Chapter 8

PARABOLIC DIFFERENTIAL EQUATIONS

Parabolic partial differential equations involve time (or a time-like quantity) as an independent variable. Therefore, the resulting initial/boundary-value problems include two types of independent variables, i.e., spatial variables (e.g., x_i, $i = 1, 2, 3$) and a temporal variable (t). In the context of the finite element method, there are two general approaches in dealing with the two types of variables. These are:

(a) Discretize the spatial variables independently from the temporal variable.

In this approach, the spatial discretization typically occurs first and yields a system of ordinary differential equations in time. These equations are subsequently integrated in time by means of some standard numerical integration method. This approach is referred to as *semi-discretization* and is used widely in engineering practice due to its conceptual simplicity and computational efficiency.

(b) Discretize spatial and temporal variables together.

Here, all independent variables are treated simultaneously, although the discretization is generally different for spatial and temporal variables. This approach yields *space-time finite elements*. Such elements are typically used for special problems, as they tend to be more complicated and expensive than those resulting from semi-discretization.

8.1 Standard semi-discretization methods

Consider the time-dependent version of the Laplace-Poisson equation in two dimensions. The initial/boundary value problem takes the form

$$
\begin{aligned}
\frac{\partial}{\partial x_1}(k\,\frac{\partial u}{\partial x_1}) \;+\; \frac{\partial}{\partial x_2}(k\,\frac{\partial u}{\partial x_2}) - f \;&=\; \rho c \frac{\partial u}{\partial t} \quad \text{in } \Omega \times I \;, \\
-k\,\frac{\partial u}{\partial n} \;&=\; \bar{q} \quad \text{on } \Gamma_q \times I \;, \\
u \;&=\; \bar{u} \quad \text{on } \Gamma_u \times I \;, \\
u(x_1, x_2, 0) \;&=\; u_0(x_1, x_2) \quad \text{in } \Omega \;,
\end{aligned}
\tag{8.1}
$$

where $u = u(x_1, x_2, t)$ is the (yet unknown) solution and $I = (0, T]$, with T being a given time. Continuous functions $k = k(x_1, x_2)$, $\rho = \rho(x_1, x_2)$, $c = c(x_1, x_2)$ and $u_0 = u_0(x_1, x_2)$ are defined in Ω and a continuous function $f = f(x_1, x_2, t)$ is defined in $\Omega \times I$. Further, continuous functions $\bar{q} = \bar{q}(x_1, x_2, t)$ and $\bar{u} = \bar{u}(x_1, x_2, t)$ are defined on $\Gamma_u \times I$ and $\Gamma_q \times I$, respectively. Equations $(8.1)_2$ and $(8.1)_3$ are the *time-dependent Neumann* and *time-dependent Dirichlet* conditions, respectively. Finally, equation $(8.1)_4$ is the *initial condition*. The strong form of the initial/boundary-value problem is stated as follows: given functions k, ρ, c, u_0, f, $\bar{q}$ and $\bar{u}$, find a function u that satisfies equations (8.1).

A Galerkin-based weighted-residual form of the above problem can be deduced by assuming that: (i) the time-dependent Dirichlet boundary conditions are satisfied *a priori* by the choice of the space of admissible solutions $\mathcal{U}$, hence the weighting function w_u vanishes, i.e., $w_u = 0$ on $\Gamma_u \times I$, (ii) the remaining weighting functions satisfy $w_\Omega = w$ in $\Omega \times I$, $w_q = w$ on $\Gamma_q \times I$, (iii) $w = 0$ on $\Gamma_u \times I$, and (iv) the initial condition is satisfied *a priori* in Ω, hence it also enters the space of admissible solutions $\mathcal{U}$.

Taking into account the preceding assumptions, one may write a weighted-residual statement of the form

$$
\begin{aligned}
\int_{\Omega \times I} w \left[-\rho c \frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1}\left(k\,\frac{\partial u}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(k\,\frac{\partial u}{\partial x_2}\right) - f \right] d(\Omega \times I) \\
- \int_{\Gamma_q \times I} w \left[k\,\frac{\partial u}{\partial n} + \bar{q} \right] d(\Gamma \times I) \;=\; 0 \;.
\end{aligned}
\tag{8.2}
$$

Clearly, equation (8.2) involves a space-time integral. Since the spatial and temporal dimensions are independent of each other, they may be readily decoupled, so that (8.2) can be rewritten as

$$\int_I \int_\Omega w \left[-\rho c \frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1}\left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2}\left(k \frac{\partial u}{\partial x_2} \right) - f \right] d\Omega dt \\ - \int_I \int_{\Gamma_q} w \left[k \frac{\partial u}{\partial n} + \bar{q} \right] d\Gamma dt \;=\; 0 \; . \quad (8.3)$$

