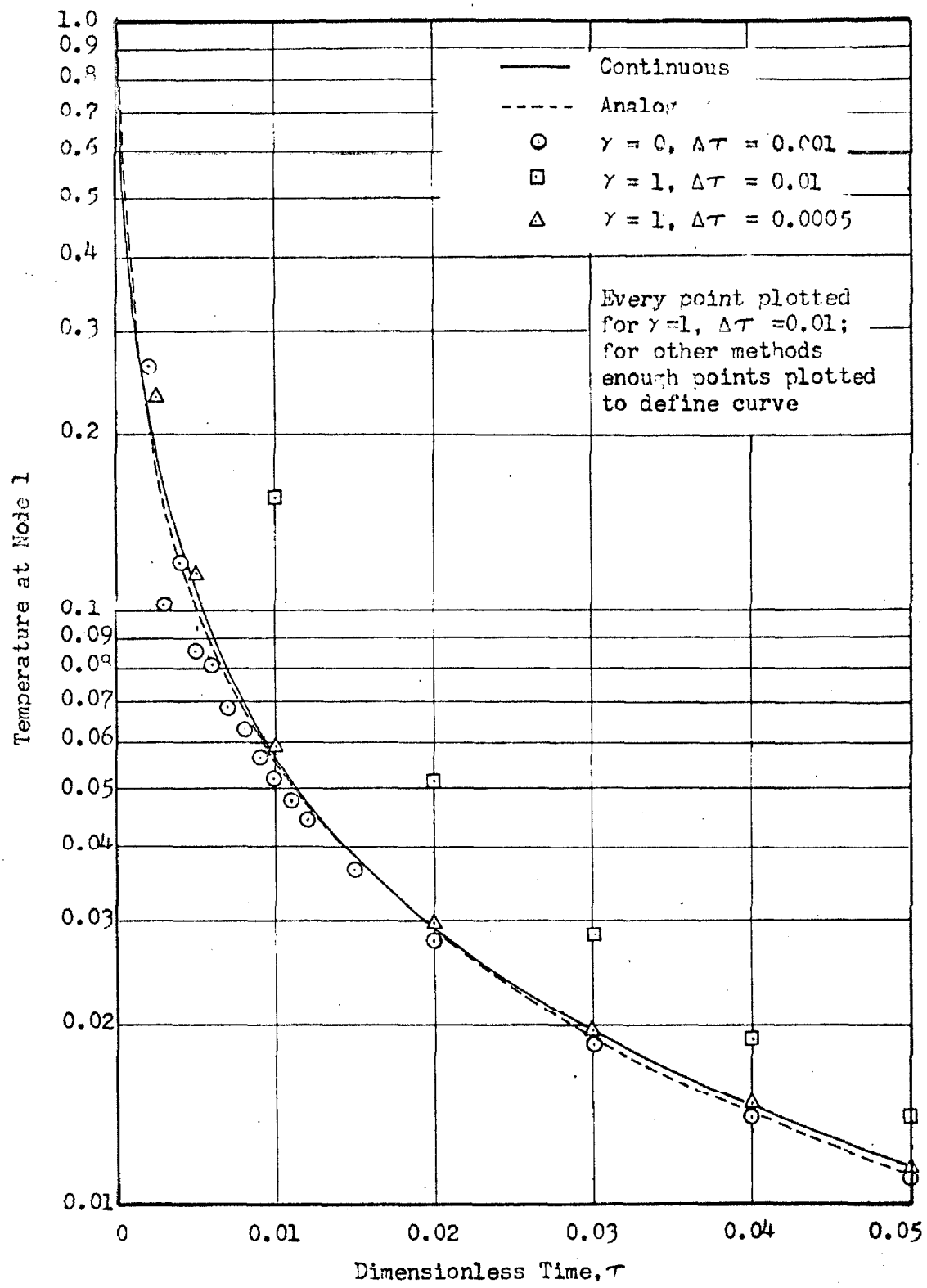


Figure V-9 . Temperature at node 1 (Fig. III-1) versus dimensionless time; intermediate time.

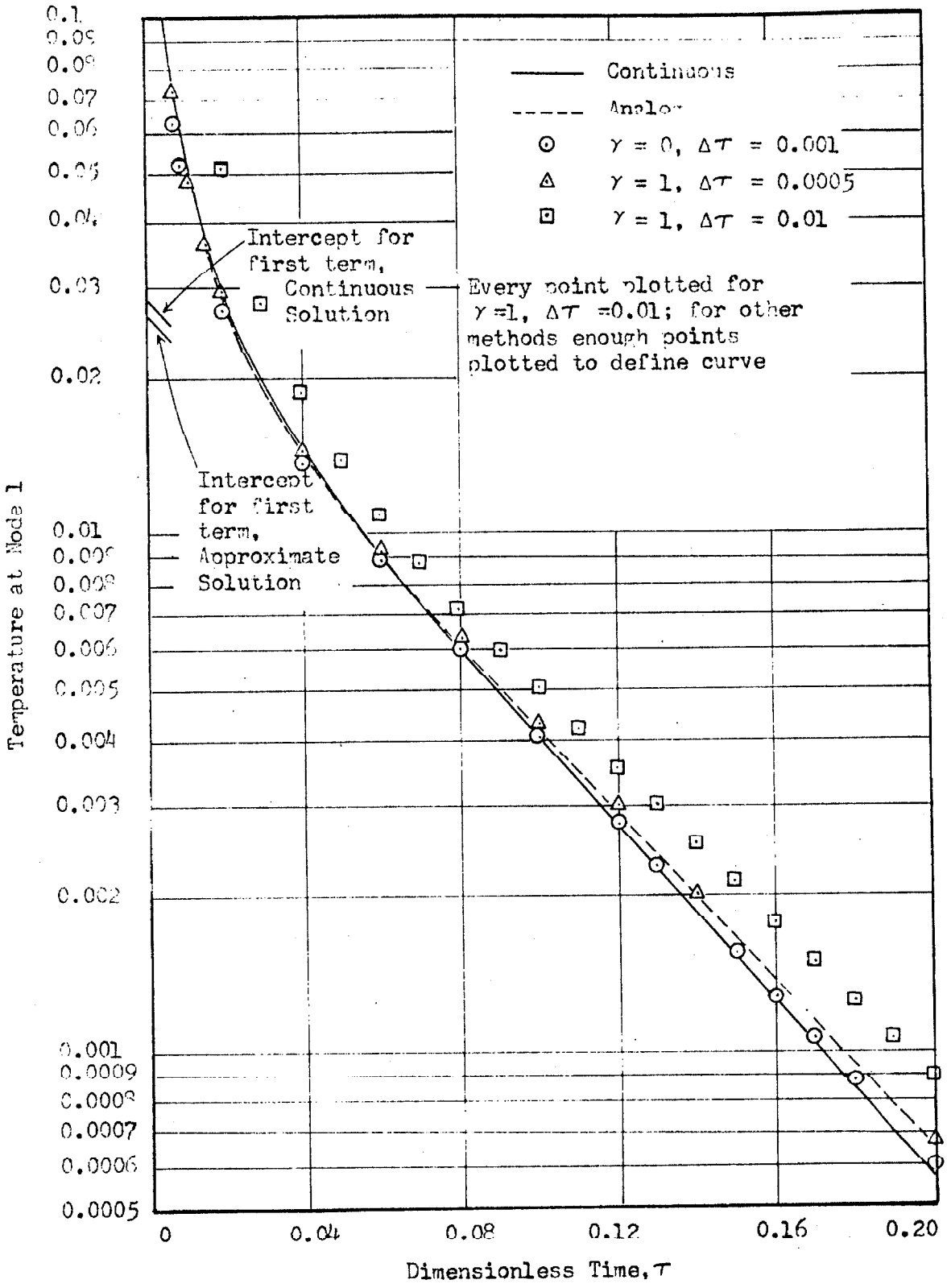


Figure V-1.9. Temperature at node 1 (Fig. III-1) versus dimensionless time; long time.

