8. TIME INTEGRATION OF A SINGLE-DEGREE-OF-FREEDOM SYSTEM

8.1 INTRODUCTION

In the previous section we described the finite element approach as a method for obtaining approximate solutions to ordinary differential equations in which the independent variable is the spatial variable, x. For transient problems another independent variable, the time, t, is introduced. With both x and t as independent variables, the problem becomes two-dimensional and both a spatial discretization of x and a temporal discretization of t is required. Finite elements and finite differences have been used as methods for spatial discretization, and either can be used for temporal discretization as well. Sometimes the finite element form is used for the space variable and a finite difference approach is used for time. To make the matter even more confusing, a time integrator is often introduced. What we show in this chapter is that a time integrator approach is equivalent to the use of a finite difference approach. Again, the notation can be confusing so we take the time to define terms and to give an example of an algorithm that is fairly standard.

Instead of moving directly to a two-dimensional problem, we step back and consider a procedure for obtaining approximate solutions to an ordinary differential equation in time. The domain of the independent variable, t, is from zero to infinity rather than the finite domain normally associated with boundary-value problems involving a spatial parameter. Finite elements could be used to discretize in time but here we introduce instead a time integrator. There are a large number of integrators in use. We focus on one class of integrator, the general trapezoidal rule, and provide a thorough analysis of the algorithm. The integrator, itself, is useful in conjunction with finite element spatial discretization but, more importantly, the terminology and concepts of the analysis are fundamental and common to procedures used to analyze all integrators. Therefore, the steps of performing the analysis of the integrator are an essential part of this chapter. However, for those readers who wish to skip the analysis, the appropriate point is indicated in the text where the description of the integrator is complete. The next chapter brings the finite element method and the time integrator together for the transient heat conduction problem.

8.2 SINGLE-DEGREE-OF-FREEDOM SYSTEM

Consider the first-order differential equation in time which can loosely be considered as a one-degree-of-freedom approximation to the one-dimensional heat conduction problem of Section 1.2:

$$C\dot{T} + KT = Q \tag{8.2-1}$$

where C is the heat capacitance, K the conductivity, and Q the forcing function. Suppose Q is a function of time, the coefficient functions C and K are constants, and the initial condition is $T(0) = T_o$. To provide a direct correlation with the multi-degree of freedom formulation which will come later in which C and K are replaced with matrices, we choose to present much of the analysis of the algorithm in terms of these parameters rather than the parameter, Ψ, defined as follows:

$$\Psi = \frac{K}{C} \tag{8.2-2}$$

The solution to the homogeneous problem governed by Q = 0 for an initial value of $T = T_0$ at $t = 0$ is

$$T = T_o\, e^{-\Psi t} \tag{8.2-3}$$

A numerical algorithm provides an approximate solution T^k at discrete times

$$t^k = ks\,, \qquad k = 0,\ 1,\ 2, \tag{8.2-4}$$

for a time step, s. The numerical values, T^k, are an approximation to the analytical solution at the discrete times:

$$T^k \qquad T(t^k) \tag{8.2-5}$$

If the backward and forward algorithms of (5.4-13) and (5.4-23), as examples of two-point stencils, are applied to (8.2-1), the results are

$$\begin{aligned} &\tfrac{C}{s}(T^k - T^{k-1}) + KT^k = Q^k && \textbf{Backward Difference} \\ &\tfrac{C}{s}(T^{k+1} - T^k) + KT^k = Q^k && \textbf{Forward Difference} \end{aligned} \tag{8.2-6}$$

respectively, in which $Q^k = Q(t^k)$. Either equation is solved consecutively to obtain T^k for k = 1, 2, ... with $T^0 = T_0$. For example, if the backward difference algorithm is used, then the equations for k = 1 and k = 2 are

$$(\tfrac{C}{s} + K)T^1 = \tfrac{C}{s}T^0 + Q^1\,, \qquad (\tfrac{C}{s} + K)T^2 = \tfrac{C}{s}T^1 + Q^2 \tag{8.2-7}$$

The procedure is continued to obtain approximate solutions up to the maximum discrete time required.

The backward and forward difference algorithms are first order accurate. The central difference algorithm of (5.4-27) is an example of a second-order accurate formula:

$$\tfrac{C}{2s}(T^{k+1} - T^{k-1}) + KT^k = Q^k \qquad \textbf{Central Difference} \tag{8.2-8}$$

The problem with (8.2-8) is that the equation at each time step involves the primary variable at three steps rather than two steps as given in (8.2-6). Only one initial value is given so a special algorithm must be used to get started. Normally, the initial condition and the differential equation, itself, is used to obtain the special algorithm. For example, if the differential equation (8.2-1) is applied at t = 0, the result is that the first derivative at t = 0 must satisfy

$$C\dot{T}_0 = -KT_0 + Q(0) \tag{8.2-9}$$

This is a mixed initial condition in that a combination of the primary variable and its derivative is prescribed. A second-order accurate algorithm, such as that given in (5.5-19), involving points at k = 1, 2 and 3 must be used for time t^0:

$$\frac{C}{2s}(-3T^0 + 4T^1 - T^2) + KT^0 = Q^0 \tag{8.2-10}$$

Then (8.2-8) is used for k = 1 and k = 2 to obtain three equations to solve for T^0, T^1 and T^2. Once this special start-up procedure is invoked, (8.2-8) can be used consecutively for k = 3,4,…. The need for a special start-up algorithm for a procedure that is second-order accurate is circumvented by the trapezoidal rule described in the next section.

