Lecture Note #2 **MTH652: Advanced Numerical Analysis**
Winter 2025 Instructor: Will Pazner

Time Discretization

We now briefly review some basic methods for approximating the solution to ODEs; in the context of time-dependent PDEs, these methods are referred to as "time stepping" or "time integration" methods.

Recall that in the previous lecture we wrote the semi-discrete heat equation as

$$\frac{\partial \boldsymbol{\eta}}{\partial t} + \bar{A}\boldsymbol{\eta} = \boldsymbol{f}.$$

where the matrix $\bar{A}$ is an SPD matrix. Our model problem will be the system of ODEs

$$(*) \begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} = B\boldsymbol{u}, \\ \boldsymbol{u}(0) = \boldsymbol{u}_0, \end{cases}$$

for any *negative-definite* B. The exact solution to this problem is given by

$$\boldsymbol{u}(t) = e^{Bt}\boldsymbol{u}_0.$$

Since A is SPD, the norm of u decreases exponentially with t.

Typically, in order to discretize $(*)$, we choose a **time step** Δt, and look for a sequence

$$\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_K,$$

such that

$$\boldsymbol{u}_i \approx \boldsymbol{u}(i\Delta t).$$

The initial condition gives the first vector in this sequence.

1 Forward and Backward Euler Methods

1.1 Forward Euler

The simplest possible method is the **forward Euler** method. The method uses a simple difference quotient to approximate the term $\partial \boldsymbol{u}/\partial t$,

$$\frac{\partial \boldsymbol{u}(t)}{\partial t} \approx \frac{\boldsymbol{u}(t+\Delta t) - \boldsymbol{u}(t)}{\Delta t}. \tag{1}$$

So, $(*)$ is approximated as

$$\frac{\boldsymbol{u}(t+\Delta t) - \boldsymbol{u}(t)}{\Delta t} \approx B\boldsymbol{u}(t).$$

Setting $t = 0$, and using the initial condition, we obtain

$$\frac{\boldsymbol{u}(\Delta t) - \boldsymbol{u}_0}{\Delta t} \approx B\boldsymbol{u}_0.$$

Rearranging, this gives

$$\boldsymbol{u}(\Delta t) \approx \boldsymbol{u}_1 := \boldsymbol{u}_0 + \Delta t B \boldsymbol{u}_0. \tag{2}$$

Repeating this process, the general method can be written as

$$\boldsymbol{u}_{i+1} = \boldsymbol{u}_i + \Delta t B \boldsymbol{u}_i = (I + \Delta t B)\boldsymbol{u}_i.$$

This is a very simple method: the solution is incremented at each time step simply by taking a combination of $\boldsymbol{u}_i$ and $B\boldsymbol{u}_i$. It is an example of an **explicit method**, requiring only the *evaluation* of the right-hand side (i.e. matrix multiplication with B), and not the solution of a system of equations (this generalizes to nonlinear ODEs as well). In the previous term, we saw that the difference quotient (1) has order of accuracy $\mathcal{O}(\Delta t)$, and indeed it can be shown that the forward Euler method is *first-order accurate*.

We know that the exact solution to our model problem has *decreasing norm*, i.e.

$$\|\boldsymbol{u}(t_2)\| \leq \|\boldsymbol{u}(t_1)\| \quad \text{whenever } t_2 > t_1.$$

We would like to have the same property at the discrete level, which can be written

$$\|\boldsymbol{u}_{i+1}\| \leq \|\boldsymbol{u}_i\|.$$

Since we have the update (2), we can bound the norm $\|\boldsymbol{u}_{i+1}\|$ using the spectrum of the symmetric matrix $I + \Delta t B$. The eigenvalues of this matrix are given by

$$\mu = 1 + \Delta t \lambda, \quad \text{where } \lambda \text{ is an eigenvalue of } B.$$

We need the condition

$$|1 + \Delta t \lambda| < 1,$$

This means that the quantity $\Delta t \lambda$ has to lie within the circle of radius 1 centered at the point -1. This means that the time step has to satisfy the condition

$$\Delta t < \frac{2}{|\lambda|}$$

If the time step is chosen too large, then the discrete solution will not be decreasing in norm; it will "blow up". The maximum (absolute) eigenvalue of the matrix B determines the time step restriction.

Note that in our example of the heat equation, the matrix B is given by

$$B = -\bar{A} = E^{-T} A E^{-1},$$

where A is the stiffness matrix, and E is the Cholesky factor of the mass matrix. We saw that this means that the largest eigenvalue of $\bar{A}$ will scale like h^{-2}. Consequently, we have the stability condition

$$\Delta t \leq C h^2.$$

(The type of stability restriction relating the mesh size h to the time step Δt is known as the *CFL condition*, after a 1928 paper of Courant, Friedrichs, and Lewy). This is a **severely restrictive** time step condition; each uniform refinement of the finite element method requires taking *four times* the number of time steps. It is not enough to have $\Delta t, h \to 0$ to obtain convergence; the resulting fully discrete method will diverge if Δt does not go to zero like the square of h.