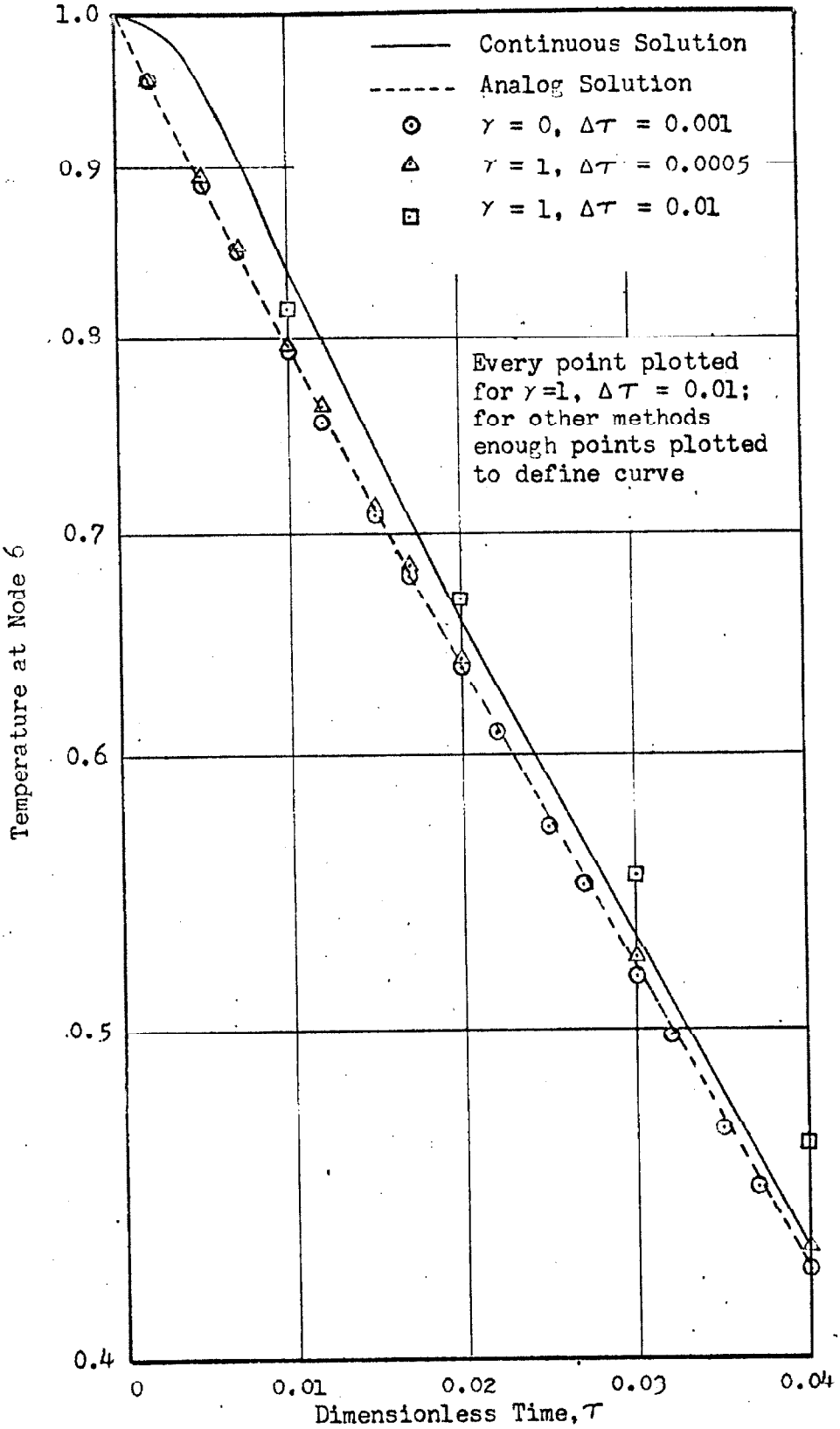


Figure V-11. Temperature at node 6 (Fig. III-1) versus dimensionless time; short time.

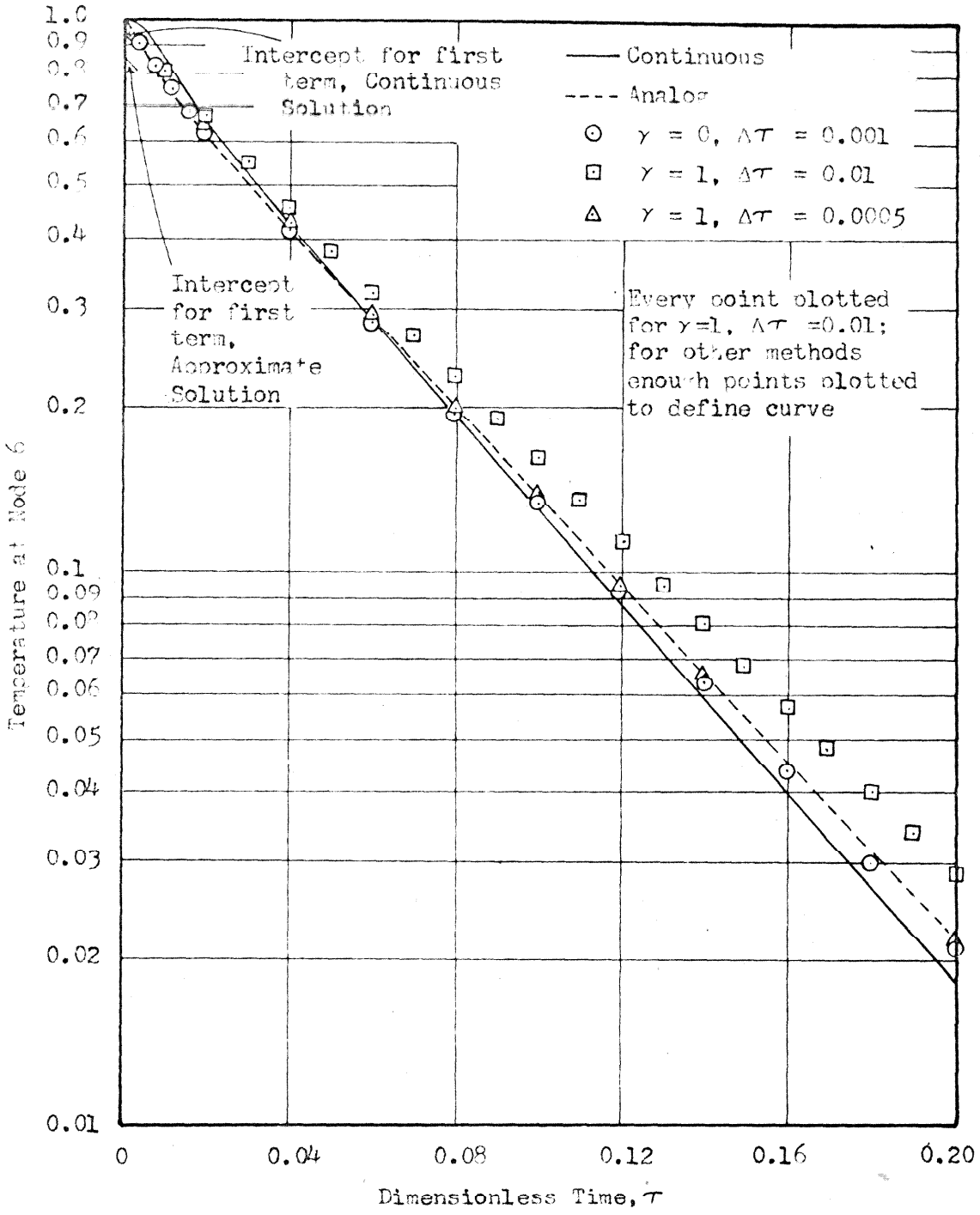
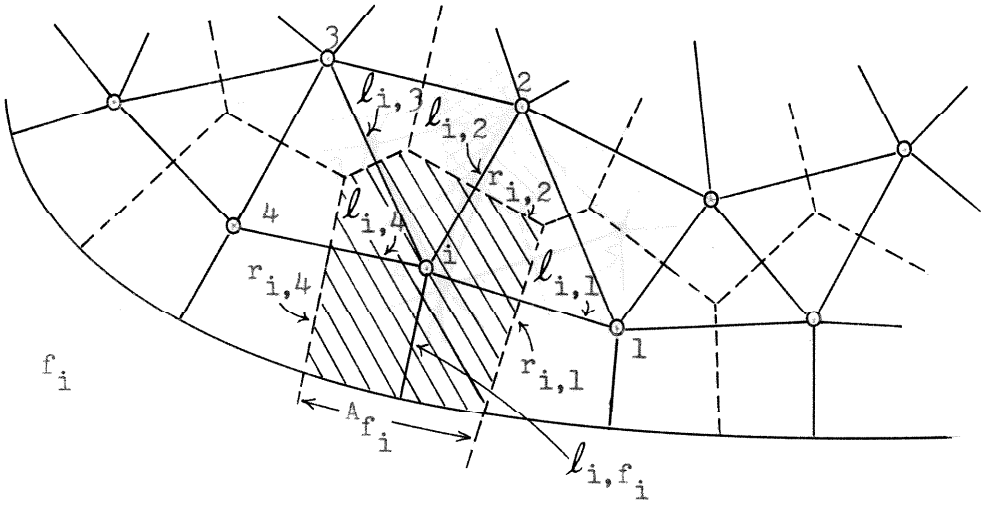