Even higher-order algorithms can be utilized, but then additional conditions are required at the boundary to provide enough equations to initiate the algorithm. These additional equations are obtained by differentiating the terms in the differential equation. For example, a condition on the second derivative is obtained from (8.2-1) to be the following:

$$C\ddot{T}_0 = -K\dot{T}_0 + \dot{Q}(0) \tag{8.2-11}$$

Then, a special algorithm must be constructed for approximating (8.2-11).

8.3 THE GENERAL TRAPEZOIDAL RULE

We study the general trapezoidal rule because the algorithm has the following desirable features:

(i) a special starting algorithm is not needed,
(ii) the algorithm contains a free parameter; particular choices yield the finite difference algorithms given above, and
(iii) a second-order accurate algorithm is obtained involving just two levels of time.

A **time integrator** is an equation involving the approximation T^k and its derivative, $\dot{T}^k$, at two or more levels of time. A time integrator by itself is not sufficient to obtain approximate solutions. In addition, a discretized form of the differential equation, itself, is needed. When the two equations are combined, the result is defined here to be a **system integrator**, which provides a sufficient number of equations to solve for the discrete approximations to the primary variable at sequential levels of time. The system integrator must be shown to be stable and convergent.

There are an infinite number of time integrators. The **general trapezoidal rule** is the following particular choice of a time integrator in which approximations to the primary variable, T, and its derivative, $\dot{T}$, are related at two consecutive levels of time:

$$T^{k+1} = T^k + s[\alpha\dot{T}^{k+1} + (1-\alpha)\dot{T}^k] , \qquad 0 \le \alpha \le 1 \tag{8.3-1}$$

The free parameter, α, provides a means for studying a class of integrator rather than focusing initially on a particular value of α. The time integrator, itself, is explicit if $\alpha = 0$ for then the updated value for T depends only on previous values of T and its derivative; otherwise it is implicit. A discussion of whether or not the system integrator is an **explicit integrator** or an **implicit integrator** is more properly described in the multi-degree-of-freedom context of the next chapter.

The increment in T is (T^{k+1} - T^k) and an interpretation of the general trapezoidal rule is given in the sketch shown in Fig. 8.3-1. If $\alpha = 0$, (8.3-1) becomes

$$\dot{T}^k = \frac{T^{k+1} - T^k}{s} \tag{8.3-2}$$

which is the **forward difference** operator. If $\alpha = 1$, (8.3-1) becomes

$$\dot{T}^{k+1} = \frac{T^{k+1} - T^k}{s} \qquad \text{or} \qquad \dot{T}^k = \frac{T^k - T^{k-1}}{s} \tag{8.3-3}$$

which is a **backward difference**.

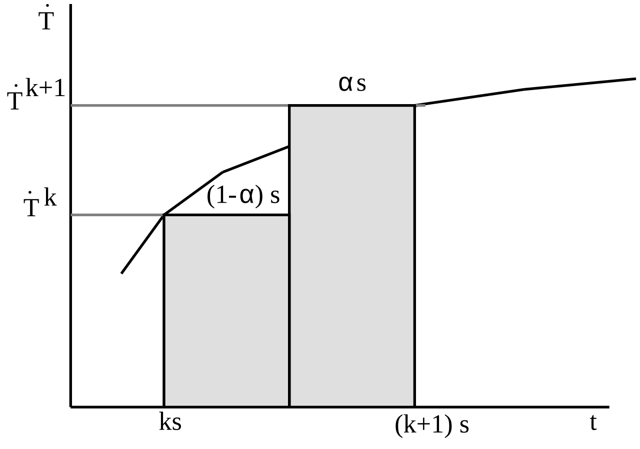

Fig. 8.3-1. Trapezoidal rule for time integration.

There are also an infinite number of ways to discretize the governing differential equation. The most straightforward way, and the way that many implicitly follow without inferring that there are alternative choices, is the following discretization of (8.2-1):

$$C\dot{T}^k + KT^k = Q^k \qquad (8.3\text{-}4)$$

in which $Q^k = Q(t^k)$.

The system integrator is obtained by combining (8.3-1) and (8.3-4). First, update the terms in (8.3-4) one time step to obtain the following:

$$C\dot{T}^{k+1} + KT^{k+1} = Q^{k+1} \qquad (8.3\text{-}5)$$

Multiply (8.3-1) by C, (8.3-4) by s(1-α), and (8.3-5) by sα. It is left as an exercise to show that when the results are combined the following algorithm, called the system integrator, is obtained:

$$AT^{k+1} = BT^k + \mathcal{F}^k, \qquad k = 0, 1, 2, ... \qquad (8.3\text{-}6)$$

in which

$$\begin{gathered} A = C + \alpha sK, \quad B = C - (1-\alpha)sK \\ \mathcal{F}^k = \alpha sQ^{k+1} + (1-\alpha)sQ^k \end{gathered} \qquad (8.3\text{-}7)$$

Since Q(t) is a given function, $\mathcal{F}^k$ is always computable for any k. With T^0 known as an initial condition, (8.3-6) provides T^1 with k = 0. The approximate solution for subsequent discrete times are obtained by merely incrementing k and applying (8.3-6) sequentially.