One may taking advantage of this decoupling and "freeze" time in order to first operate on the space integrals, i.e., on the integro-differential equation

$$\int_\Omega w \left[-\rho c \frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1}\left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2}\left(k \frac{\partial u}{\partial x_2} \right) - f \right] d\Omega \\ - \int_{\Gamma_q} w \left[k \frac{\partial u}{\partial n} + \bar{q} \right] d\Gamma \;=\; 0 \; , \quad (8.4)$$

which, upon using integration by parts, the divergence theorem, and assumption (iii) takes the form

$$\int_\Omega w \rho c \frac{\partial u}{\partial t} \, d\Omega + \int_\Omega \left[\frac{\partial w}{\partial x_1} k \frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2} k \frac{\partial u}{\partial x_2} \;+\; wf \right] d\Omega \;+\; \int_{\Gamma_q} w \bar{q} \, d\Gamma \;=\; 0 \; . \quad (8.5)$$

The Galerkin weighted-residual form can be now stated as follows: given k, ρ, c, f, and $\bar{q}$, find a function $u \in \mathcal{U}$, such that

$$\int_I \left[\int_\Omega w \rho c \frac{\partial u}{\partial t} \, d\Omega + \int_\Omega \left[\frac{\partial w}{\partial x_1} k \frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2} k \frac{\partial u}{\partial x_2} \;+\; wf \right] d\Omega \;+\; \int_{\Gamma_q} w \bar{q} \, d\Gamma \right] dt \;=\; 0 \; , \quad (8.6)$$

for all $w \in \mathcal{W}$. Here, the space of admissible solutions $\mathcal{U}$ and the space of weighting functions $\mathcal{W}$ are defined respectively as

$$\mathcal{U} \;=\; \left\{ u \in H^1(\Omega \times I) \quad | \quad u = \bar{u} \text{ on } \Gamma_u \times I \quad , \quad u(x_1, x_2, 0) = u_0 \right\} \; , \quad (8.7)$$

and

$$\mathcal{W} \;=\; \left\{ w \in H^1(\Omega \times I) \quad | \quad w = 0 \text{ on } \Gamma_u \times I \quad , \quad w(x_1, x_2, 0) = 0 \right\} \; . \quad (8.8)$$

A Bubnov-Galerkin approximation of the weak form (8.6) can be effected by writing

$$\begin{aligned} u \doteq u_h \;&=\; \sum_{I=1}^{N} \varphi_I(x_1, x_2) \, u_I(t) \;+\; u_b(x_1, x_2, t) \; , \\ w \doteq w_h \;&=\; \sum_{I=1}^{N} \varphi_I(x_1, x_2) \, w_I(t) \; , \end{aligned} \quad (8.9)$$

where $\varphi_I = 0$ on Γ_u and $u_I(0) = w_I(0) = 0$. Also, the function $u_b(x_1, x_2, t)$ is chosen to satisfy the time-dependent Neumann condition $(8.1)_3$ and the initial condition $(8.1)_4$. It is clear from (8.9) that the approximation induces a separation of spatial and temporal variables, which plays an essential role in the ensuing developments.

Substitution of u_h and w_h into the weak form (8.6) leads to

$$\begin{aligned}\int_I \Bigg[\sum_{I=1}^{N} w_I \int_\Omega \varphi_I \rho c (\sum_{J=1}^{N} \varphi_J \dot{u}_J + \dot{u}_b) \, d\Omega \\ + \sum_{I=1}^{N} w_I \int_\Omega \{\varphi_{I,1} \; \varphi_{I,2}\} k \left(\sum_{J=1}^{N} \left\{ \begin{array}{c} \varphi_{J,1} \\ \varphi_{J,2} \end{array} \right\} u_J + \left\{ \begin{array}{c} u_{b,1} \\ u_{b,2} \end{array} \right\} \right) d\Omega \\ + \sum_{I=1}^{N} w_I \int_\Omega \varphi_I f \, d\Omega + \sum_{I=1}^{N} w_I \int_{\Gamma_q} \varphi_I \bar{q} \, d\Gamma \Bigg] \, dt \;=\; 0 \; . \qquad (8.10)\end{aligned}$$