Figure V-12. Temperature at node 6 (Fig. III-1) versus dimensionless time; long time.



——— Connectors, Network Legs
----- Perpendicular Bisectors

Energy Balance for i^{th} Node:

$$\frac{A_{f_i}(t_{f_i}-t_i)_\theta}{\frac{1}{h}+\frac{\ell_{i,f_i}}{k}} + k\sum_{j=1}^{4}\frac{r_{ij}}{\ell_{ij}}(t_j-t_i)_\theta = \text{Net Energy in at Time}$$

$$C_p\sigma(\text{Area})_i\frac{dt_i}{d\theta} = \text{Accumulation at Time}$$

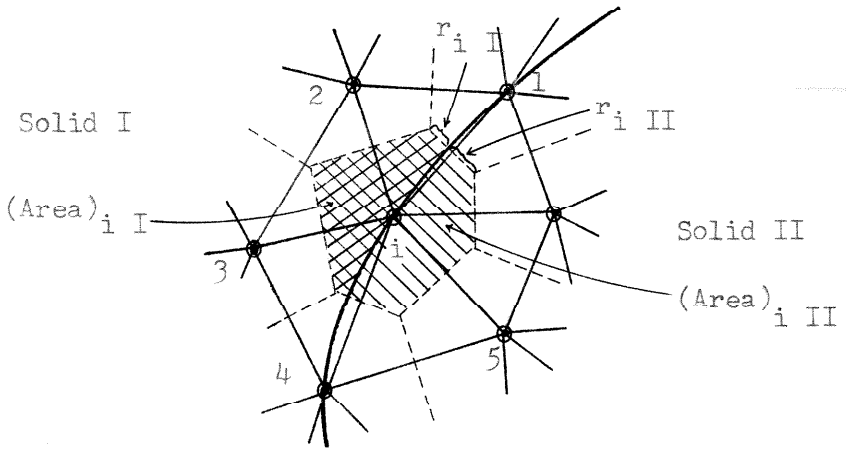
$(\text{Area})_i$ = Cross-Hatched Area

$t^*_{i,}$ = Temperature at Boundary Surface of ith Adjacent Node

$$h(t_{f_i}-t^*_i) = \frac{k(t_i-t_i)_\theta}{\ell_{i,f_i}}$$

$$\text{or} \quad t^*_{i_\theta} = \frac{\left[\frac{h\ell_{i,f_i}}{k}\left(t_{f_i}\right)+t_i\right]_\theta}{\frac{h\ell_{i,f_i}}{k}+1}$$

Figure V-13. Modification to MacNeal's approximation for adjacent node on curved boundary with finite non-zero heat-transfer coefficient; analogous to mesh $\Delta\xi/2$.



Dimensionless Capacity:

$$A_i = \left[(Area)_{i\ I}\, C_{p\ I}\, \sigma_I + (Area)_{i\ II}\, C_{p\ II}\, \sigma_{II} \right] \frac{1}{C_{p\ 0}\, \sigma_0 L^2}$$

Dimensionless Conductance:

$$y_{i\ I} = \frac{1}{k_0} \left(k_I r_{i\ I} + k_{II} r_{i\ II} \right)$$

MacNeal Interfacial Node

Figure V-14. MacNeal node for interfacial boundaries between uniform solids of different thermal properties.

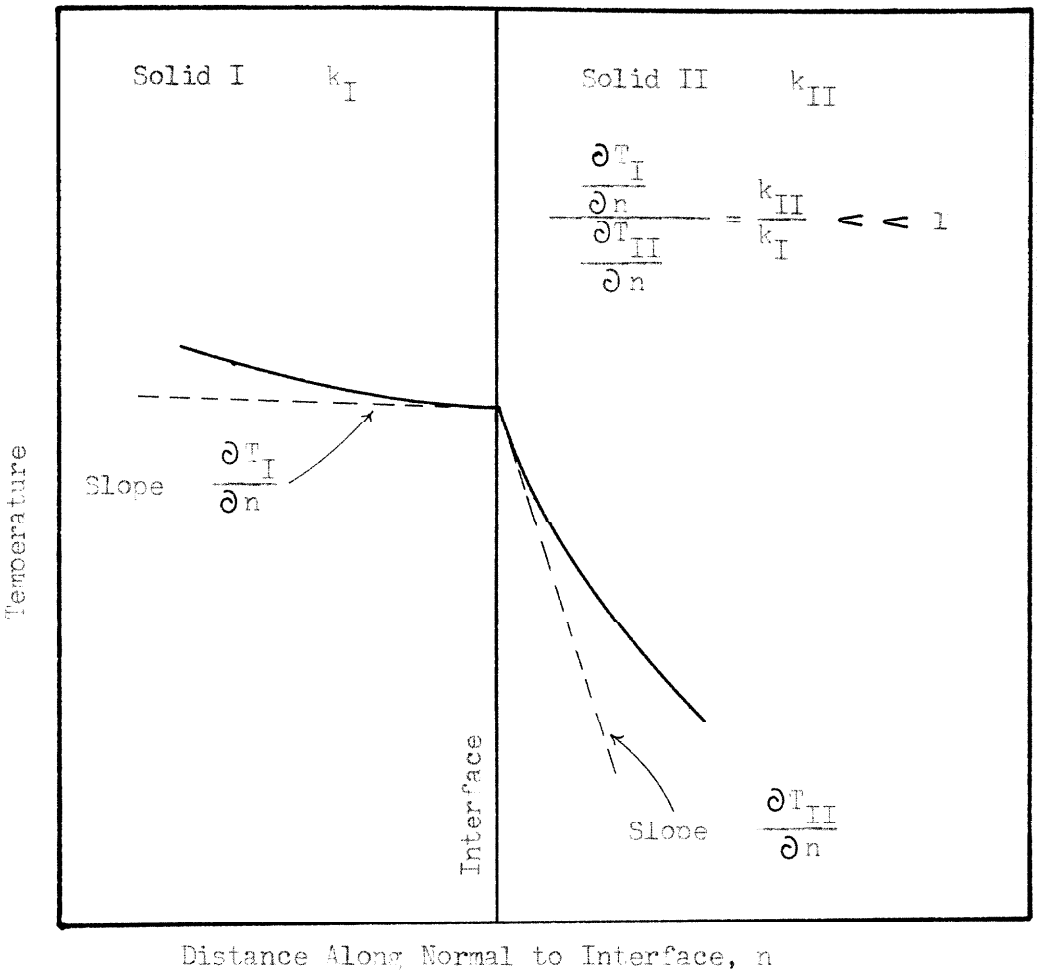
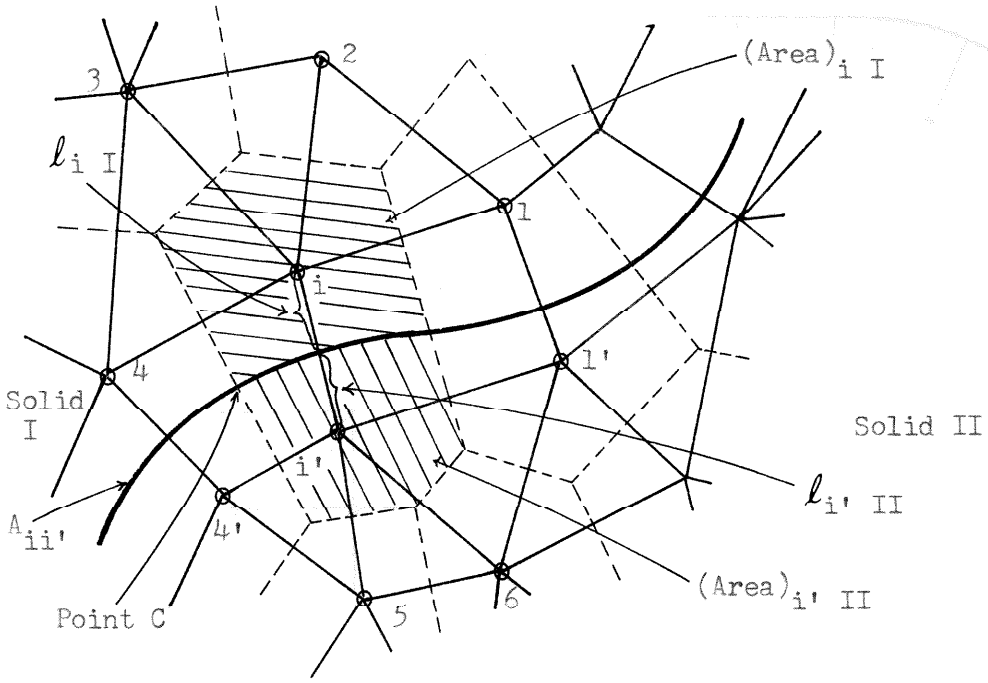