The result of completing Exercise 2 is a series of plots showing the effects of the parameters α and s on the numerical solutions obtained from the algorithm of (8.3-6). The following subsections provide a detailed analysis of the system integrator so if the reader prefers to see, instead, the combination of the time integrator with finite elements, this is the appropriate point to go to the next chapter.

8.4 STABILITY ANALYSIS OF THE GENERAL TRAPEZOIDAL RULE

The use of the trapezoidal rule (8.3-1) and the approximation (8.3-4) to the governing equation yields the discretized system of (8.3-6). For the elementary case of zero forcing function, the exact solution to the system integrator is given by (8.3-9), a form sufficiently simple that specific conclusions can be made concerning stability, consistency and convergence. First, we consider stability. If the approximate solution remains bounded for all discrete time, t^k, the numerical algorithm is said to be **stable**.

To analyze the system integrator, consider the homogeneous form of the differential equation. Define a dimensionless time step, $\bar{s}$, to be

$$\bar{s} = s\Psi , \qquad \Psi = \frac{K}{C} \tag{8.4-1}$$

Then the system time integration algorithm of (8.3-6) becomes

$$T^{k+1} = \lambda T^k \qquad \text{with} \qquad \lambda = \frac{B}{A} = \frac{[1- \bar{s}(1- \alpha)]}{[1+ \bar{s}\alpha]} \tag{8.4-2}$$

and $T^0 = T_0$. Note that the exact solution $T^e = T_0 e^{-\Psi t}$ yields $T^{e(k+1)} = e^{-\bar{s}} T^{e(k)}$. The Taylor expansion

$$e^{-\bar{s}} = 1 - \bar{s} + \tfrac{1}{2}\bar{s}^2 - \tfrac{1}{6}\bar{s}^3 + \tag{8.4-3}$$

provides a method for easily comparing exact and approximate solutions for various choices of α. If approximations to the derivative are required, (8.3-1) is used with the following result:

$$\dot{T}^k = -\Psi T^k \tag{8.4-4}$$

Particular values for α yield special cases that are referred to by name:

Explicit, Forward Difference or Forward Euler: $\alpha = 0$

$$\lambda = 1- \bar{s} \tag{8.4-5}$$

Trapezoidal Rule, Mid-point Rule or Crank-Nicolson: $\alpha = \frac{1}{2}$

$$\lambda = \frac{[1 - \tfrac{1}{2}\bar{s}]}{[1 + \tfrac{1}{2}\bar{s}]} \approx 1 - \bar{s} + \tfrac{1}{2}\bar{s}^2 - \tfrac{1}{4}\bar{s}^3 + \tag{8.4-6}$$

Note the agreement with the exact solution through second-order in $\bar{s}$.

Galerkin: $\alpha = \frac{2}{3}$

$$\lambda = \frac{[1 - \tfrac{1}{3}\bar{s}]}{[1 + \tfrac{2}{3}\bar{s}]} \tag{8.4-7}$$

Backward Difference or Backward Euler: $\alpha = 1$

$$\lambda = \frac{1}{1+\bar{s}} \approx 1 - \bar{s} + \bar{s}^2 - \qquad (8.4\text{-}8)$$

The system integrator algorithm, given in (8.4-2), is so simple that the following exact solution to the discretized equation is obtained:

$$T^k = \lambda T^{k-1} = \lambda^2 T^{k-2} = \quad = \lambda^k T^0 \qquad (8.4\text{-}9)$$

Even though (8.4-9) is the exact solution to the approximating equation, in general the numerical solution will not equal the exact solution at the discrete time, t^k. Stability implies that the solution remains bounded for all time. Therefore, a consideration of the right most term in (8.4-9) implies that the algorithm is stable provided $|\lambda| \leq 1$.

To visualize the stability properties of the general trapezoidal rule, consider the form for λ in (8.4-2) and plot λ as a function of $\bar{s}$ for various values of α as shown in Fig. 8.4-1. The unstable regions consist of $\lambda > 1$ and $\lambda < -1$. We choose the particular values of $\alpha = 0$, 1/2 and 1 to illustrate the use of the plot. In particular, we note the following limiting cases:

$$\begin{aligned} \lambda\big|_{\alpha=1} &= \frac{1}{1+\bar{s}} & \lim_{\bar{s}\to\infty} \lambda\big|_{\alpha=1} &= 0 \\ \lambda\big|_{\alpha=0} &= 1-\bar{s} & \lim_{\bar{s}\to\infty} \lambda\big|_{\alpha=0} &= -\infty \qquad (8.4\text{-}10) \\ \lambda\big|_{\alpha=1/2} &= \frac{[1-\frac{1}{2}\bar{s}]}{[1+\frac{1}{2}\bar{s}]} & \lim_{\bar{s}\to\infty} \lambda\big|_{\alpha=1/2} &= -1 \end{aligned}$$

In general, by considering the general features illustrated in Fig. 8.4-1, and with the use of (8.4-2), we see that

$$\lim_{\bar{s}\to\infty} \lambda = \frac{(1-\alpha)}{\alpha} < -1 \quad \text{for} \quad \alpha < \tfrac{1}{2} \qquad (8.4\text{-}11)$$

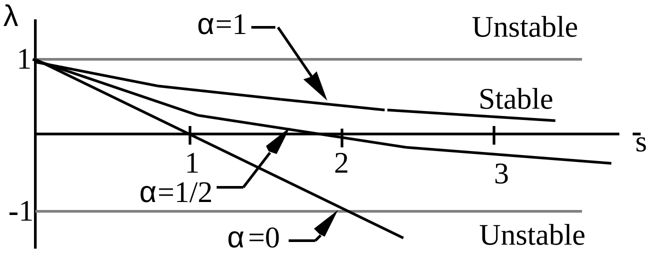

Fig. 8.4-1. Stability diagram for trapezoidal rule.

