Numerical methods for the heat equation

Consider the following initial-boundary value problem (IBVP) for the one-dimensional heat equation

$$\begin{cases} \dfrac{\partial U}{\partial t} = \alpha \dfrac{\partial^2 U}{\partial x^2} + q(x) \qquad t \geq 0 \qquad x \in [0, L] \\ U(x,0) = U_0(x) \\ U(0,t) = g_0(t) \\ U(L,t) = g_L(t) \end{cases} \tag{1}$$

where $q(x)$ is the internal heat generation and α the thermal diffusivity.

The IBVP (1) describes the propagation of temperature in a one-dimensional slab of width L initially at temperature $U_0(x)$ with Dirichlet boundary conditions $U(0,t) = g_0(t)$ and $U(L,t) = g_L(t)$. You have learned in AM 212A that it is possible to compute the analytical solution of the problem (1) using many different techniques. For example, if we set $q(x) = 0$, and $g_0(t) = g_L(t) = 0$ then it is easy to show that

$$U(x,t) = \frac{2}{L} \sum_{k=1}^{\infty} e^{-\alpha k^2 \pi^2 t / L^2} \sin\left(\frac{k\pi}{L} x\right) \int_0^L U_0(x) \sin\left(\frac{k\pi}{L} x\right) dx, \tag{2}$$

where $\sin(k\pi x/L)$ are eigenfunctions of the eigenvalue problem (see [1, p. 48])

$$\frac{d^2 X(x)}{dx^2} + \beta X(x) = 0, \qquad X(0) = 0, \qquad X(L) = 0, \tag{3}$$

with eigenvalues $\beta_k = k^2 \pi^2 / L^2$.

- **Energy decay:** It is straightforward to show that in the case of no heat generation and zero Dirichlet boundary conditions the $L^2([0,L])$ norm of the solution to (1) i.e.,

$$\|U\|^2_{L^2([0,L])} = \int_0^L U(x,t)^2 dx \tag{4}$$

decays monotonically to zero as time increases. This can be seen directly from the analytical solution (2). Alternatively, we can derive an evolution equation for (4) and solve it. To this end, let us multiply the heat equation by $U(x,t)$ and integrate it over the spatial domain $[0,L]$. This yields

$$\int_0^L U(x,t) \frac{\partial U(x,t)}{\partial t} dx = \alpha \int_0^L U(x,t) \frac{\partial^2 U(x,t)}{\partial x^2} dx. \tag{5}$$

By integrating by parts and recalling (4) we obtain

$$\frac{d}{dt} \|U\|^2_{L^2([0,L])} = -2\alpha \underbrace{\int_0^L \left(\frac{\partial U}{\partial x}\right)^2 dx}_{\left\|\frac{\partial U}{\partial x}\right\|^2_{L^2([0,L])}} + 2\alpha \underbrace{\left[U \frac{\partial U}{\partial x}\right]_{x=0}^{x=L}}_{=0}. \tag{6}$$

At this point we use the Poincaré inequality[1]

$$\left\|\frac{\partial U}{\partial x}\right\|^2_{L^2([0,L])} \geq C \|U\|^2_{L^2([0,L])} \quad \Leftrightarrow \quad -\left\|\frac{\partial U}{\partial x}\right\|^2_{L^2([0,L])} \leq -C \|U\|^2_{L^2([0,L])} \tag{7}$$

[1]The Poincaré inequality holds for all differentiable functions u with zero boundary conditions.

to obtain

$$\frac{d}{dt}\left\|U\right\|^2_{L^2([0,L])} + 2\alpha \left\|U\right\|^2_{L^2([0,L])} \leq 0 \qquad \Rightarrow \qquad \left\|U\right\|^2_{L^2([0,L])} \leq \left\|U_0\right\|^2_{L^2([0,L])} e^{-2\alpha C t}. \tag{8}$$

Hence the "energy" of the solution, i.e., the L^2 norm (4) decays to zero as $t \to \infty$.

Finite-difference approximation

To solve the IBVP (1) with finite differences, let us consider the following an evenly-spaced grid in $[0, L]$, i.e.,

$$x_j = j\Delta x \qquad \Delta x = \frac{L}{N+1} \qquad j = 0, \ldots, N+1. \tag{9}$$

On this grid, we approximate the second derivative in (1) by using, e.g., the second-order finite difference formula

$$\left.\frac{\partial^2 U(x,t)}{\partial x^2}\right|_{x=x_j} \simeq \frac{U(x_{j-1},t) - 2U(x_j,t) + U(x_{j+1},t)}{\Delta x^2} \qquad j = 1, \ldots, N. \tag{10}$$

A substitution of (10) into (1) yields the so-called *semi-discrete* form[2]

$$\begin{cases} \dfrac{du_j}{dt} = \alpha \dfrac{u_{j-1}(t) - 2u_j(t) + u_{j+1}(t)}{\Delta x^2} + q(x_j) \qquad j = 1, \ldots, N \\ u_j(0) = U_0(x_j) \qquad j = 1, \ldots, N \\ u_0(t) = g_0(t) \\ u_{N+1}(t) = g_L(t) \end{cases} \tag{11}$$

where $u_j(t)$ represents an approximation of the exact solution $U(x_j, t)$, i.e., the exact solution evaluated at the grid point x_j. The system (11) can be written in a matrix-vector form as

$$\begin{cases} \dfrac{d\boldsymbol{u}}{dt} = \alpha \boldsymbol{D}^2_{\text{FD}} \boldsymbol{u} + \boldsymbol{h}(t) \\ \boldsymbol{u}(0) = \boldsymbol{U}_0 \end{cases} \tag{12}$$

where[3]

$$\boldsymbol{D}^2_{\text{FD}} = \frac{1}{\Delta x^2} \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & 1 & -2 & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 & -2 \end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}, \qquad \boldsymbol{h}(t) = \begin{bmatrix} q(x_1) + \alpha g_0(t)/\Delta x^2 \\ q(x_2) \\ q(x_3) \\ \vdots \\ \vdots \\ q(x_{N-1}) \\ q(x_N) + \alpha g_L(t)/\Delta x^2 \end{bmatrix}. \tag{13}$$