1.2 Backward Euler

The **backward Euler** method is equally simple to derive (but its implementation is more difficult because it is an **implicit method**, requiring the solution of a system of equations at each time step). Instead of the forward difference quotient (1), consider the *backwards difference* quotient

$$\frac{\partial \boldsymbol{u}(t)}{\partial t} \approx \frac{\boldsymbol{u}(t) - \boldsymbol{u}(t - \Delta t)}{\Delta t}. \tag{3}$$

Using this as an approximation for the left-hand side of $(*)$ (evaluated at the point $t + \Delta t$), we obtain

$$\frac{\boldsymbol{u}(t + \Delta t) - \boldsymbol{u}(t)}{\Delta t} \approx B\boldsymbol{u}(t + \Delta t).$$

Rearranging,

$$\boldsymbol{u}(t + \Delta t) - \Delta t B\boldsymbol{u}(t + \Delta t) \approx \boldsymbol{u}(t).$$

This gives the update

$$(I - \Delta t B)\boldsymbol{u}_{i+1} = \boldsymbol{u}_i,$$

which requires the solution of a (positive-definite) system of equations at each time step.

For the backward Euler method, the stability of the method can be related to the eigenvalues of the matrix $(I - \Delta t B)^{-1}$, since the update is given by

$$\boldsymbol{u}_{i+1} = (I - \Delta t B)^{-1}\boldsymbol{u}_i.$$

The eigenvalues are given by

$$\mu = \frac{1}{1 - \Delta t \lambda}, \quad \text{where } \lambda \text{ is an eigenvalue of } B.$$

Since B is negative-definite, $\lambda < 0$, and $1 - \Delta t\lambda > 1$, so $|\mu| < 1$ unconditionally. We therefore say that the backward Euler method is **unconditionally stable** (for this model problem). This means that the method will converge for $\Delta t, h \to 0$, without a condition on Δt relative to h. (However, it still is only first-order accurate, like the forward Euler method).

1.3 Application to the finite element discretization of the heat equation

Since the semi-discrete heat equation can be written as

$$\frac{\partial \boldsymbol{\eta}}{\partial t} + \bar{A}\boldsymbol{\eta} = \boldsymbol{f}$$

it is straightforward to apply the forward and backward Euler methods. The change of variables to $\boldsymbol{\eta}$ is useful for the analysis (since the matrix $\bar{A}$ is SPD), but it is not practical, because it used the Cholesky factorization of M. Instead, we can consider the original system of ODEs (we take $\boldsymbol{f} = 0$ here for simplicity — the forcing term can easily be incorporated)

$$M\frac{\partial \boldsymbol{u}}{\partial t} + A\boldsymbol{u} = \boldsymbol{f}.$$

Then, the forward and backward Euler methods are

$$M\boldsymbol{u}_{i+1}^{FE} = M\boldsymbol{u}_i^{FE} - \Delta t A\boldsymbol{u}_i^{FE}, \tag{4}$$
$$M\boldsymbol{u}_{i+1}^{BE} = M\boldsymbol{u}_i^{BE} - \Delta t A\boldsymbol{u}_{i+1}^{BE}. \tag{5}$$

So, the forward Euler method gives the update

$$\boldsymbol{u}_{i+1}^{FE} = \boldsymbol{u}_i^{FE} - \Delta t M^{-1} A\boldsymbol{u}_i^{FE},$$

whereas the backward Euler method gives the update

$$\boldsymbol{u}_{i+1}^{BE} = (M + \Delta t A)^{-1}\boldsymbol{u}_i^{BE}.$$

Both of these methods require solving linear systems each time step, but forward Euler requires solving a system with the mass matrix, and backward Euler requires solving a system with the matrix $M + \Delta t A$. Recall that the mass matrix is well-conditioned ($\kappa(M) = \mathcal{O}(1)$), so this system is "easy" to solve using e.g. the conjugate gradient method. On the other hand, the matrix $M + \Delta t A$ may be ill-conditioned (depending on Δt). We estimate the Rayleigh quotient

$$\frac{\boldsymbol{u}^T(M + \Delta t A)\boldsymbol{u}}{\|\boldsymbol{u}\|^2}.$$

Recall that

$$\frac{\boldsymbol{u}^T M\boldsymbol{u}}{\|\boldsymbol{u}\|^2} \approx h^2,$$
$$h^2 \lesssim \frac{\boldsymbol{u}^T A\boldsymbol{u}}{\|\boldsymbol{u}\|^2} \lesssim 1,$$

and so

$$(1 + \Delta t)h^2 \lesssim \frac{\boldsymbol{u}^T(M + \Delta t A)\boldsymbol{u}}{\|\boldsymbol{u}\|^2} \lesssim h^2 + \Delta t.$$

We can roughly estimate the condition number as

$$\kappa(M + \Delta t A) \approx \frac{\Delta t}{h^2}.$$

Unless $\Delta t \sim h^2$ (which is also the explicit CFL condition), then the condition number of the system matrix increases, and unpreconditioned CG will require an increasing number of iterations to solve the system. This means that if we want the backward Euler method to be computationally efficient (with large time steps), we will need a good preconditioner for the matrix $M + \Delta t A$.