This equation may be rewritten as

$$\int_I \left[\sum_{I=1}^{N} w_I \{ \sum_{J=1}^{N} (M_{IJ} \dot{u}_J + K_{IJ} u_J) - F_I \} \right] \;=\; 0 \; , \qquad (8.11)$$

where

$$M_{IJ} \;=\; \int_\Omega \varphi_I \rho c \varphi_J \, d\Omega \; , \qquad (8.12)$$

$$K_{IJ} \;=\; \int_\Omega \{\varphi_{I,1} \; \varphi_{I,2}\} \, k \left\{ \begin{array}{c} \varphi_{J,1} \\ \varphi_{J,2} \end{array} \right\} d\Omega \; , \qquad (8.13)$$

and

$$F_I \;=\; -\int_\Omega \varphi_I \rho c \dot{u}_b \, d\Omega - \int_\Omega \{\varphi_{I,1} \; \varphi_{I,2}\} \, k \left\{ \begin{array}{c} u_{b,1} \\ u_{b,2} \end{array} \right\} d\Omega \; - \int_\Omega \varphi_I f \, d\Omega \; - \int_{\Gamma_q} \varphi_I \, \bar{q} \, d\Gamma \; . \qquad (8.14)$$

The arrays $[\mathbf{M}]$, $[\mathbf{K}]$ and $[\mathbf{F}_I]$ are termed the *mass* (or *capacitance*) matrix, the *stiffness* matrix and the *forcing* vector, respectively. Clearly, both $[\mathbf{M}]$ and $[\mathbf{K}]$ are symmetric, while it is easy to establish that $[\mathbf{M}]$ is also positive-definite as long as $\rho c > 0$, and $[\mathbf{K}]$ is positive-semidefinite provided $k > 0$, as argued for the steady problem.

In conclusion, one has arrived at the semi-discrete form (8.11), which may be also written as

$$\int_I [\mathbf{w}]^T ([\mathbf{M}][\dot{\mathbf{u}}] + [\mathbf{K}][\mathbf{u}] - [\mathbf{F}]) \, dt \;=\; 0 \; , \qquad (8.15)$$

where $[\mathbf{u}] = [u_1(t)\ u_2(t)\ \ldots\ u_N(t)]^T$ and $[\mathbf{w}] = [w_1(t)\ w_2(t)\ \ldots\ w_N(t)]^T$. Equation (8.15) is now an integro-differential equation in time only, as all the spatial derivatives and integrals have been evaluated and "stored" in the arrays $[\mathbf{M}]$, $[\mathbf{K}]$ and $[\mathbf{F}]$.

Once the spatial problem has been discretized, one may proceed to the temporal problem. Here, there are two distinct options:

(a) Discretize $[\mathbf{u}]$ and $[\mathbf{w}]$ in time according to some polynomial series, i.e.,

$$[\mathbf{u}] \ \doteq\ [\hat{\mathbf{u}}] \ =\ \sum_{n=1}^{M} [\boldsymbol{\alpha}_n] t^n \qquad , \qquad [\mathbf{w}] \ \doteq\ [\hat{\mathbf{w}}] \ =\ \sum_{n=1}^{M} [\boldsymbol{\beta}_n] t^n \ , \tag{8.16}$$

where $[\boldsymbol{\alpha}_n]$ is a vector to be determined and $[\boldsymbol{\beta}_n]$ is an arbitrary vector. These approximate functions are then substituted into the semi-discrete form (8.15) and the resulting system is solved for the values of $[\boldsymbol{\alpha}_n]$. This is essentially a Bubnov-Galerkin approximation in time.

(b) Apply a standard discrete time integrator directly on the semi-discrete form (8.15). This amounts to choosing $\mathbf{w}$ to consist of Dirac-delta functions at discrete times t_1, t_2, ..., t_n, t_{n+1}, ... , which would imply that the system of ordinary differential equations

$$[\mathbf{M}][\dot{\mathbf{u}}] + [\mathbf{K}][\mathbf{u}] \ =\ [\mathbf{F}] \tag{8.17}$$

is to be exactly satisfied at these times.

In the remainder of this section, the second option is pursued. To this end, recall that the general solution of the homogeneous counterpart of (8.17), i.e., when $[\mathbf{F}] = [\mathbf{0}]$, is of the form

$$[\mathbf{u}(t)] \ =\ \sum_{I=1}^{N} c_I e^{-\lambda_I t} [\mathbf{z}_I] \ , \tag{8.18}$$

where the pairs $(\lambda_I, [\mathbf{z}_I])$, $I = 1, 2, \ldots, N$, are to be determined. Upon substituting a typical such pair $(\lambda, [\mathbf{z}])$ into the homogeneous equation, one gets

$$e^{-\lambda t}(-\lambda[\mathbf{M}] + [\mathbf{K}])[\mathbf{z}] \ =\ [\mathbf{0}] \ , \tag{8.19}$$