The conclusion is that the system trapezoidal rule is stable for all values of the dimensionless time step $\bar{s}$ if $\alpha \geq \frac{1}{2}$. Such an integrator is said to be **unconditionally stable**. For values of $\alpha < \frac{1}{2}$, the integrator is stable provided the time step is sufficiently small. Then the integrator is said to be **conditionally stable**.

Based on the sketch, the maximum time step that yields a stable algorithm for $0 < \alpha < \frac{1}{2}$ can be obtained by selecting a value for the time step such that λ = -1 in (8.4-2). The result is

$$-1 = \frac{1 - \bar{s}_c(1-\alpha)}{1 + \bar{s}_c\alpha} \qquad (8.4\text{-}12)$$

where $\bar{s}_c$ denotes the critical time step. The solution for $\bar{s}_c$ is

$$\bar{s}_c = \frac{2}{(1-2\alpha)} \qquad (8.4\text{-}13)$$

For example, if α = 0 (forward difference) the critical time step is

$$\bar{s}_c \big|_{\alpha=0} = 2 \qquad (8.4\text{-}14)$$

which is consistent with Fig. 8.4-1 for the case of $\alpha = 0$.

For **conditionally stable integrators** the dimensionless time step must be chosen such that $\bar{s} \leq \bar{s}_c$. For **unconditionally stable integrators**, any time step can be used. However, for large time steps, it is to be expected that the approximate solution will not be very accurate.

8.5 ERROR AND CONVERGENCE

In Chapt. 5, the method for drawing conclusions concerning convergence involved first obtaining the local truncation error, next using the local truncation error to obtain an expression for the error of the numerical solution, and finally showing convergence with respect to refinement of the step size. The same order of analysis is followed here with application to the system trapezoidal integrator.

Again, for ease of reference, we repeat the key equations. Recall that the governing differential equation is

$$C\dot{T} + KT = Q(t) \qquad (8.5\text{-}1)$$

which yields the following equations at the discrete times, t^k and t^{k+1}:

$$C\dot{T}(t^k) + KT(t^k) = Q(t^k)\ , \qquad C\dot{T}(t^{k+1}) + KT(t^{k+1}) = Q(t^{k+1}) \tag{8.5-2}$$

The approximate solution satisfies the following system integrator:

$$AT^{k+1} = BT^k + F^k\ , \qquad k = 0,\ 1,\ 2,\ .... \tag{8.5-3}$$

in which

$$\begin{gathered} A = C + \alpha sK\ , \quad B = C - (1-\alpha)sK \\ F^k = \alpha sQ^{k+1} + (1-\alpha)sQ^k \end{gathered} \tag{8.5-4}$$

If the exact solution is substituted in the system integrator, the equation will not be satisfied. The residual for each discrete time, t^k, is proportional to the **local truncation error**, τ^k. We note that the forcing term, F^k, contains the time step, s. This term must be factored out to obtain a form for the local truncation error which will yield the correct rate of convergence. Therefore, the local truncation error in this context is defined to be

$$\tau^k = \tfrac{1}{s}\left[AT(t^{k+1}) - BT(t^k) - \alpha sQ(t^{k+1}) - (1-\alpha)sQ(t^k)\right] \tag{8.5-5}$$

With the use of (8.5-2), the force terms are eliminated and the result is

$$\begin{aligned} \tau^k = \tfrac{1}{s}\big[& AT(t^{k+1}) - BT(t^k) - \alpha sC\dot{T}(t^{k+1}) \\ & - \alpha sKT(t^{k+1}) - (1-\alpha)sC\dot{T}(t^k) - (1-\alpha)sKT(t^k)\big] \end{aligned} \tag{8.5-6}$$

Next, we substitute in the expressions for A and B and collect terms to obtain

$$\tau^k = \tfrac{C}{s}\left[T(t^{k+1}) - T(t^k) - \alpha s\dot{T}(t^{k+1}) - (1-\alpha)s\dot{T}(t^k)\right] \tag{8.5-7}$$

We use Taylor series to express all terms at the common discrete time of t^k:

$$\begin{aligned} & T(t^k) = T(t^k) \\ & T(t^{k+1}) = T(t^k) + s\dot{T}(t^k) + \frac{s^2}{2}\ddot{T}(t^k) + \frac{s^3}{6}\dddot{T}(t^k) + \\ & \dot{T}(t^k) = \dot{T}(t^k) \\ & \dot{T}(t^{k+1}) = \dot{T}(t^k) + s\ddot{T}(t^k) + \frac{s^2}{2}\dddot{T}(t^k) + \end{aligned} \tag{8.5-8}$$

After some algebraic manipulations when (8.5-8) is substituted in (8.5-7), the local truncation error becomes

$$\tau^k = Cs\left[\left(\frac{1}{2}-\alpha\right)\ddot{T}(t^k) + \left(\frac{1}{6}-\frac{\alpha}{2}\right)s\dddot{T}(t^k) + \qquad \right] \tag{8.5-9}$$

If $\tau^k \to 0$ as $\bar{s} \to 0$, the system integrator is said to display **consistency**. The analysis above indicates that the combination of the general trapezoidal rule as a time integrator and the particular choice of (8.5-2) as a discretized form of the differential equation provides a consistent system integrator for any choice of α.