Figure V-15. Continuous temperature distribution across interfacial boundary of uniform solids with very different thermal properties.



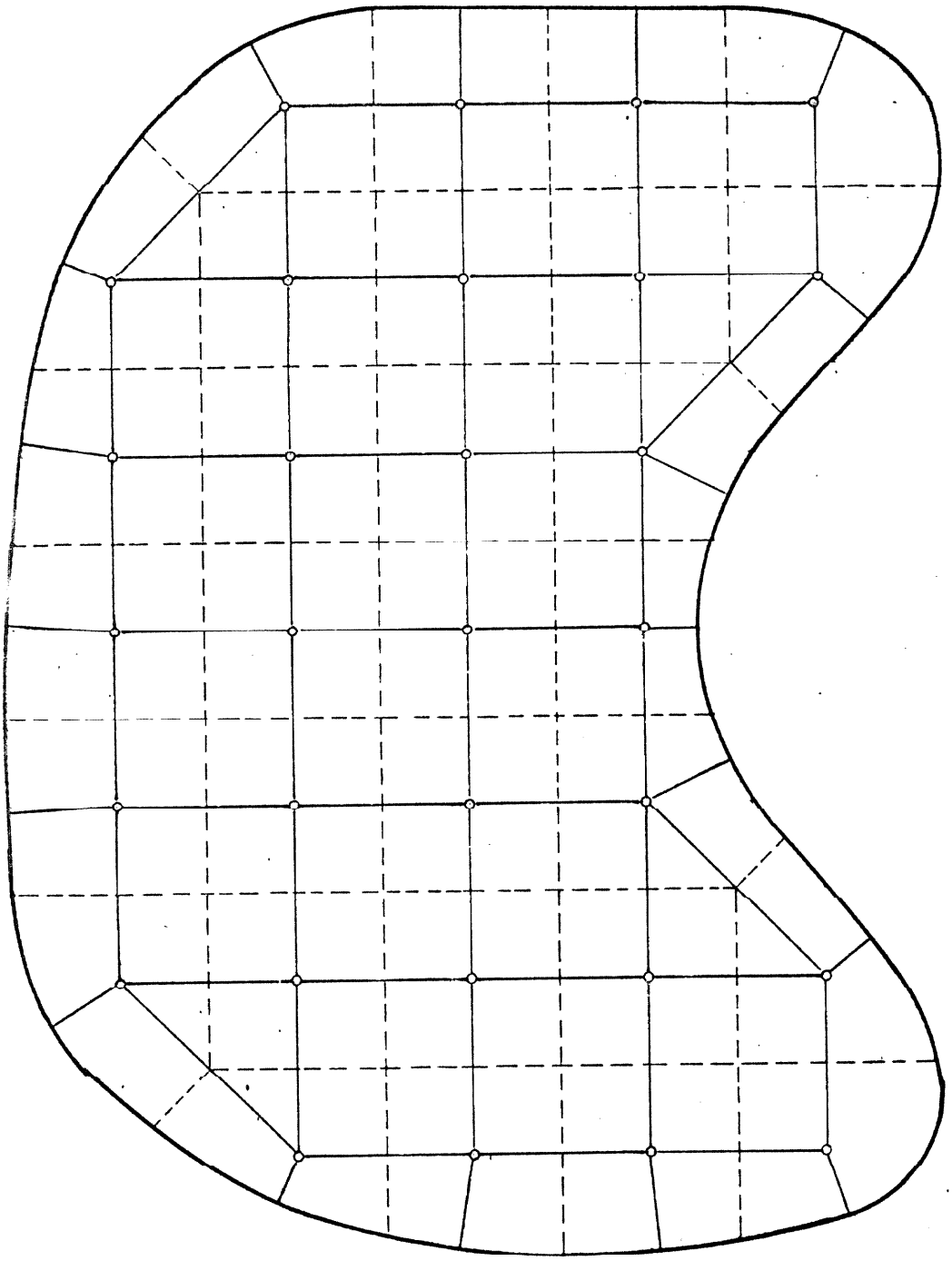
$$A_{i\ I} = \frac{C_{p\ I}\ \sigma_I\ (\text{Area})_{i\ I}}{C_{p\ 0}\ \sigma_0 L^2}$$

$$A_{i'\ II} = \frac{C_{p\ II}\ \sigma_{II}(\text{Area})_{i'\ II}}{C_{p\ 0}\ \sigma_0 L^2}$$

$$y_{ii'} = \frac{A_{ii'}}{k_0}\left[\frac{1}{\frac{\ell_{i\ I}}{k_I} + \frac{\ell_{i'\ II}}{k_{II}}}\right]$$