hence,

$$\lambda[\mathbf{M}][\mathbf{z}] \ =\ [\mathbf{K}][\mathbf{z}] \ . \tag{8.20}$$

Equation (8.20) corresponds to the general symmetric linear eigenvalue problem, which can be solved for the eigenpairs $(\lambda_I, [\mathbf{z}_I])$, $I = 1, 2, \ldots, N$. For notational simplicity, define the $(N \times N)$ arrays

$$[\mathbf{\Lambda}] \; = \; \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{bmatrix} \tag{8.21}$$

and

$$[\mathbf{Z}] \; = \; [\mathbf{z}_1 \; \mathbf{z}_2 \; \ldots \; \mathbf{z}_N] \; , \tag{8.22}$$

so that the eigenvalue problem (8.20) may be conveniently rewritten as

$$[\mathbf{M}][\mathbf{Z}][\mathbf{\Lambda}] \; = \; [\mathbf{K}][\mathbf{Z}] \; . \tag{8.23}$$

Given that $[\mathbf{M}]$ and $[\mathbf{K}]$ are symmetric, standard orthogonality properties of the eigenpairs $(\lambda_I, \mathbf{z}_I)$, $I = 1, 2, \ldots, N$, where λ_I are assumed distinct, yield the diagonalizations

$$[\mathbf{Z}]^T[\mathbf{M}][\mathbf{Z}] \; = \; \begin{bmatrix} m_1 & & & \\ & m_2 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix} \tag{8.24}$$

and

$$[\mathbf{Z}]^T[\mathbf{K}][\mathbf{Z}] \; = \; \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_N \end{bmatrix} \; , \tag{8.25}$$

where $\lambda_I = \dfrac{k_I}{m_I}$, and $m_I > 0$, $k_I \geq 0$, thus $\lambda_I \geq 0$. Indeed, starting from (8.20) for an eigenpair $(\lambda_I, \mathbf{z}_I)$ and premultiplying both sides by $[\mathbf{z}_J]^T$ yields

$$\lambda_I [\mathbf{z}_J]^T[\mathbf{M}][\mathbf{z}_I] \; = \; [\mathbf{z}_J]^T[\mathbf{K}][\mathbf{z}_I] \; . \tag{8.26}$$

Conversely, starting from (8.20) for an eigenpair $(\lambda_J, \mathbf{z}_J)$ and premultiplying both sides by $[\mathbf{z}_I]^T$ leads to

$$\lambda_J [\mathbf{z}_I]^T[\mathbf{M}][\mathbf{z}_J] \; = \; [\mathbf{z}_I]^T[\mathbf{K}][\mathbf{z}_J] \; . \tag{8.27}$$

Taking into account the symmetry of $[\mathbf{M}]$ and $[\mathbf{K}]$, equations (8.26) and (8.27) imply that

$$(\lambda_I - \lambda_J)[\mathbf{z}_J]^T[\mathbf{M}][\mathbf{z}_I] \;=\; 0 \; . \tag{8.28}$$

Equation (8.28) proves the diagonalization of $[\mathbf{M}]$, as in (8.24), provided that $\lambda_I \neq \lambda_J$. Either (8.26) or (8.27) may be subsequently invoked to deduce the simultaneous diagonalization of $[\mathbf{K}]$, as in (8.25).

The solution of the non-homogeneous equations (8.17) is attained by employing the classical technique of variation of parameters, according to which it is assumed that

$$[\mathbf{u}(t)] \;=\; [\mathbf{Z}][\mathbf{v}(t)] \; , \tag{8.29}$$

where $[\mathbf{Z}]$ is obtained from the homogeneous problem. Substituting (8.29) into (8.17) gives rise to

$$[\mathbf{M}][\mathbf{Z}][\dot{\mathbf{v}}] + [\mathbf{K}][\mathbf{Z}][\mathbf{v}] \;=\; [\mathbf{F}] \; . \tag{8.30}$$