Next, we discuss error and convergence and show why an expression for the local truncation error in necessary. Define the discrete **error**, e^k, to be the difference between the exact and approximate solution at the discrete time, t^k, as follows:

$$e^k \equiv T(t^k)- T^k \tag{8.5-10}$$

Suppose the system integrator is stable and the local truncation error satisfies the inequality

$$|\tau^k| \le c_\tau \bar{s}^r \tag{8.5-11}$$

If r > 0 in (8.5-11), the system integrator is consistent. Now invoke **Lax's Equivalence Theorem** which states that the error satisfies the inequality

$$|e^k| \le c_e \bar{s}^r \tag{8.5-12}$$

The implication of (8.5-12) is that as $\bar{s}$ approaches zero, so does the error, or the approximate solution converges to the exact solution. In other words, stability and consistency imply **convergence**. Based on (8.5-12), the exponent r is called the **rate of convergence**. The coefficients, c_τ and c_e, depend on the data for the problem but do not depend on the time step, s.

Now return to (8.5-9). The rates of convergence and the coefficients for two cases follow:

$$\begin{aligned} &\alpha \neq \frac{1}{2}, \quad r=1, \quad c_\tau = C\left(\frac{1}{2} - \alpha\right)\ddot{T}(t^k) \\ &\alpha = \frac{1}{2}, \quad r=2, \quad c_\tau = C\left(\frac{1}{6}-\frac{\alpha}{2}\right)\dddot{T}(t^k) \end{aligned} \tag{8.5-13}$$

We have derived the result that the system integrator is always consistent and first-order accurate. For the particular choice of α = 1/2 the system integrator is second-order accurate. The data for the problem include C and affect the exact solution $T(t^k)$ and, therefore, the data determine the coefficients c_τ and c_e.

For the sake of completeness, **Lax's Equivalence Theorem** is proven in the next Section.

8.6 LAX'S EQUIVALENCE THEOREM

The key results that we want from an integration algorithm is that the numerical solution remains bounded in time (stability) and that the numerical solution improves if a smaller time step is used (convergence). In most cases, it is fairly easy to show stability, but a proof of convergence normally involves the use of consistency as indicated in the following theorem.

Theorem: If a system integrator is stable and consistent with local truncation error $c_\tau s^r$ then the approximate solution converges to the exact solution as s approaches zero and the rate of convergence is r.

Proof: The system integrator can be cast in the form:

$$0 = AT^{k+1} - BT^{k} - F^{k} \qquad \text{(i)}$$

and the equation for the local truncation error, τ^k, is obtained from the following equation:

$$s\tau^{k} = AT(t^{k+1}) - BT(t^{k}) - F^{k} \qquad \text{(ii)}$$

Define the error to be

$$e^{k} = T(t^{k}) - T^{k} \qquad \text{(iii)}$$

Subtract corresponding terms in (i) and (ii) and use (iii) to obtain

$$s\tau^{k} = Ae^{k+1} - Be^{k} \qquad \text{(iv)}$$

Shift down one time step, recall from (8.4-2) that λ = B/A, and rearrange terms to obtain

$$e^{k} = \lambda\, e^{k-1} + \frac{s}{A}\tau^{k-1} \qquad \text{(v)}$$

Shift back one time step again:

$$e^{k-1} = \lambda\, e^{k-2} + \frac{s}{A}\tau^{k-2} \qquad \text{(vi)}$$

Substitute (vi) in (v) to obtain

$$e^{k} = \lambda^{2} e^{k-2} + \frac{s}{A}(\tau^{k-1} + \lambda\tau^{k-2}) \qquad \text{(vii)}$$

Repeat the procedure k - 2 times:

$$e^{k} = \lambda^{k} e^{0} + \frac{s}{A} \sum_{i=0}^{k-1} \lambda^{i} \tau^{k-i} \qquad \text{(viii)}$$

But $e^{0} = 0$ because the approximate solution begins with the exact inital condition. Recall that $A = C + \alpha sK$, that $\Lambda = K/C$ and that $\bar{s} = s\Lambda$. Then (viii) becomes

$$e^{k} = \frac{s}{C(1 + \alpha \bar{s})} \sum_{i=0}^{k-1} \lambda^{i} \tau^{k-i} \qquad \text{(ix)}$$