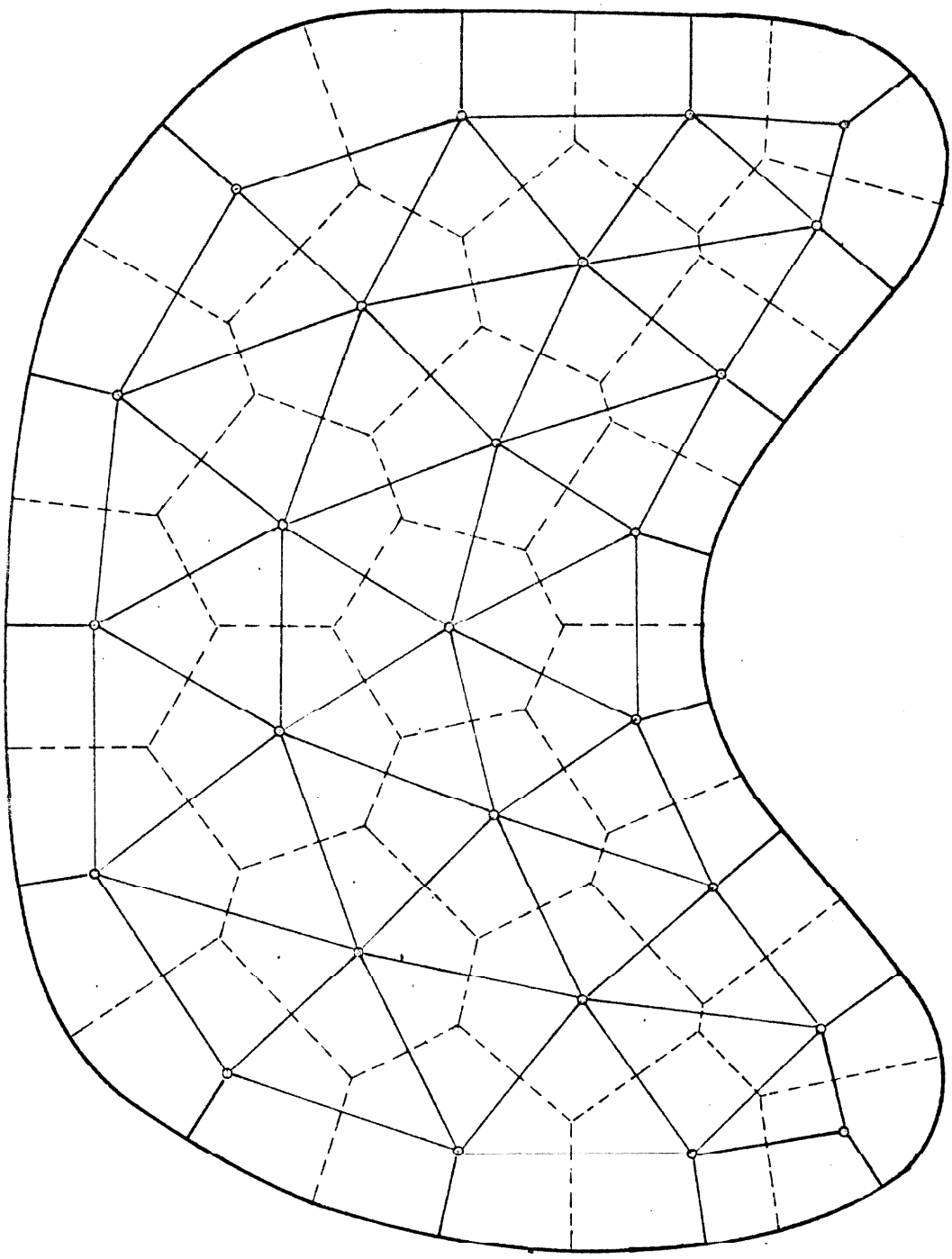
(Other conductances in similar fashion)

Figure V-16. Modified nodes for interfacial boundaries between uniform solids of different thermal properties.



Number of Adjacent Nodes:	18
Number of Interior Nodes:	12
Total Number of Nodes:	30

Figure V-17. Network to approximate solid with curved boundary; uses only square interior nodes.



Number of Adjacent Nodes	17
Number of Interior Nodes	9
Total Number of Nodes	26

Figure V-18. Network to approximate solid with curved boundary; distorted triangular network of distorted hexagonal nodes.

REFERENCES

1. Carslaw, H. S. and J. C. Jaeger, Conduction of Heat in Solids, Clarendon Press, Oxford (1947).

2. O'Brien, G. G., M. A. Hyman, and S. Kaplan, "A Study of the Numerical Solution of Partial Differential Equations," Journal of Mathematics and Physics, Vol. 29, pp. 223-251 (1950).

3. Richtmyer, R. D., Difference Methods for Initial-Value Problems, Interscience Publishers, Inc., New York (1957).

4. Todd, John, "A Direct Approach to the Problem of Stability in the Numerical Solution of Partial Differential Equations," Communications on Pure and Applied Mathematics, Vol. IX, pp. 597-612 (1956).

5. Franklin, J. N., "Numerical Stability in Digital and Analog Computation for Diffusion Problems," Journal of Mathematics and Physics, vol. XXXVII, No. 4, pp. 305-315 (Jan., 1959).

6. Varga, Richard, Matrix Iterative Analysis, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1962).

7. _____, "On Higher Order Stable Implicit Methods for Solving Parabolic Partial Differential Equations," Journal of Mathematics and Physics, Vol. XL, No. 3, pp. 220-231 (Oct., 1961).

8. MacNeal, R. H., "An Asymmetrical Finite Difference Network," Quarterly of Applied Mathematics, Vol. XI, No. 3, pp. 295-310 (Oct., 1952).

9. Longwell, P. A., Transient Temperature Distribution in Rocket Motors with Complex Cross-Section, NAVWEPS Report No. 7098, U. S. Naval Ordnance Test Station, Propulsion Development Department, China Lake, California (Dec. 1, 1959).

10. Forsythe, G. E. and W. R. Wasow, Finite-Difference Methods for Partial Differential Equations, John Wiley and Sons, Inc., New York (1960).

11. Perlis, Sam, Theory of Matrices, Addison-Wesley Publishing Co., Inc., Reading, Mass. (1952).

12. Jury, Eliahu I., Sampled-Data Control Systems, John Wiley and Sons, Inc., New York (1958).

13. Ragazzini, John R. and Gene F. Franklin, Sampled-Data Control Systems, McGraw-Hill Book Co., Inc., New York (1958).

14. Wilts, Charles H., Principles of Feedback Control, Addison-Wesley Publishing Co., Inc., Reading, Mass. (1960).

15. Crandall, S. H., Engineering Analysis, McGraw-Hill Book Co., Inc., New York (1956).

16. Faddeeva, V. N., Computational Methods of Linear Algebra, Dover Publications, Inc., New York (1959).

17. Lapidus, Leon, Digital Computation for Chemical Engineers, McGraw-Hill Book Co., Inc., New York (1962).

18. Ralston, Anthony and Herbert S. Wilf, Ed., Mathematical Methods for Digital Computers, John Wiley and Sons, Inc., New York (1960).

19. McDowell, E. R. H., "Transient Heat Transfer in Solids," Student Report No. 866, California Institute of Technology, Department of Chemical Engineering, Pasadena Calif. (May 27, 1960).

20. Clazie, Ronald, K. C. Hwang, and P. A. Longwell, "Applying Boundary Conditions to Graphical Solutions in Cartesian Coordinates," Ms. 5216, California Institute of Technology, Pasadena, Calif.

20a. _____, abbreviated form of above reference can be found in American Society of Mechanical Engineers, paper 63-HT-8.

21. Hildebrand, F. B., Methods of Applied Mathematics, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1952).

22. Muir, Sir Thomas, The Theory of Determinants in the Historical Order of Development, MacMillan and Co., Ltd., London (1923).

23. Rutherford, D. E., "Some Continuant Determinants Arising in Physics and Chemistry," Proceedings of the Royal Society of Edinburgh, Section A, Vol. LXII, pp. 229-236 (1943-1949).

24. _____, "Some Continuants Arising in Physics and Chemistry--II," Proceedings of the Royal Society of Edinburgh, Section A, Vol. LXIII, pp. 232-241 (1949-1952).

25. Aseltine, John A., Transform Method in Linear System Analysis, McGraw-Hill Book Co., Inc., New York (1958).

26. Churchill, Ruel V., Operational Mathematics, McGraw-Hill Book Co., Inc., New York (1958).

27. Dwight, Herbert Bristol, Tables of Integrals and Other Mathematical Data, The MacMillan Co., New York (1961).

28. Churchill, Ruel V., Complex Variables and Applications, McGraw-Hill Book Co., Inc., New York (1960).

29. Marcus, M., Basic Theorems in Matrix Theory, United States Department of Commerce, National Bureau of Standards, Applied Mathematics Series 57 (Jan. 22, 1960).

30. Todd, John, "The Condition of a Certain Matrix," Proceedings of the Cambridge Philosophical Society, Vol. 46, pp. 116-118 (1950).

31. _____, Ed., Survey of Numerical Analysis, McGraw-Hill Book Co., Inc., New York (1962).

32. Mickley, H. S., T. K. Sherwood, and C. E. Reed, Applied Mathematics in Chemical Engineering, McGraw-Hill Book Co., Inc., New York (1957).

33. Bellman, Richard, Introduction to Matrix Analysis, McGraw-Hill Book Co., Inc., New York (1960).

34. Federal Works Agency, Works Project Administration for the City of New York, Tables of Circular and Hyperbolic Sines and Cosines for Radian Arguments (1939).