Premultiplying this equation by $[\mathbf{Z}]^T$ leads to

$$[\mathbf{Z}]^T[\mathbf{M}][\mathbf{Z}][\dot{\mathbf{v}}] + [\mathbf{Z}]^T[\mathbf{K}][\mathbf{Z}][\mathbf{v}] \;=\; [\mathbf{Z}]^T[\mathbf{F}] \; , \tag{8.31}$$

or, taking into account the earlier orthogonality conditions,

$$\begin{bmatrix} m_1 & & & \\ & m_2 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix} \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \vdots \\ \dot{v}_N \end{bmatrix} + \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_N \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{bmatrix} \; , \tag{8.32}$$

where $g_i = [\mathbf{z}_i]^T[\mathbf{F}]$. It is now concluded that the original system of coupled linear ordinary differential equations (8.15) has been reduced to a set of N uncoupled scalar ordinary differential equations of the form

$$m_I\dot{v}_I + k_I v_I = g_I \; , \tag{8.33}$$

where $I = 1, 2, \ldots, N$. Therefore, in order to understand the behavior of the original system, one only needs to study the solution of a single scalar ordinary differential equation of the form

$$m\dot{v} + kv \;=\; g \; , \tag{8.34}$$

with initial condition $v(0) = v_0$.

The general solution of equation (8.34) can be obtained using the method of variation of parameters, and is given by

$$v(t) \ = \ e^{-\lambda t} y(t) \ , \tag{8.35}$$

where $\lambda = \dfrac{k}{m}$ and $y = y(t)$ is a function to be determined. Upon substituting the general solution into (8.34) and simplifying, the resulting expression leads to

$$\dot{y}(t) \ = \ \frac{1}{m} e^{\lambda t} g \ . \tag{8.36}$$

This equation can be integrated in the time interval $I_{n+1} = (t_n, t_{n+1}]$ and results in

$$y(t) \ = \ y_n + \int_{t_n}^{t} \frac{1}{m} e^{\lambda \tau} g(\tau)\, d\tau \ , \tag{8.37}$$

where $y_n = y(t_n)$. Hence, one obtains from (8.35) the solution for $v(t)$ in convolution form as

$$v(t) \ = \ e^{-\lambda t} y_n + \int_{t_n}^{t} \frac{1}{m} e^{\lambda(\tau - t)} g(\tau)\, d\tau \ . \tag{8.38}$$

Noting, further, that $v(t_n) = v_n = e^{-\lambda t_n} y_n$, it follows that the preceding solution can be also expressed as

$$v(t) \ = \ e^{-\lambda(t - t_n)} v_n + \int_{t_n}^{t} \frac{1}{m} e^{-\lambda(t - \tau)} g(\tau)\, d\tau \ . \tag{8.39}$$

Now, setting $t = t_{n+1}$, it is readily seen from (8.39) that

$$v_{n+1} \ = \ e^{-\lambda \Delta t_n} v_n + \int_{t_n}^{t_n+1} \frac{1}{m} e^{-\lambda(t_{n+1} - \tau)} g(\tau)\, d\tau \ , \tag{8.40}$$

where $\Delta t_n = t_{n+1} - t_n$. The ratio $\dfrac{v_{n+1}}{v_n} = r$ is termed *the amplification factor*. In the homogeneous case ($g = 0$), equation (8.40) immediately implies that $r = e^{-\lambda \Delta t_n}$, i.e., the exact solution experiences exponential decay. This, in turn, implies that $r \rightarrow 1$ when $\lambda \Delta t_n \rightarrow 0$ and $r \rightarrow 0$ when $\lambda \Delta t_n \rightarrow \infty$.

8.2 Stability of classical time integrators

In this section, attention is focused on the application of certain discrete time integrators to the scalar first-order differential equation (8.34), which, as argued in the preceding section, fully represents the general system (8.17) obtained through the semi-discretization of the weak form (8.2).

The first discrete time integrator is the *forward Euler method*, according to which the time derivative $\dot{v}$ can be approximated at time t_{n+1} by using a Taylor series expansion of $v(t)$ at t_n as

$$v(t_{n+1}) \;=\; v(t_n) + \Delta t_n \dot{v}(t_n) + o(\Delta t_n^2)\ , \tag{8.41}$$

which, upon ignoring the second-order terms in $\Delta t_n = t_{n+1} - t_n$, leads to

$$\dot{v}(t_n) \;\doteq\; \frac{v_{n+1} - v_n}{\Delta t_n}\ . \tag{8.42}$$

Upon writing (8.17) at t_n with $\dot{v}(t_n)$ computed from (8.42), it is concluded that

$$m\,\frac{v_{n+1} - v_n}{\Delta t_n} + k v_n \;=\; g_n\ , \tag{8.43}$$

where $v_k = v(t_k)$. This equation may be trivially rewritten as

$$v_{n+1} \;=\; (1 - \lambda \Delta t_n) v_n + \frac{\Delta t_n}{m}\, g_n\ , \tag{8.44}$$

where, again $\lambda = \dfrac{k}{m}$. In the homogeneous case ($g = 0$), it is seen from (8.44) that the discrete amplification ratio r_f of the forward Euler method is given by

