(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 thc 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, ..., S; and for methods C and F based on mesh $\Delta\xi/2$, m = 1/2, 3/2, ..., 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 m^{th} 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 interrelationships 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.

$$T_P - t_P = 0 \tag{IV-210}$$

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 thc problcms 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 n^{th} 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 m^{th} point at the n^{th} 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 factorsratio $e^{-\psi_j^2 r}/q_j$ 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 eigenvector 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

$$V = J\left|(g_J - a_J)c_{mJ} + g_J\left\{c_{mJ} - b_J(m\Delta\xi)\right\} + g_J\, c_{mJ}\left\{1 - \left(\frac{e^{-\nu_J^2}}{q_J^{1/\Delta\tau}}\right)^{\tau_0}\right\}\right| |q_1^n| \geq |v_{m,n}| \tag{IV-218}$$

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

$$\frac{V}{2Jq_1^{n_0}} \geq \left|g_J\, c_{mJ}\left\{1 - \left(\frac{e^{-\nu_J^2}}{q_J^{1/\Delta\tau}}\right)^{\tau_0}\right\}\right| \tag{IV-220}$$

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:

$$\min_{1 \leq j \leq J} |g_j \, c_{mj} \, q_j^{n_0}| \gg \operatorname{MAX}_{J < j \leq S} |g_j \, c_{mj} \, q_j^{n_0}| \qquad \text{(IV-221)}$$

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.

$$V \gg g_S \, c_{mS} \, Y^{n_0} \qquad \text{(IV-222)}$$

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 J^{th} 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

$$\sin S\alpha_j = \sin S\psi_j = 0 \qquad \text{(IV-224)}$$

which is satisfied by

$$\psi_j = \alpha_j = \frac{(j-1)\pi}{S} \qquad j = 1, \ldots, \infty \qquad \text{(IV-225)}$$

For an infinite H (problem II) by taking limits the equations become

$$\cos S\psi_j = \cos S\alpha_j = 0 \qquad \text{(IV-266)}$$

which is satisfied by

$$\psi_j = \alpha_j = (2j-1)\pi/2S \qquad j = 1, \ldots, \infty \qquad \text{(IV-227)}$$

Thus each function is zero at an even multiple of π/2S, including zero, and infinite at an odd multiple of π/2S.

The continuous function in Figure IV-7 shows that H(ψ) rises with increasing slope from 0 at a ψ of 0 to ∞ at a ψ of π/10. Between ψ's of π/10 and π/5 the function H(ψ) is negative and is not shown. At a ψ of π/5, this function H(ψ) becomes zero and again rises to infinity at a ψ of 3π/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,-ψ_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

$$\alpha_6 = \pi + \sqrt{-1}\,\mathscr{J}(\alpha_6) \qquad \text{(IV-228)}$$

Substituting this into the characteristic equation for method G from Table IV-12 and using complex variable theory, the equation is modified to

$$5[\sinh \mathscr{J}(\alpha_6)]\ [\tanh 5\mathscr{J}(\alpha_6)] = H \qquad \text{(IV-229)}$$

A plot of this as a function of α is in Figure IV-10. The function $\{5[\sinh \mathscr{J}(\alpha_6)]\ [\tanh 5\mathscr{J}(\alpha_6)]\}$ rises from 0 at an $[\mathscr{J}(\alpha_6)]$ of zero to infinity at an $[\mathscr{J}(\alpha_6)]$ 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 ν (equation IV-153),is more convenient to use than α. It is defined by:

$$\mu = \alpha S \tag{IV-230}$$

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 Sπ. 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 ν and μ instead of ψ and α, and using straightforward algebraic manipulations, there is obtained:

$$(\nu_j-\mu_j)(\nu_j+\mu_j) = (\nu_j^2-\mu_j^2) = \mu_j(D_j-\mu_j)\left[\frac{\mu_j^2-\nu_j^2}{\mu_j(\mu_j-\nu_j\ \tan\nu_j\ \mathrm{ctn}\,\mu_j)}\right] \qquad \text{(IV-231)}$$