35. Federal Works Agency, Works Projects Administration for the City of New York, Table of Arc Tangent (1942).

PROPOSITION I

Proposition

The stability of numerical methods which are of a stepping or iterative nature, and which are used to find approximate solutions for a wide variety of linear problems, is determined by the eigenvalues of a characteristic matrix. Specifically, if the approximate numerical method is a stepping method which can be formulated equivalent to

$$y_{n+1} = A\, y_n + f_n \qquad n \geq 0 \qquad \text{(PI-1)}$$

it is stable if an only if all the eigenvalues of the A matrix are less than one in absolute value. Under these circumstances the solution vector will converge to a vector that depends on f_n . In equation PI-1 y_{n+1}, y_n, and f_n are S-dimensional vectors, and A is a constant S x S matrix. Since the absolute value of the eigenvalues of a matrix are bounded by easily calculated norms of the matrix, a useful sufficient condition for stability of many approximate methods, including approximations for partial differential equations, can be found. In certain cases, some generalizations to non-linear equations can also be made.

Uses

Approximate methods which can be formulated equivalent to equation PI-1 include difference approximations of initial-value problems for both linear partial differential equations and systems of linear ordinary differential equations with constant coefficients, for example using Euler's method (PI. 1); and methods of iteration or successive

approximation for linear algebraic equations, for example, the Gauss-Seidel method (PI. 2). In many of such numerical methods equation PI-1 is used directly to step out the approximation. In these cases the matrix A is known explicitly, and its norms can be found using the definitions in equations III-10 and III-11. If the minimum of these norms is less than one, the stepped-out calculation must be stable. But if it is greater than one,the stepped-out calculation can be stable or unstable. For some numerical methods, although the stepping calculation is equivalent to equation PI-1, this equation is not used to carry out the approximation, and the A matrix is not known explicitly. In these cases some algebraic manipulation is required to find the relationship between the eigenvalues of matrix A and the other known matrices; however, these manipulations are often successful in giving useful stability criteria for the approximation.

In addition, a stability analysis for approximations to non-linear equations of the above type can also be made. For systems of non-linear ordinary differential equations, which could have been obtained from differencing of some variables in a non-linear partial differential equation, a difference equation for the propagation of the error based on the Jacobian matrix of the non-linear functions can be derived. In this case the constant matrix A in equation PI-1 is then related to the Jacobian evaluated at the time, and based on this linearization the stability of a method can be studied, providing elements of the y vector

do not change too much. In the same way, for a system of non-linear algebraic equations the A matrix can be taken as the Jacobian of the system evaluated close to or at the solution value of y. If then the starting estimated solution is sufficiently close to the exact solution, a method of successive approximations is stable and converges to the exact solution.

Advantages

The main advantages of using the eigenvalues of the A matrix for studying stability are:

(1) Such a study gives rigorous stability criteria which apply to all numerical approximations of a stepping nature for a wide variety of linear problems, and which apply regardless of the nature of the problem being approximated, or the origin of the system of difference equations. Further, any instability occurring in the calculation is seen to be caused by the eigenvalues of the stepping procedure, and not by round-off error. An exception is that, for some specific problems, where the initial conditions are such that all the eigenvalues with absolute values larger than one are associated with a zero weighting, round-off error can cause these weights to change slightly, and these eigenvalues become active, giving an unstable result.

(2) Stability criteria can be easily found because the absolute value of the maximum eigenvalue can be bounded by easily calculated

matrix norms. Moreover, the cause of an instability often becomes immediately obvious, and possible ways of obtaining stable solutions can sometimes be seen.

Although proofs of stability have been based on a matrix analysis of eigenvalues for methods of iteration and of approximations for ordinary differential equations, they have not been used (until recently) to derive stability criteria for approximate methods for partial differential equations. It is for this type of approximate solution that the matrix norms are particularly useful, since they reduce the problem of stability to a trivial calculation for many cases, instead of a relatively complicated Fourier analysis usually used. Further, the instability obtained in approximate solutions for partial differential solutions is shown to be caused in the same way as that which occurs in the solution of systems of ordinary differential equations or in successive approximate calculations.

Argument

The solution of the difference equation PI-1 is:

$$y_n = C\,Q^n C^{-1} y_0 + C \sum_{p=1}^{n} Q^{n-p} C^{-1} f_{p-1} \qquad \text{(PI-2)}$$

where Q is a diagonal matrix of eigenvalues q_j or Jordan boxes containing eigenvalues and C is the matrix of eigenvectors and principal vectors.

If the elements in y_n are not to go to infinity as n goes to infinity, when n is bounded, then all the eigenvalues q_j must be less than one in absolute value. Further, if the minimum norm is less than one in these circumstances, then the eigenvalues are all less than one in absolute value, and the procedure is stable.

References

PI. 1 Ralston, Anthony and Herbert S. Wilf, Ed., Mathematical Methods for Digital Computers, John Wiley and Sons, Inc., New York (1960).

PI. 2 Faddeeva, V. N., Computational Methods of Linear Algebra, Dover Publications, Inc., New York (1959).

PROPOSITION II

Proposition

The z-transform operational calculus (PII. 1) can be used to find complete analytic solutions of systems of linear difference equations which are represented by a partial difference equation together with boundary equations and for which an initial condition is known. This solution with z-transforms follows that of Laplace transforms for partial differential equations (PII. 2).

Uses

Partial difference equations are usually obtained by a direct differencing of a partial differential equation. From this partial difference equation a system of difference equations can be generated which, together with approximations for boundary conditions, can be used to step out a numerical approximate solution to the partial differential equation. The complete analytic solution of such a partial difference equation can be compared to a corresponding solution of the partial differential equation, leading to a detailed understanding of the accuracy of difference approximations to partial differential equations.

The solution of a partial difference equation is in terms of the eigenvector and eigenvalues of the characteristic matrix of the system of difference equations. These eigenvalues and eigenvectors are obtained

from a z-transform solution of the partial difference equation which, together with boundary equations, generates the matrix. Consequently this z-transform procedure can also be used to solve some matrix eigenvector-eigenvalue problems.

Advantages

The two main advantages of z-transform solutions of partial difference equations over other methods of solution, such as separation of variables or direct matrix methods, are:

(1) The complete analytic solution which fits the initial condition and which contains the particular solution is obtained by the procedure. Thus, no additional work is necessary to find or use orthogonality relationships to fit the initial condition, or to obtain a simple expression of a finite summation for the particular solution.

(2) The z-transform technique provides a methodical procedure for solving partial difference equations which requires a minimum of ingenious and adroit manipulations. The main obstacle to carrying the solution to completion is in finding a solution for the transformed partial difference equation, which is an ordinary difference equation for problems in two variables. The reason for this difficulty, which does not occur for Laplace transforms of partial differential equations, is that solutions of ordinary difference equations are not as well known as those of differential equations.

Argument

No special analytic justification appears to be necessary for applying z-transforms to partial difference equations. The only difference between their application here, compared to their customary application, is that the transformed equation is an ordinary difference equation or a partial difference equation of one less variable than the original equation, rather than an algebraic equation. However, this does not lead to any mathematical inconsistencies in the mathematical theory of the transform or in the inversion of the transform.

This technique has been used in this thesis (Chapter IV, section D) to solve the partial difference equation of diffusion in one space dimension in Cartesian coordinates. These solutions are the same as those obtained by other methods, and also give the same numerical results as are obtained by stepping out the difference solutions. Thus, there appears to be little doubt about the applicability of this method to the partial difference equation of diffusion or to partial difference equations which result from a direct differencing of partial differential equations which can be solved with Laplace transforms. However, as with any transform technique, partial difference solutions obtained with z-transforms should be checked to be sure of their validity.

References

PII.1 Ragazzini, John R., and G. F. Franklin, Sampled-Data Control Systems, McGraw-Hill Book Co., Inc., New York (1958).

PII.2 Churchill, Ruel V., Operational Mathematics, McGraw-Hill Book Co., Inc., New York (1958).

PROPOSITION III

Proposition

The sediment by extraction test for residual fuel oils, ASTM D473-59 (PIII. 1) when applied to petroleum tars resulting from the thermal cracking of straight-run residuums and catalytic cycle oils is a measure of the compatibility of the fuel oil with benzene rather than a measure of solid material that would separate in the absence of benzene. The fact that most such thermally cracked tars are incompatible with benzene accounts for the relatively poor repeatability and reproducibility of up to about 40 per cent error reported for this test.