We perform the following steps involving inequalities and absolute values:

$$\begin{aligned}
|e^{k}| &\le \left| \frac{s}{C} \sum_{i=0}^{k-1} \lambda^{i} \tau^{k-i-1} \right| && \text{since } \bar{s}\alpha \ge 0 \\
&\le \frac{s}{C} \sum_{i=0}^{k-1} |\lambda^{i}|\,|\tau^{k-i-1}| && \left(\frac{s}{C} > 0\right) \\
&\le \frac{s}{C} \sum_{i=0}^{k-1} |\tau^{k-i-1}| && \text{with the use of stability} \\
&\le \frac{sk}{C} \max_{i=0}^{k-1} |\tau^{i}| && \text{for fixed domain of time, t} \quad [0, T] \\
&\le \frac{t^{k}}{C} c_{\tau} s^{r} && \text{with the use of consistency and } sk = t^{k} \\
&\to 0 \text{ as } s \to 0 && \text{for fixed } t^{k} \text{ provided } r > 0.
\end{aligned}$$

Therefore the error goes to zero as the time step goes to zero. This is the desired result that stability and consistency imply convergence, and the rate of convergence is r. **EOP**

8.7 DERIVATION BASED ON LOCAL TRUNCATION ERROR

An alternative approach, outlined in Chapter 5, for developing a system time integrator is to propose a stencil for a numerical algorithm with free parameters. The free parameters are then chosen to ensure stability and convergence, and, if desired, a specified order of convergence. Here, we outline the approach for choosing the parameters in the stencil. The final result is that the two approaches are just different ways of developing equivalent numerical algorithms.

Suppose a two-point stencil (two time levels) involving a linear combination of forcing terms is proposed for the differential equation of (8.2-1) as follows:

$$\beta T^{k} + \gamma T^{k+1} = (1-\alpha)Q^{k} + \alpha Q^{k+1}\,, \qquad 0 \le \alpha \le 1 \tag{8.7-1}$$

The **local truncation error**, τ, is the residual obtained when the exact solution is substituted into the approximating equation. If the parameters β and γ are chosen to ensure that the local truncation error approaches zero as the time step, s, goes to zero, then the algorithm is said to satisfy **consistency**. The local truncation error based on (8.7-1) is:

$$\tau^{k} = \beta T(t^{k}) + \gamma T(t^{k+1}) - (1-\alpha)Q(t^{k}) - \alpha Q(t^{k+1}) \tag{8.7-2}$$

Note that the forcing terms do not have the time step, s, as a coefficient. Utilize (8.5-2) and (8.5-8) to obtain

$$\begin{aligned} \tau^{k} = {} & T(t^{k})[\beta + \gamma - K] \\ & + \dot{T}(t^{k})[s\gamma - C - \alpha K s] \\ & + \ddot{T}(t^{k})[\frac{s^2}{2}\gamma - \alpha C s - \alpha K \frac{s^2}{2}] \\ & + \dddot{T}(t^{k})[\frac{s^3}{6}\gamma - \alpha C \frac{s^2}{2} - \alpha K \frac{s^3}{6}] + \end{aligned} \tag{8.7-3}$$

For consistency τ^{K} must go to zero as the time step, s, goes to zero so the first two square brackets must be zero. The result of solving for β and γ is

$$\beta = \frac{(1-\alpha)Ks - C}{s} \qquad \text{and} \qquad \gamma = \frac{\alpha Ks + C}{s} \tag{8.7-4}$$

Suppose (8.7-4) is substituted in (8.7-1). After multiplying through by the time step, s, the resulting algorithm is

$$AT^{k+1} = BT^{k} + F^{k} \tag{8.7-5}$$

in which

$$\begin{aligned} A = C + \alpha sK\,, \quad & B = C - (1-\alpha)sK \\ F^{k} = \alpha sQ^{k+1} & + (1-\alpha)sQ^{k} \end{aligned} \tag{8.7-6}$$

which is identical to the system time integrator called the general trapezoidal rule of (8.3-6) and (8.3-7).

In summary, the preceding development suggests that at least three general approaches are available for deriving a suitable algorithm for time integration:

(i) The first approach is to use algorithms from the finite difference literature.

(ii) The second approach is to introduce a time integrator of which a fairly general form is the following:

$$T^{k+1} = T^k + s[\gamma_0 \dot{T}^{k+1} + \gamma_1 \dot{T}^k + \gamma_2 \dot{T}^{k-1} + \quad] \tag{8.7-7}$$

The time integrator is to be combined with a discretized form of the governing equation which can be a relation of the following type:

$$\begin{aligned} C[\alpha_1 \dot{T}^{k+1} + \alpha_2 \dot{T}^k + \alpha_3 \dot{T}^{k-1} + \quad] + K[\beta_1 T^{k+1} + \beta_2 T^k + \beta_3 T^{k-1} + \quad] \\ -[a_1 Q^{k+1} + a_2 Q^k + a_3 Q^{k-1} + ...] = 0 \end{aligned} \tag{8.7-8}$$

(iii)The third approach is to take linear combinations of T and Q at various time steps:

$$aT^{k+1} + bT^k + cT^{k-1} + \qquad [\alpha Q^{k+1} + \beta Q^k + \gamma Q^{k-1} + \quad] = 0 \tag{8.7-9}$$

The free parameters in (8.7-7), (8.7-8) or (8.7-9) are chosen so that consistency and the desired order of accuracy is achieved.

If the same order of accuracy is specified, the three approaches will yield equivalent algorithms which may appear to be entirely different. A reason for the different approaches is often a matter of history and the specific scientific or engineering discipline that is involved.