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

or, expanding the second bracketed term using trigonometric series, which shows that $(\mu_j^2-\nu_j^2)/\mu_j$ is a factor of $(\mu_j-\nu_j\ \tan\nu_j\ \mathrm{ctn}\,\mu_j)$ we obtain:

$$(\nu_j-\mu_j)(\nu_j+\mu_j) =$$

$$\mu_j(D_j-\mu_j)\left[1-\frac{\nu_j^2}{3}-\frac{\nu_j^2}{45}(\mu_j^2+\nu_j^2)-\frac{\nu_j^2}{945}(2\mu_j^4-5\nu_j^2\mu_j^2+2\nu_j^4\cdots\right]$$

(IV-232)

The expanded term depends only slightly on S because the order of magnitude of μ_j does not change with S, and ν_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 $(\nu_j^2 - \mu_j^2)$ or $(\nu_j - \mu_j)$ 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
$$D_j - \mu_j = \frac{S \sin \frac{\mu_j}{S}}{1-2r(1-\cos \frac{\mu_j}{S})} - \mu_j \qquad \text{(IV-234)}$$

$$= \frac{\mu_j^3}{6S^2}(6r-1) + \frac{\mu_j^5}{120S^4}(1-30r \left\{1-4r\right\}) \cdots$$

Graphical A $\quad r = 1/2$

$$D_j - \mu_j = S \tan \frac{\mu_j}{S} - \mu_j = \frac{\mu_j^3}{3S^2} + \frac{2\mu_j^5}{15S^4} \cdots \qquad \text{(IV-235)}$$

$$|\mu_j| < \frac{S\pi}{2}$$

$$\text{C} \qquad D_j - \mu_j = 2S \tan \frac{\mu_j}{2S} - \mu_j = \frac{\mu_j^3}{12S^2} + \frac{\mu_j^4}{120S^4} \cdots \qquad \text{(IV-236)}$$

$$|\mu_j| < S\pi$$

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 $(\nu_j^2-\mu_j^2)$ 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 ν_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 $(\nu_j^2-\mu_j^2)$ and $(\nu_j-\mu_j)$ 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:

$$\nu_{j+1} \simeq \mu_{j+1} \simeq \mu_j + \pi \simeq \nu_j + \pi \qquad \text{(IV-245)}$$

The first root μ_1 or ν_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 n^{th} power, and determine how the transient solution changes with time. In the solution to the partial differential equation the factors to the n^{th} 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

$$(1 - \cos\alpha_j) \le 2 \qquad\qquad j = 1, \ldots, S \qquad\qquad \text{(IV-248)}$$

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

$$\frac{1-4r(1-\gamma)}{1+2r\gamma} = -\Upsilon \le q_{min} \qquad\qquad \text{(IV-249)}$$

where

$$\Upsilon > 0 \qquad\qquad \text{(IV-250)}$$

r and γ must satisfy the following equivalent inequalities:

$$r[1 - \gamma(\Upsilon + 1)] \le \frac{\Upsilon + 1}{4} \qquad\qquad \text{(IV-251)}$$

or

$$\gamma \geq \frac{4r - (1+\Upsilon)}{4r\,(\Upsilon+1)} \qquad \text{(IV-252)}$$

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 Υ/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 $-\Upsilon$ for large S. This is the basis for a prior statement in Chapter III that $|\lambda_{min}|$ for a Υ/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 Υ'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 Υ 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 Υ/A matrix is $4S^2$ or less, this graph can be used to find Υ 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 Υ of zero, the sufficient condition for non-oscillatory solutions, this asymptote is one,$(\gamma=1)$; for an Υ 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 Υ is very close to q_{min}. First, increasing the weighting γ at a constant difference modulus r reduces Υ rapidly for values of Υ close to one; as Υ approaches zero, a reduction in γ does not reduce Υ as much as before. But reducing r at constant weighting factor lowers Υ less when Υ 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 $|q_{min}|$, Υ, must also be decreased to keep $\Upsilon^{\tau_0/\Delta\tau}$ 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

$$q_{min} = q_{S+1} = \frac{1-2r(1-\gamma)[1+\cosh \mathscr{g}(\alpha_{S+1})]}{1+2r\gamma[1+\cosh \mathscr{g}(\alpha_{S+1})]} \qquad \text{(IV-253)}$$

where $\mathscr{g}(\alpha_{S+1})$ satisfies equation IV-229.