$$r_f \;=\; 1 - \lambda \Delta t_n\ . \tag{8.45}$$

Equation (8.45) implies that for finite values of λ, the limiting case $\Delta t_n \to 0$ leads to $r_f \to 1$, which is consistent with the exact solution, as argued earlier in this section. However, the limiting case $\Delta t_n \to \infty$ leads to $r_f \to -\infty$, which reveals that the discrete solution does not predict exponential decay in the limit of an infinitely large time step Δt_n. As can be easily inferred from (8.44), the forward Euler method is only conditionally stable. Indeed, ignoring the inhomogeneous term, it is clear that for $\lambda \Delta t_n > 1$, the discrete solution exhibits oscillations with respect to $v = 0$ (which are, of course, absent in the exact exponentially decaying solution). For $1 < \lambda \Delta t_n < 2$, these oscillations are decaying, hence the discrete solution is *stable*. However, for $\lambda \Delta t_n > 2$, the oscillations grow in magnitude with each time step and the solution becomes *unstable*, i.e., instead of decaying, it artificially grows toward infinity. Therefore, the forward Euler method is referred to as a *conditionally stable* method, which means that its time step Δt_n needs to be controlled in order to satisfy the condition $\Delta t_n < \dfrac{2}{\lambda} = \Delta t_{cr}$. In systems with many degrees of freedom, such as (8.17), the *critical step-size* Δt_{cr} is defined as

$$\Delta t_{cr} \;=\; \frac{2}{\lambda_{max}}\ , \tag{8.46}$$

where λ_{max} is the maximum eigenvalue of problem (8.20). This implies that in order to guarantee stability for the forward Euler method, one needs to know (or estimate) the maximum eigenvalue of (8.20). Fortunately, there exist inexpensive methods of estimating λ_{max} in finite element approximations, a fact that significantly enhances the usefulness of the forward Euler method.

A simple scaling argument can be made for the dependence of λ_{max} on the element size h. To this end, recall that, since the interpolation functions φ_I are dimensionless, the components of the stiffness matrix in the two-dimensional transient heat conduction problem are of order $o(1)$, while the components $[M_{IJ}]$ of the mass matrix are of order $o(h^2)$. This implies that, by virtue of its definition, λ is of order $o(h^{-2})$, hence Δt_{cr} is of order $o(h^2)$. This means that, when using forward Euler integration in the solution of the two-dimensional transient heat conduction equation, the critical step-size must be reduced quadratically under mesh refinement, i.e., halving the mesh-size necessitates reduction of the step-size by a factor of four. Similar scaling arguments can be made for one- or three-dimensional versions of the transient heat conduction equation.

An alternative discrete time integrator is the *backward Euler method*, which can be deduced by writing $v(t_n)$ using a Taylor series expansion at t_{n+1} as

$$v(t_n) \ = \ v(t_{n+1}) - \Delta t_n \dot{v}(t_{n+1}) + o(\Delta t_n^2) \ , \tag{8.47}$$

which, upon ignoring the second-order terms in Δt_n leads to

$$\dot{v}(t_{n+1}) \ \doteq \ \frac{v_{n+1} - v_n}{\Delta t_n} \ . \tag{8.48}$$

Writing now (8.34) at t_{n+1}, with $\dot{v}(t_{n+1})$ estimated from (8.48), results in

$$m\,\frac{v_{n+1} - v_n}{\Delta t_n} + kv_{n+1} = g_{n+1} \tag{8.49}$$

or, upon solving for v_{n+1},

$$v_{n+1} \ = \ \frac{1}{1 + \lambda\Delta t_n}\, v_n + \frac{\Delta t_n}{1 + \lambda\Delta t_n}\frac{1}{m}\, g_{n+1} \ . \tag{8.50}$$

For the homogeneous problem, equation (8.50) implies that in the limiting cases $\Delta t_n \to 0$ and $\Delta t_n \to \infty$, the discrete amplification ratio r_b, defined as

$$r_b \ = \ \frac{1}{1 + \lambda\Delta t_n} \ , \tag{8.51}$$

satisfies $A \rightarrow 1$ and $A \rightarrow 0$, respectively. This means that the backward Euler method is consistent with the exact solution in both extreme cases. In addition, as seen from (8.51), this method is *unconditionally stable*, in the sense that it yields numerical approximations to $v(t)$ that are decaying in time (without any oscillations!) regardless of the step-size Δt_n. Figure 8.1 shows the amplification factor for the two methods, as well as for the exact solution.