Argument

Many authors, as summarized in Sachanen (PIII. 2, P. III. 3), have reported incompatibility of some hydrocarbons with residual fuel oils. For example, the addition of a low-boiling paraffin hydrocarbon, such as normal pentane, to either a straight-run or cracked residuum causes organic material to be precipitated or flocculated. Such an incompatibility also can occur when two fuel oils, one of straight-run stock, the other of cracked residuum, are mixed. In this case the deposit is usually shown by rapid fouling of heater surfaces (PIII. 2).

An explanation (PIII. 3) of this behavior is that the organic material precipitated is made up of long-chain molecules with polycyclic nuclei which were originally in colloidal solution or fine

suspension in the residual. These materials are precipitated as they are lyophobic with respect to paraffins. Some of the precipitated solids can be colloidally dispersed by the addition of aromatic hydrocarbons such as benzene. Although any distinction between these materials is arbitrary, solids precipitated with normal pentane, and which can be dispersed in benzene, are referred to as asphaltenes; those that cannot be dispersed in benzene are called carbenes. Cracked residuums contain more carbenes than the straight-run residuums and they are thought to be formed by dehydrogen polymerizations and condensation reactions during the cracking operation.

Secondly, with the advent of residuum strippers and catalytic cracking units in petroleum processing in the early 1950's, the feed to thermal crackers changed drastically. Instead of having a feed of straight-run oils boiling above about 650° F, the feed now is made up of straight-run residuum boiling above 900-1050° F and a heavy cycle oil from the catalytic cracking unit usually boiling at about 650° F. This heavy cycle oil contains mostly relatively high-boiling aromatic and unsaturated hydrocarbons instead of paraffinic hydrocarbons as before. The effect of this change in feed is that the thermal crackers must have more severe cracking conditions to approach gasoline yields obtained previously and that the residual thermally cracked tar probably contains more carbenes than before. Further, the carbenes produced in thermal cracking probably contain less hydrogen than previously,

and their molecules probably have longer chains with possibly more rings. However, the cracked oils formed by thermal cracking of these feeds are high-boiling aromatic hydrocarbons which act as effective peptizers. Thus the carbenes remain peptized and do not deposit after the cracking has stopped.

However, benzene, with the lowest molecular weight of the aromatic hydrocarbons, is not as effective a peptizer for these carbenes. When it is added, it dilutes and replaces the heavy aromatic hydrocarbons that are adsorbed by the carbenes, and the carbenes precipitate. As this process takes time the amount of carbenes which are precipitated is a function of both the time of contact and the relative amount of benzene added.

Finally, in the ASTM test the volume of dilution is reasonably well controlled, but the time of contact can vary widely. The extraction is carried out by refluxing benzene into the oil sample contained in a porous thimble. Thus, the time of contact is a function of the rate of refluxing, the viscosity of the oil, and the permeability of the porous thimble. This probably accounts for the rather large range of confidence limits shown in the ASTM procedure.

Suggestions

Since sediment by extraction is a specification of No. 6 residual fuel oil (ASTM D396-62T), an effort should be made to remove the ambiguity from its meaning and to make the test more reproducible. It probably should be decided if the fuel oil compatibility with a light aromatic solvent such as benzene is related to an important fuel oil quality. Fuel oil compatibility with other fuel oils can be measured more directly using blends of the test fuel oil with two standard fuel oils of different blends in the NBTL heater tube test (PIII.3). Further, it has been pointed out that fuel oils containing aromatic stocks such as

catalytic cycle oils are compatible with most any other fuel.

If it is decided that the compatibility of the fuel with benzene is important or that this property correlates with other important qualities of fuel oil, then a better test should be devised. A possibility is that the oil could be diluted with benzene in a fixed ratio and allowed to age a certain time before filtering through a porous thimble. Filtration immediately after dilution would give a good indication of solid material originally separated in the oil. After aging for as much as 12 to 24 hours at a fixed temperature the amount of sediment would be a measure of the incompatibility of the fuel with benzene. Neither of these amounts of sediment would agree even in order of magnitude with the sediment as measured by the sediment by extraction test ASTM D473-59.

References

PIII.1 ASTM Standards on Petroleum Products and Lubricants (1961) and supplement (1962).

PIII.2 Sachanen, A. N., The Chemical Constituents of Petroleum, Reinhold Publishing Corp., New York (1945).

PIII.3 Guthrie, Virgil B., Ed., Petroleum Products Handbook, McGraw-Hill Book Co., Inc. New York (1960).

PROPOSITION IV

Proposition

In the teaching of science and engineering, the transient and evolutionary character of all scientific theories should be emphasized. It should also be pointed out that not only are the new, more speculative theories open to question, but often even the basic concepts of accepted theories must be significantly modified, or rejected, in light of new experimental data, or because of entirely new theories better able to account for the phenomena. Moreover, theories rejected earlier are sometimes revived in a modified form. This transient nature of theories should be contrasted to the fact that accurate measurements, observations, and descriptions of physical phenomena are always valid.

Argument

Although most scientists would agree that all theories are always tentative, in the past scholars have taken the attitude that some theoretical hypotheses were unchallengeable. Such attitudes undoubtedly retarded the advance of science significantly.

One of the most glaring examples of this attitude was that of the ancient Greeks to the effect that all planetary movement is described by motion compounded from circular motion. This hypothesis was developed by Ptolemy in the second century A. D. to describe planetary motion with a fairly complicated method using epicycles. This technique, with the earth in a fixed position, gave relationships which agreed to within the accuracy of the astronomical observations then available, and lasted virtually unchanged for almost 1400 years until Copernicus simplified the theory by assuming that the

earth moved around the sun. Finally, Kepler, in the early seventeenth century, using the accurate data of Tycho Brahe, discovered the elliptical nature of the orbits.

This discovery, which in all probability helped Newton to formulate his laws of mechanics, undoubtedly was retarded because the scholars themselves considered only circular, or compounded circular, motions. The refusal to depart from one line of thought must have been due in part to limitations which they unconsciously placed on themselves. Such unconscious limitations often can prevent one from finding an obvious solution. For example, in the problem of trying to arrange six matches to form four triangles, the false assumption is often made that this must be done in two dimensions. Whereas, if the problem is expanded into three dimensions the solution becomes trivial.

Another explanation of the long time which was required for the solution of the planetary orbits is given in Astronomy , by Fred Hoyle (PIV.1). He remarks that if the sun-centered solar system with circular planetary orbits, postulated tentatively in the second century B.C. by Aristarchus, had been accepted and used, the history of astronomy and of science would have been changed radically. The reason that this theory was rejected then and for several hundred years thereafter was that it did not describe planetary motion as accurately as the developed epicycle theory. This leads Hoyle to state that, "Here we have a remarkable example to show that it does not always pay to know too much about the facts of the situation, " * and "The Greeks . . . were attempting to represent phenomena that were far too complicated for them. " *

---

* Fred Hoyle, Astronomy , Doubleday and Co., Inc., Garden City, New York (1962), p. 91.

However, it was not too complicated for Kepler who had even more accurate data to work with. Indeed, Hoyle states that " . . . Kepler had a far greater respect for facts than the average scientist has . . . " * Thus, it appears that it was not wholly the accurate data that retarded the understanding of planetary motion, but a good portion of the responsibility must go to the unquestioning acceptance of the hypothesis of circular motion.

Although our understanding of both scientific phenomena and scientific theory have greatly increased since the days of Kepler, we must be watchful lest any of even our most successful theories become dogma. However, even today, one hears of data which have not been reported because of lack of agreement with current theory. Indeed, if the data are clearly in error they should be reported as such; if, however, they are accurate, it is most important that they be reported so that the theory not be used inappropriately, and so that the need for a better theory be recognized. Moreover, in many publications empirical relationships are often presented in an apologetic manner. It is rather the theorist who should apologize for not having a satisfactory theory to explain the relationship. In fact, many of our theories have been developed as explanations for such empirical relationships. Feinberg and Goldbar, writing on the conservation laws of physics (PIV.2), express the opinion that the proper relationship between theoretical and empirical results can become obscured today. They comment, " Consequently, there has been a tendency to forget that the basis of the conservation laws is, after all, empirical, . . . "

The predicament of the ancient Greeks, of trying to represent facts too complicated for them, and facts which were so accurate that partially developed theories were not satisfactory, also is the predicament of modern engineers

---

* Ibid., p. 119.

and scientists. Further, Hoyle's hope that "discordant facts will not appear until worthwhile theories have had a chance to establish themselves" * is probably almost never possible, and moreover, it is only these "discordant facts" which indicate which theories are worthwhile and which allow us to use the theories intelligently.