However, a good simple upper bound for $[\cosh \mathscr{g}(\alpha_{S+1})]$ 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 (γ=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

$$g_j\, c_{mj}(\cos\alpha_j)^n + g_{S+1-j}\, c_{m\ S+1-j}(-\cos\alpha_j)^n =$$

$$\left[g_j + (-1)^n\, g_{S+1-j} \left(\frac{c_{m\ S+1-j}}{c_{mj}}\right)\right] c_{mj}(\cos\alpha_j)^n \qquad \text{(IV-256)}$$

$$0 \leq m \leq S$$

$$n \geq 0$$

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, $-\nu_j^2$, and $[(1/\Delta\tau)\ \ln q_j]$. The difference between these quantities is the natural logarithm of the ratio $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ 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 $e^{-\upsilon_j^2 n\Delta\tau}$. 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.

$$\lim_{\substack{\Delta\tau \to 0 \\ n \to \infty \quad n\Delta\tau \to \tau}} q_j^n \longrightarrow e^{-2S^2\left(1-\cos\frac{\mu_j}{S}\right)\tau} \qquad \text{(IV-258)}$$

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.

$$\lim_{\substack{\Delta\tau\to 0\\ n\to\infty \quad n\Delta\tau\to\tau\\ S\to\infty}} q_j^{\,n} \longrightarrow e^{-\nu_j^2\tau} \qquad \text{(IV-259)}$$

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:

$$\ln\frac{e^{-\nu_j^2}}{q_j^{1/\Delta\tau}} = (\mu_j-\nu_j)(\mu_j+\nu_j) + \mu_j^4\left[(\tfrac{1}{2}-\gamma)\Delta\tau - \frac{1}{12S^2}\right] +$$

$$+\,\mu_j^6\left[(\tfrac{1}{3}-\gamma+\gamma^2)(\Delta\tau)^2 - \frac{(\frac{1}{2}-\gamma)\Delta\tau}{12S^2} + \frac{1}{360S^4}\right]\cdots \qquad \text{(IV-260)}$$

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 $\ln(e^{-\nu_j^2}/q_j^{1/\Delta\tau})$ 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 $(\mu_j-\nu_j)(\mu_j+\nu_j)$, 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 ν 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 $\frac{1}{2}$.

$$1 \leq \gamma < \tfrac{1}{2} \qquad \text{(IV-261)}$$

Further if the weighting γ is selected according to

$$\gamma_o = \tfrac{1}{2} - \frac{1}{12S^2\Delta\tau} = \frac{6r-1}{12r} \qquad \text{(IV-262)}$$

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 $e^{-\upsilon_j^2}/q_j^{1/\Delta\tau}$ is less than zero, the ratio $e^{-\upsilon_j^2}/q_j^{1/\Delta\tau}$ 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 $(e^{-\nu_j^2}/q_j^{1/\Delta\tau})$ to the j^{th} trigonometric roots and the differencing parameters. This expansion was shown to be suitable for estimating $e^{-\nu_j^2}/q_j^{1/\Delta\tau}$ 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 $(\mu_j^2 - \nu_j^2)$ 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 Υ 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 Υ 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:

$$0 \le H_0 \text{ , } H_S \le \infty \qquad \text{(IV-263)}$$

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 j^{th} eigenfunction $b_j(\xi)$ gives the relative weighting of the j^{th} continuous damping factor $e^{-\psi_j^2 r}$ for all points ξ within the solid; the j^{th} 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 semilogarithmic graph for the corresponding terms in the transient solutions for the m^{th} point is the product of the j^{th} initial vector component times the m^{th} component of the j^{th} 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 j^{th} damping factor in the solution for the m^{th} 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 $(\mu_j+\nu_j)$ with $2\mu_j$:

$$c_{mj}-b_j(m\Delta\xi) \simeq$$

$$(\nu_j-\mu_j)(1-\frac{m}{S})\left[1-\frac{(\mu_j-\nu_j)^2(1-\frac{m}{S})^2}{6}\right]\sin(1-\frac{m}{S})\mu_j \qquad \text{(IV-267)}$$

(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 j^{th} transient solution term in the equation for the m^{th} 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 $(\mu_j-\nu_j)$ 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 $(\mu_j-\nu_j)$; 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

$$\cos(S-m)\alpha_{S+1} = (-1)^{S-m} \cosh(S-m) \mathscr{J}(\alpha_{S+1}) \qquad \text{(IV-268)}$$

Since the hyperbolic cosine is large for large $[\mathscr{J}(\alpha_{S+1})]$ 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 $(\mu_j - \upsilon_j)$ 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 S^{th} 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 n^{th} 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 m^{th} component of its eigenvector $g_j c_{mj}$ determines the intercept on the semi-logarithmic graph of the j^{th} term in the transient solution for the m^{th} 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:

$$\sigma(\xi)d\xi \;=\; d\xi \;= dV \qquad \text{(IV-270)}$$

where dV is a dimensionless differential volume.