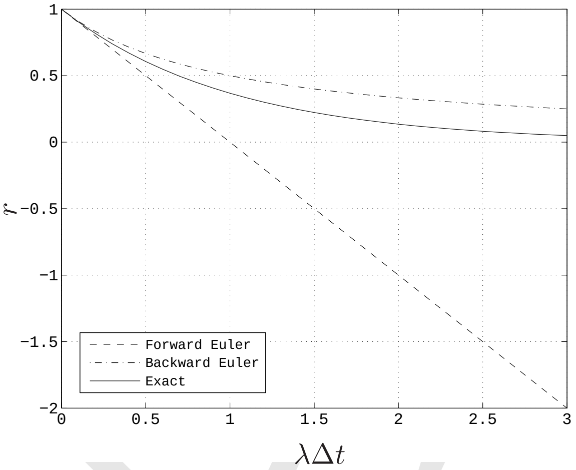

Figure 8.1: *Amplification factor r as a function of $\lambda \Delta t$ for forward Euler, backward Euler and the exact solution of the homogeneous counterpart of* (8.34)

Returning to the system of ordinary differential equations in (8.17), one may use forward Euler integration at time t_n, which leads to

$$[\mathbf{M}] \, \frac{[\mathbf{u}_{n+1}] - [\mathbf{u}_n]}{\Delta t_n} + [\mathbf{K}][\mathbf{u}_n] \;=\; [\mathbf{F}_n] \; , \tag{8.52}$$

hence

$$[\mathbf{M}][\mathbf{u}_{n+1}] \;=\; [\mathbf{M}][\mathbf{u}_n] - \Delta t_n [\mathbf{K}][\mathbf{u}_n] + \Delta t_n [\mathbf{F}_n] \; . \tag{8.53}$$

It is clear that computing $[\mathbf{u}_{n+1}]$ requires the factorization of $[\mathbf{M}]$, which may be performed once and be used repeatedly for $n = 1, 2, \ldots$. In fact, the factorization itself may become unnecessary if $[\mathbf{M}]$ is diagonal, in which case $[\mathbf{M}]^{-1}$ can be obtained from $[\mathbf{M}]$ by merely inverting its diagonal components. In this case, it is clear that the advancement of the solution from $[\mathbf{u}_n]$ to $[\mathbf{u}_{n+1}]$ does not require the solution of an algebraic system. For this reason, the resulting semi-discrete method is termed *explicit*. A diagonal estimate of the mass matrix $[\mathbf{M}]$ can be easily computed using nodal quadrature, i.e., by evaluating the integral expression that defines its components using an integration rule that takes the element nodes as its sampling points. This observation can be readily justified by recalling the definition of the components M_{IJ} of the mass matrix in (8.12) and the properties of the element interpolation functions.

The backward Euler method can also be applied to (8.17) at time t_{n+1}, resulting in

$$[\mathbf{M}]\,\frac{[\mathbf{u}_{n+1}]-[\mathbf{u}_n]}{\Delta t_n}+[\mathbf{K}][\mathbf{u}_{n+1}] \;=\; [\mathbf{F}_{n+1}]\ , \tag{8.54}$$

which implies that

$$([\mathbf{M}]+\Delta t_n[\mathbf{K}])[\mathbf{u}_{n+1}] \;=\; [\mathbf{M}][\mathbf{u}_n]+\Delta t_n[\mathbf{F}_n]\ . \tag{8.55}$$

The above system requires factorization of $[\mathbf{M}]+\Delta t_n[\mathbf{K}]$, which cannot be circumvented by diagonalization, as in the forward Euler case. Hence, the resulting semi-discrete method is termed *implicit*, in the sense that advancement of the solution from $[\mathbf{u}_n]$ to $[\mathbf{u}_{n+1}]$ cannot be achieved without the solution of algebraic equations.

Explicit and implicit semi-discrete methods give rise to vastly different computer code architectures. In the former case, emphasis is placed on the control of step-size Δt_n, so that is always remain below the critical value Δt_{cr}. In the latter, emphasis is placed on the efficient solution of the resulting algebraic equations.

8.3 Weighted-residual interpretation of classical time integrators

It is interesting to re-derive the discrete time integrators of the previous section using a weighted-residual formalization. To this end, start from equation (8.15) and consider the time interval $I=(t_n,t_{n+1}]$, where

$$\int_{t_n}^{t_{n+1}}[\mathbf{w}]^T([\mathbf{M}][\dot{\mathbf{u}}]+[\mathbf{K}][\mathbf{u}]-[\mathbf{F}])\,dt \;=\; 0\ . \tag{8.56}$$

Now, choose a linear polynomial interpolation of $[\mathbf{u}]$ in time, namely

$$[\mathbf{u}] \;\doteq\; [\hat{\mathbf{u}}] \;=\; \left(1-\frac{t-t_n}{\Delta t_n}\right)[\mathbf{u}_n]+\frac{t-t_n}{\Delta t_n}\,[\mathbf{u}_{n+1}]\ , \tag{8.57}$$

where $[\mathbf{u}_n]$ is known from the integration in the previous time interval $(t_{n-1},t_n]$.