8.8 DIRECT PLOTS OF ERROR

Suppose that more detail about numerical error is required than that provided by the rate of convergence. For example, considerable insight into the behavior of an algorithm is often provided by a direct plot of error for a problem with a known analytical solution. Suppose the error is normalized by the maximum absolute value of the exact solution, T_m, to obtain a dimensionless plot of error, i.e., define a measure of error to be

$$E = \left| \frac{T^k - T(t^k)}{T_m} \right| \tag{8.8-1}$$

This expression could be plotted as a function of discrete time for any set of problem data. As an example, consider the homogeneous problem with initial value, T_0. The exact solutions for arbitrary time and for discrete times are given as follows:

$$T(t) = T_0 e^{-\Lambda t} \quad \text{and} \quad T(t^k) = T_0 e^{-\Lambda ks} = T_0\, e^{-k\bar{s}} \qquad (8.8\text{-}2)$$

The approximate solution is

$$T^k = T^0 \lambda^k \;, \qquad T^0 = T_0 \qquad (8.8\text{-}3)$$

The maximum value of T is the intial value, T_0, so an explicit form for the measure of error is available for this case as follows:

$$E = |\lambda^k - e^{-k\bar{s}}| \quad \text{with} \quad \lambda = \frac{[1- \bar{s}\,(1-\alpha)]}{[1+ \bar{s}\alpha]} \qquad (8.8\text{-}4)$$

It follows that

$$\lim_{\bar{s} \to 0} E = 0 \quad => \quad T^k \to T(t^k) \text{ as } \bar{s} \to 0 \qquad (8.8\text{-}5)$$

which is expected because of the proof of convergence.

Plots of E as a function of $\bar{s}$ for various values of α can often provide more detailed information about the integrator than that provided by an error analysis. Sometimes good qualitative solutions are obtained even with a low-order algorithm.

8.9 CONCLUDING REMARKS

For an elementary model problem, several approaches have been introduced with regard to the development of time integration algorithms. First, the finite difference method can be used directly to provide an algorithm. Second, a time integrator can be combined with a discretized form of the differential equation to obtain a system time integrator. Third, a linear combination of the dependent variable and the forcing term at consecutive time levels can be postulated with arbitrary coefficients. Consistency is invoked to obtain the equations used for determining the coefficients.

It is critical to determine if the system integrator is stable or not, and if stable, whether or not it is conditionally or unconditionally stable. If the system is conditionally stable, then the time step must be less than a critical value.

The local truncation error is related to the residual obtained if the exact solution is substituted in the approximate governing equation. If the local truncation error approaches zero as the time step approaches zero the algorithm is said to be consistent. Consistency and stability ensures that the approximate solution converges to the exact solution as the time step approaches zero.

A demonstration of convergence and the determination of the rate of convergence is still not sufficient for most engineers. Approximate solutions are often obtained for much longer periods of time than would be considered appropriate based on the usual mathematical analysis of algorithms. By obtaining numerical plots of error versus time, it is often possible to show that useful information can be extracted for extended solution times.

The most important aspect of this chapter involves the approaches for developing and analyzing integrators. This is particularly crucial because the literature provides an overwhelming number of integrators that appear to be different. If these integrators are analyzed in a systematic fashion, it can be shown that the differences are rather superficial and result from the historical development of a particular discipline.

8.10 EXERCISES

1. Show that the combination of (8.3-1), (8.3-2) and (8.3-3) yields (8.3-4).

2. Consider the single-degree-of-freedom problem

$$\dot{T} + \Lambda T = 0 \,, \qquad T(0) = 1 \qquad \text{and} \qquad t \quad [0,8]$$

with $\Lambda = 1$. Overlay plots of the exact solution and approximate solutions as functions of time obtained from the general trapezoidal rule for the following values of α and time step, s:

(a) $\alpha = 0$ s = 4, 2, 1, 0.5, 0.25, 0.125
(b) $\alpha = 1/2$ s = 4, 2, 1, 0.5, 0.25, 0.125
(c) $\alpha = 1$ s = 4, 2, 1, 0.5, 0.25, 0.125

For each case plot the error as a function of time

$$E = \left| T^k - T(t^k) \right| / T_{ref}$$

where T_{ref} is a reference temperature and determine whether or not such a measure of error is useful for drawing conclusions concerning the accuracy of the approximate solution.

3. Perform an error analysis of the system integrator obtained when the general trapezoidal integrator is combined with the following discretized form of the differential equation

$$C\dot{T}^k + K[\eta T^k + (\eta - 1)T^{k-1}] = Q^k \,, \qquad 0 \le \eta \le 1$$

Now there are two free parameters, α and η. Is it possible to obtain a rate of convergence higher than two?

The following is the original version:

8.5 ERROR AND CONVERGENCE

In Chapt. 5, the method for drawing conclusions concerning convergence involved first obtaining the local truncation error, next using the local truncation error to obtain an expression for the error of the numerical solution, and finally showing convergence with respect to refinement of the step size. The same order of analysis is followed here with application to the system trapezoidal integrator.