The integration is:

$$a_j = \int_0^1 b_j(\xi) \,[T(\xi,0) - T_P(\xi,0)]\, d\xi \qquad \text{(IV-271)}$$

The summation of the eigenvectors to find the initial vector coefficients g_j is

$$g_j = \sum_{m=\frac{1}{2}}^{S-\frac{1}{2}} c_{mj} \,[t_{m,0} - t_{P\;m,0}]A_m \qquad \text{(IV-272)}$$

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, ..., 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 j^{th} effective initial vector component with a component of the j^{th} 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:

$$g_j \, c_{mj} \, q_j^{\,n} + g_{S+1-j} \, c_{m\,S+1-j} \, q_{S+1-j}^{\,n} =$$

$$\left[g_j + \left(\frac{q_{S+1-j}}{q_j} \right)^n \frac{c_{m\,S+1-j}}{c_{mj}} \right] c_{mj} \, q_j^{\,n} \qquad \text{(IV-278)}$$

$$1 \leq j \leq \frac{S}{2}$$

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,

$$V = J \left| g_J\, c_{mJ} - a_J\, b_J(m\Delta\xi) + g_J\, c_{mJ} \left\{ 1 - \left(\frac{e^{-\nu_J^2}}{q_J} \right)^{\tau_0} \right\} \right| q_1^{n_0} \geq |v_{m,n}| \qquad \text{(IV-282)}$$

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

$$\Phi \frac{\mu_J}{S^2} > | g_J \, c_{mJ} - a_J \, b_J(m\Delta\xi)| \qquad \text{(IV-286)}$$

The ratio $e^{-\nu_J^2} / q_J^{1/\Delta\tau}$ can be found from equation IV-260,

$$\ln \frac{e^{-\nu_J^2}}{q_J^{1/\Delta\tau}} = \mu_J^2 - \nu_J^2 + \mu_J^4 \left[(\tfrac{1}{2} - \gamma)\Delta\tau - \frac{1}{12S^2} \right] + \qquad \text{(IV-287)}$$

$$+ \mu_J^6 \left[(\tfrac{1}{3} - \gamma - \gamma^2)(\Delta\tau)^2 - \frac{(\tfrac{1}{2} - \gamma)\Delta\tau}{12S^2} + \frac{1}{360S^4} \right] \cdots$$

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 $|q_{min}|$, Υ, is found for method C from Figure IV-14 or equation IV-252, which is, when solved for Υ:

$$\Upsilon = \frac{4r(1-\gamma)-1}{4r\gamma + 1} \geq |q_{min}| \qquad \text{(IV-288)}$$

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 J^{th} 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 Υ from equation IV-288 obtaining:

$$\frac{V}{\Phi} >> \left[\frac{4S^2\Delta\tau(1-\gamma)-1}{4S^2\Delta\tau\gamma + 1} \right]^{\tau_0/\Delta\tau} = (\Upsilon)^{\tau_0/\Delta\tau} \qquad \text{(IV-295)}$$

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 ½ 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:

$$\frac{V}{\Phi} >> \Upsilon^{\tau_0/\Delta\tau_{Ex}} = (4S^2\Delta\tau_{Ex} - 1)^{\tau_0/\Delta\tau_{Ex}} \qquad \text{(IV-303)}$$

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 J^{th} 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

$$r_{Im} > 3.5\, r_{Ex} \tag{IV-305}$$

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:

$$\frac{V}{\Phi} >> \left[\frac{3S^2\Delta\tau - 1}{3S^2\Delta\tau + 1}\right]^{\tau_0/\Delta\tau} \tag{IV-306}$$

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 $e^{-\upsilon_J^2}/q_J^{1/\Delta\tau}$ is calculated and the accuracy of the J^{th} 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 thc 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:

$$N_{Im} = \frac{7S^3\tau_1}{r_{Im}} = \frac{7S\tau_1}{\Delta\tau_{Im}} \qquad \text{(IV-307)}$$

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.