Different discrete time integrators can be deduced by appropriate choices of the weighting function $[\mathbf{w}]$. Specifically, let

$$[\mathbf{w}] \doteq [\hat{\mathbf{w}}] = \delta(t_n^+)[\mathbf{c}] \, , \tag{8.58}$$

in (t_n, t_{n+1}), where $\mathbf{c}$ is an arbitrary constant vector. Substituting (8.57) and (8.58) into (8.56), one obtains (8.52), thus recovering the semi-discrete equations of the forward Euler rule. Alternatively, setting

$$[\mathbf{w}] \doteq [\hat{\mathbf{w}}] = \delta(t_{n+1})[\mathbf{c}] \, , \tag{8.59}$$

one readily obtains (8.54), namely the semi-discrete equations of the backward Euler rule.

More generally, let

$$[\mathbf{w}] \doteq [\hat{\mathbf{w}}] = \delta(t_{n+\alpha})[\mathbf{c}] \, , \tag{8.60}$$

where $0 < \alpha \leq 1$. Substituting (8.57) and (8.60) into (8.56) leads to

$$[\mathbf{M}] \, \frac{[\mathbf{u}_{n+1}] - [\mathbf{u}_n]}{\Delta t_n} + [\mathbf{K}] \Big[(1-\alpha)[\mathbf{u}_n] + \alpha[\mathbf{u}_{n+1}] \Big] = [\mathbf{F}_{n+\alpha}] \, , \tag{8.61}$$

which corresponds to the *generalized trapezoidal rule.* For the special case $\alpha = 1/2$, one recovers the *Crank-Nicolson rule.*

Finally, one may choose to use a smooth interpolation for the weighting function $[\mathbf{w}]$ in $(t_n, t_{n+1}]$. Indeed, let

$$[\mathbf{w}] \doteq [\hat{\mathbf{w}}] = \frac{t - t_n}{\Delta t_n} [\mathbf{w}_{n+1}] \, , \tag{8.62}$$

where $\mathbf{w}_{n+1}$ is an arbitrary constant vector. In this case, one recovers the Bubnov-Galerkin method in time. In particular, substituting (8.57) and (8.62) into (8.56) leads to

$$\int_{t_n}^{t_{n+1}} \frac{t - t_n}{\Delta t_n} [\mathbf{w}_{n+1}]^T \left[[\mathbf{M}] \, \frac{[\mathbf{u}_{n+1}] - [\mathbf{u}_n]}{\Delta t_n} + [\mathbf{K}] \left\{ \left(1 - \frac{t - t_n}{\Delta t_n} \right) [\mathbf{u}_n] + \frac{t - t_n}{\Delta t_n} \, [\mathbf{u}_{n+1}] \right\} - [\mathbf{F}] \right] \, dt = 0 \, . \tag{8.63}$$

Upon integrating (8.63) in time and recalling that $\mathbf{w}_{n+1}$ is arbitrary, one finds that

$$\frac{1}{2} \, [\mathbf{M}]([\mathbf{u}_{n+1}] - [\mathbf{u}_n]) + [\mathbf{K}](\frac{1}{6} \, [\mathbf{u}_n] + \frac{1}{3} \, [\mathbf{u}_{n+1}])\Delta t_n - \int_{t_n}^{t_{n+1}} \frac{t - t_n}{\Delta t_n} \, [\mathbf{F}] \, dt = 0 \tag{8.64}$$

or

$$([\mathbf{M}] + \frac{2}{3} \, \Delta t_n [\mathbf{K}])[\mathbf{u}_{n+1}] = ([\mathbf{M}] - \frac{1}{3} \, \Delta t_n [\mathbf{K}])[\mathbf{u}_n] + [\bar{\mathbf{F}}] \, , \tag{8.65}$$

where $[\bar{\mathbf{F}}] = \int_{t_n}^{t_{n+1}} 2 \, \frac{t - t_n}{\Delta t_n} \, [\mathbf{F}] \, dt$. When $[\mathbf{F}]$ is constant, the Bubnov-Galerkin method coincides with the generalized trapezoidal rule with $\alpha = 2/3$.