Note also that the equations (4) and (5) can be written in variational form. The forward Euler method is: given $u_i \in V_h$, find $u_{i+1} \in V_h$ such that, for all $v_h \in V_h$,

$$(u_{i+1}, v_h) = (u_i, v_h) - \Delta t a(u_i, v_h).$$

The backward Euler method is: find $u_{i+1} \in V_h$ such that

$$(u_{i+1}, v_h) = (u_i, v_h) - \Delta t a(u_{i+1}, v_h).$$

2 Qualitative behavior of the heat equation

For simplicity, we will consider the 1D heat equation with zero forcing term,

$$\begin{aligned} u_t(x,t) - u_{xx}(x,t) &= 0, \\ u(x,0) &= u_0(x). \end{aligned}$$

Similar analysis can be extended to the more general case; we use this example to keep the exposition simple. Suppose the initial condition is given as a sum of sine waves,

$$u_0(x) = \sum_{j=0}^{N} a_j \sin(jx).$$

An easy calculation shows that the exact solution to this PDE is given by

$$u(x,t) = \sum_{j=0}^{N} a_j e^{-j^2 t} \sin(jx). \tag{6}$$

In each term in the sum, the *frequency* of the sine wave is given by j. Equation (6) tells us that the coefficients of the jth term decay like $e^{-j^2 t}$; in other words, high frequencies decay much faster than low frequencies. This means that the solution will become smoother with time. For small times, there can be some "initial transients" (large temporal derivatives), but these get damped as time increases.

This qualitative behavior can motivate choices for spatial and temporal discretization. If the solution is dominated by lower frequencies with increasing time, then it makes sense to use a coarser mesh and a larger time step. At the beginning of the simulation, it makes sense to use a smaller time step and finer mesh to adequately resolve the initial transients.

3 Crank-Nicolson and Theta Methods

We can consider a general class of methods ("θ" methods) by introducing a parameter $\theta \in [0,1]$. These methods take the form

$$\frac{\boldsymbol{u}_{i+1} - \boldsymbol{u}_i}{\Delta t} = \theta B \boldsymbol{u}_i + (1-\theta) B \boldsymbol{u}_{i+1}.$$

The difference between the theta method and the Euler methods is that the right-hand side is now a convex combination of the forward and backward Euler methods. If $\theta = 1$, then the method recovers forward Euler; if $\theta = 0$, then the method recovers backward Euler. If $\theta < 1$, then the method is **implicit**: each step requires the solution of a system of equations with B. The special case $\theta = \frac{1}{2}$ is known as the **Crank–Nicolson method**.

The update can be written as

$$(I - (1-\theta)\Delta t B)\boldsymbol{u}_{i+1} = \boldsymbol{u}_i + \theta \Delta t B \boldsymbol{u}_i = (I + \theta \Delta t B)\boldsymbol{u}_i,$$

or

$$\boldsymbol{u}_{i+1} = (I - (1-\theta)\Delta t B)^{-1}(I + \theta \Delta t B)\boldsymbol{u}_i.$$

The norm of the operator $(I - (1-\theta)\Delta t B)^{-1}(I + \theta \Delta t B)$ can be bounded by

$$\max_{\lambda} \frac{1 + \theta \Delta t \lambda}{1 - (1-\theta)\Delta t \lambda}, \quad \text{where } \lambda \text{ is an eigenvalue of } B.$$

For stability, we need (letting $\mu = \Delta t \lambda$)

$$|1 + \theta\mu| \leq |1 - \mu + \theta\mu|,$$

which is satisfied when

$$(2\theta - 1)\mu \geq -2.$$

Since B is negative-definite, $\mu < 0$, and so this is satisfied *unconditionally* whenever $2\theta - 1 \leq 0$, i.e. $\theta \leq \frac{1}{2}$.

This shows that the Crank-Nicolson method is unconditionally stable. It has the additional advantage that it is second-order accurate (as opposed to every other θ method—all others are first-order accurate). This can be seen using a simple Taylor series argument; Crank-Nicolson is analogous to the *trapezoidal rule* for quadrature, and the proof of order of accuracy is essentially the same. There is actually a deep and interesting connection between quadrature rules (for approximating integrals) and numerical methods for ODEs, and quadrature rules can generate entire classes of ODE methods (e.g. "implicit Runge–Kutta methods").