Again, for ease of reference, we repeat the key equations. Recall that the governing homogeneous differential equation is

$$\dot{T} + \Lambda T = 0 \qquad (8.5\text{-}1)$$

The approximate solution satisfies the following discretized form of the equation:

$$\dot{T}^k + \Lambda T^k = 0 \qquad (8.5\text{-}2)$$

With the use of the general trapezoidal rule the system time integration algorithm is

$$T^{k+1} = \lambda T^k \qquad \text{with} \qquad \lambda = \frac{[1-\bar{s}(1-\alpha)]}{[1+\bar{s}\alpha]} \qquad (8.5\text{-}3)$$

and $T^k = T^0\lambda^k$ is the solution to the numerical algorithm. We rewrite the system integrator as follows:

$$[1+\bar{s}\alpha]\, T^{k+1} - [1-\bar{s}(1-\alpha)]\, T^k = 0 \qquad (8.5\text{-}4)$$

If the exact solution is substituted in the system integrator, in general, the equation will not be satisfied. When this approach of combining an integrator with a discretized form of the governing equation is used, the **local truncation error**, τ^k, at the discrete time, t^k, is defined to be the residual divided by $\bar{s}^q$ in which q is the order of the differential equation. For the first order differential equation considered here, q = 1 and the local truncation error is

$$\tau^k = \frac{1}{\bar{s}}\{[1+\bar{s}\alpha]\, T(t^{k+1}) - [1-\bar{s}(1-\alpha)]\, T(t^k)\} \qquad (8.5\text{-}5)$$

We use Taylor series to obtain

$$T(t^{k+1}) = T(t^k) + s\dot{T}(t^k) + \frac{s^2}{2!}\ddot{T}(t^k) + \frac{s^3}{3!}\dddot{T}(t^n) + \frac{s^4}{4!}T^{iv}(t^k) + \qquad (8.5\text{-}6)$$

The solution to the differential equation is T(t). If T is sufficiently smooth, the terms in the differential equation can be differentiated sequentially with the following implications:

$$\begin{aligned} &\dot{T} + \Lambda T = 0 &&=> &&\dot{T} = -\Lambda T \\ &\ddot{T} + \Lambda \dot{T} = 0 &&=> &&\ddot{T} = -\Lambda \dot{T} = \Lambda^2 T \qquad (8.5\text{-}7) \\ &\dddot{T} + \Lambda \ddot{T} = 0 &&=> &&\dddot{T} = -\Lambda \ddot{T} = -\Lambda^3 T \quad \text{etc.} \end{aligned}$$

The substitution of (8.5-7) in (8.5-6) yields the following alternative representation for the Taylor series:

$$T(t^{k+1}) = T(t^k)[\ 1 - \bar{s} + \frac{\bar{s}^2}{2!} - \frac{\bar{s}^3}{3!} + \frac{\bar{s}^4}{4!} - \quad] \qquad (8.5\text{-}8)$$

After some algebraic manipulations when (8.5-8) is substituted in (8.5-5), the local truncation error becomes

$$\tau^k = \Lambda T(t^k)[(\frac{1}{2} - \alpha)\bar{s} - (\frac{1}{6} - \alpha^2)\bar{s}^2 + \quad] \qquad (8.5\text{-}9)$$

If $\tau^k \to 0$ as $\bar{s} \to 0$, the system integrator is said to display **consistency**. The analysis above indicates that the combination of the general trapezoidal rule as a time integrator and the particular choice of (8.5-2) as a discretized form of the differential equation provides a consistent system integrator for any choice of α.

Next, we discuss error and convergence and show why an expression for the local truncation error in necessary. Define the discrete **error**, e^k, to be the difference between the exact and approximate solution at the discrete time, t^k, as follows:

$$e^k \equiv T(t^k) - T^k \qquad (8.5\text{-}10)$$

Suppose the system integrator is stable and the local truncation error satisfies the inequality

$$|\tau^k| \le c_\tau \bar{s}^r \qquad (8.5\text{-}11)$$

If r > 0 in (8.5-11), the system integrator is consistent. Now invoke **Lax's Equivalence Theorem** which states that the error satisfies the inequality

$$|e^k| \le c_e \bar{s}^r \qquad (8.5\text{-}12)$$

The implication of (8.5-12) is that as $\bar{s}$ approaches zero, so does the error, or the approximate solution converges to the exact solution. In other words, stability and consistency imply **convergence**. Based on (8.5-12), the exponent r is called the **rate of convergence**. The coefficients, c_τ and c_e, depend on the data for the problem but do not depend on the time step, s.

Now return to (8.5-9). The rates of convergence and the coefficients for two cases follow:

$$\begin{aligned} &\alpha \neq \frac{1}{2}\ , \qquad r = 1\ , \qquad c_\tau = \Lambda T(t^k)\left(\frac{1}{2} - \alpha\right) \\ &\alpha = \frac{1}{2}\ , \qquad r = 2\ , \qquad c_\tau = \frac{1}{12}\Lambda T(t^k) \end{aligned} \tag{8.5-13}$$

We have derived the result that the system integrator is always consistent and first-order accurate. For the particular choice of $\alpha = 1/2$ the system integrator is second-order accurate. The data for the problem affect the exact solution $T(t^k)$ and Λ and, therefore, the data determine the coefficients c_τ and c_e.

For the sake of completeness, **Lax's Equivalence Theorem** is proven in the next Section.