The way to maintain scientific progress is not to suppress these facts, but for those involved in both the theory and practice of science and engineering to develop a proper perspective. This means emphasizing that too much significance can be attached to a successful theory and that " no theory ever proposed has been found ultimately to fit all the facts . . ." * With these ideas in mind no useful theory should be prematurely rejected nor should any theory, concept, or hypothesis, even though very successful, be elevated to dogma.

Suggestions

Several suggestions to bring out the tentative nature of theory are:

(1) The basic texts that first introduce students to scientific theory should give an indication of the evolution of the theories. They should not present the material in such a fashion that it appears to the student that all the answers have been found, but they should mention the discordant facts as well as the successes of the theories.

(2) On advanced levels, not only should the more complicated theories be presented, but the contradictions between these theories and empirical results should be emphasized. Further, it should be pointed out that many of these theories are almost useless in obtaining results for practical problems.

---

* Ibid., p. 91.

(3) In history courses, the evolution of scientific theories and of technology should be made a part of the material covered. Not only should mention be made of the significant effect of these ideas on social, economic, and political factors, but also the effect of the evolutionary changes in science upon science itself should be brought out. In philosophy courses additional emphasis should be given to exploring proper relationships and perspectives between theory and empiricism.

References

PIV.1 Hoyle, Fred, Astronomy, Doubleday and Co., Inc., Garden City, New York (1962).

PIV.2 Feinberg, Gerald and Maurice Goldbar, "The Conservation Laws of Physics," Scientific American, Vol. 209, No. 4 (Oct., 1963), p. 36.

PROPOSITION V

Proposition

In a recent study of the steady state of a jacketed tubular reactor with a packed bed, in which an exothermic chemical reaction is occurring under flow conditions, Barkelew (PV.1) presents relations which show conditions which give an "unstable" reactor design. In an earlier study of both a linearized transient solution and the steady-state solution of an exothermic chemical reaction occurring in a jacketed empty tubular reactor, Bilous and Amundson (PV.2) show that a "parameter sensitivity" can occur.

The proposition is that the "instability" of Barkelew and the "parameter sensitivity" of Bilous are one and the same. Further, Barkelew's relations are suitable for empty tubes, and only under certain circumstances are they actually valid for packed-tube reactors. A brief analysis is also presented that should allow Barkelew's results to be generalized for other conditions in packed tubes and should give additional information about controlling the reaction.

Argument

In both studies, the equations are for a plug-flow reactor:

$$C_p \sigma \frac{\partial T}{\partial \theta} = - A h (T - T_j) - C_{p\,f} G \frac{\partial T}{\partial z} + Q R \qquad \text{(PV-1)}$$

$$\sigma_f \frac{\partial x}{\partial \theta} = - R - G \frac{\partial x}{\partial z} \qquad \text{(PV-2)}$$

for the first order reaction

$$A \longrightarrow B$$

where C_p = heat capacity

σ = density

A = tube surface area per unit volume

h = heat-transfer coefficient

T = temperature of fluid in cross-section

G = mass velocity

θ = time

z = distance along tube

Q = heat of reaction

R = rate of reaction

x = fraction of reactant in fluid

The difference between the equations is that the unsubscripted C_p and σ in the above equations are averaged values for both the fluid and the bed for the packed reactor; for the empty tube they are just the fluid quantities.

The expression used for the reaction rate in Barkelew's study can be applied to other than first-order reactions; but his reaction rate uses an exponential temperature dependency that is the same as the Arrhenius expression only when there is little difference between the temperature of the reactants and of the coolant. Bilous uses the conventional Arrhenius temperature dependency for the rate.

In the Bilous study the equations were solved for the steady-state temperature and concentration distributions using an analog computer. These distributions, under some conditions, showed that the maximum temperature in the reactor would increase very rapidly with small increases in the wall or jacket temperature or with small reductions in the heat-transfer coefficient.

On the basis of this behavior, Bilous defines "parametric sensitivity" as the phenomena occurring when " . . . reactor operation, without being unstable in the literal sense, will be extremely sensitive to small changes in the operating characteristics of the system." He further shows analytically that this "parametric sensitivity" is related to the transient response of the system to either a step change or a continuing sinusoidal input to the reactor. This transient response is based on a linearization of the above equations about the steady-state solution.

Barkelew numerically integrated the dimensionless equations equivalent to equations PV-1 and PV-2 with their time derivatives equal to zero for several hundred combinations of several dimensionless parameters. His steady-state temperature distributions show a behavior like that of the analog solution, and Barkelew states that " . . . there are indeed regions of extreme sensitivity to changes in parameters." His precise definition of "stability" is based on a graph that relates the maximum temperature to dimensionless ratios.

Thus, as both studies are based on the same partial differential equations, the steady-state solutions would be expected to be similar. Further, that Barkelew's "stability" is essentially the same as Bilous' "parametric sensitivity" can be seen by using Barkelew's defining graph to predict the conditions of parametric sensitivity for the problems solved by the analog. Such a numerical check does predict the parametric sensitivity of case B in Bilous. A similar calculation for Bilous' case A does not check because of the difference in Barkelew's expression for the rate of reaction mentioned above.

In both studies, the partial differential equations solved neglect several effects, including the radial temperature distribution. The temperature at each cross-section of the tube is assumed to be constant except for the drop

across a thin film at the wall. Although this is a good assumption for the empty-tube reactor with large Reynolds numbers, a significant radial temperature distribution can occur in a packed-bed reactor. As Barkelew points out, this error for the packed beds would tend to make an unstable reactor appear to be stable.

A logical extension of Barkelew's analysis that avoids this ambiguity would be to estimate a radial temperature distribution at the cross-section of maximum temperature. If we assume that at this cross-section the temperature gradient along the tube is zero, then replacing the term $-h A (T-T_j)$ with a radial energy flow term $(ke/r)(\partial/\partial r)\, r(\partial T/\partial r)$, there is obtained

$$ke\left[\frac{d^2T}{dr^2} + \frac{1}{r}\frac{d^2T}{dr}\right] + Q R = 0 \qquad (PV-3)$$

where ke = average thermal conductivity of fluid and bed

r = radial coordinate

The boundary equations for equation PV-3 are:

at the tube wall $$ke\frac{dT}{dr} = h(T-T_j) \qquad (PV-4)$$

at the center $$\frac{dT}{dr} = 0 \qquad (PV-5)$$

If one assumes that the reaction rate is either constant or a linear function of temperature, equation PV-3 can be solved analytically giving an estimated radial temperature distribution for that cross-section. In solving this equation the maximum temperature and concentration at this cross-section can be found from Barkelew's relationships and used in estimating the reaction rate. From the solution to equation PV-3 it should be apparent if a significant radial

temperature distribution does occur and if Barkelew's results wrongly predict a stable reactor design.

Conclusions

The fact that "parametric sensitivity" and "unstable" conditions are the same is important, because then Barkelew's results for a fairly wide range of kinetics and parameters are related to the transient response for empty tubular reactors. It should be pointed out that the transient response of a packed-bed reactor is expected to occur at a different rate than that of an empty reactor, even though the steady-state conditions of both might be similar. Also, a fairly simple method of estimating the radial temperature distribution at the cross-section of maximum temperature for a packed-bed reactor has been proposed to check the assumption of negligible radial temperature distribution.

References

PV.1 Barkelew, C. H., "Stability of Chemical Reactors," Chem. Eng. Progr. Symposium Series, No. 25, 55, 37 (1959).

PV.2 Bilous, O. and N. R. Amundson, "Chemical Reactor Stability and Sensitivity: II, Effect of Parameters on Sensitivity of Empty Tubular Reactors," A.I.Ch.E. Journal, Vol. 2, No. 1, p. 117.