[2]The system (11) is called "semi-discrete" form of the IBVP (1) because we discretized only the dependence of the solution on the spatial variable x. If, in addition, we discretize (10) time using a time-stepping scheme then we obtain the so-called "fully discrete form" of the IBVP (1). The semi-discrete form (10) is also known as *method of lines* (MOL). The reason for such a definition is that the finite-difference solution of the heat equation is computed by solving a finite-dimensional system of ODEs, each one of which represents the dynamics of $U(x,t)$ at a particular grid point x_j. This corresponds to a "line" emanating from $U(x_j, 0)$.

[3]Recall that the differentiation matrix $\boldsymbol{D}^2_{\text{FD}}$ corresponding to the second-order finite difference discretization is tridiagonal and negative definite.

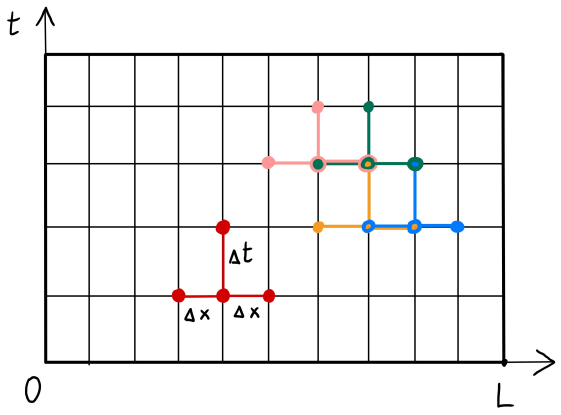

Figure 1: Finite-difference stencil corresponding to the forward-in-time centered-in-space discretization (17). We sketch the coupling of the system as we march forward in time by a few time steps.

In this way, we reduced the IBVP (1) to an *initial value problem* for a linear ODE, i.e., equation (12). Such an initial value problem can be solved using any time-stepping method we studied for initial value problems. For example, if we use the Euler forward scheme we obtain the *fully discrete form*

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \alpha \Delta t \boldsymbol{D}_{\text{FD}}^2 \boldsymbol{u}^k + \Delta t \boldsymbol{h}(t_k), \tag{14}$$

where

$$\boldsymbol{u}^k = \boldsymbol{u}(t_k). \tag{15}$$

On the other hand, if we use the two-step Adams-Bashforth method we obtain

$$\boldsymbol{u}^{k+2} = \boldsymbol{u}^{k+1} + \frac{\alpha \Delta t}{2} \left[3 \left(\boldsymbol{D}_{\text{FD}}^2 \boldsymbol{u}^{k+1} + \boldsymbol{h}^{k+1} \right) - \left(\boldsymbol{D}_{\text{FD}}^2 \boldsymbol{u}^k + \boldsymbol{h}^k \right) \right]. \tag{16}$$

Remark: Clearly, one could use higher-order finite-difference formulas to approximate the second-order derivative $\partial^2 U(x,t)/\partial x^2$. This yields other differentiation matrices, and requires some care when handling boundary conditions.

Local truncation error. The local truncation error (LTE) of a finite difference scheme is the residual arising when we ideally insert the exact solution to the problem into the fully discrete form. For illustration purposes let us compute the local truncation error of the so-called "centered in space forward in time" finite-difference scheme (see Figure 1)

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} = \alpha \frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{\Delta x^2}, \tag{17}$$

where u_k^j is an approximation of $U(x_j, t_k)$. By plugging in the *exact* solution $U(x,t)$ into (17) we obtain the LTE

$$\tau(x_j, t_k) = \frac{U(x_j, t_{k+1}) - U(x_j, t_k)}{\Delta t} - \alpha \frac{U(x_{j-1}, t_k) - 2U(x_j, t_k) + U(x_{j+1}, t_k)}{\Delta x^2}. \tag{18}$$

Let us define $U_j^k = U(x_j, t_k)$ and expand

$$U_j^{k+1} = U(x_j, t_k + \Delta t) \qquad \text{and} \qquad U_{j\pm 1}^k = U(x_j \pm \Delta x, t_k) \tag{19}$$

in Taylor series in Δx and Δt. This yields

$$\frac{U_j^{k+1}-U_j^k}{\Delta t}=\frac{1}{\Delta t}\left(\Delta t\frac{\partial U_j^k}{\partial t}+\frac{\Delta t^2}{2}\frac{\partial^2 U_j^k}{\partial t^2}+\cdots\right)=\frac{\partial U_j^k}{\partial t}+\frac{\Delta t}{2}\frac{\partial^2 U_j^k}{\partial t^2}+\cdots. \tag{20}$$

Similarly,

$$\begin{aligned}\frac{U_{j-1}^k-2U_j^k+U_{j+1}^k}{\Delta x^2}=&\frac{1}{\Delta x^2}\left(-\Delta x\frac{\partial U_j^k}{\partial x}+\frac{\Delta x^2}{2}\frac{\partial^2 U_j^k}{\partial x^2}-\frac{\Delta x^3}{6}\frac{\partial^3 U_j^k}{\partial x^3}+\frac{\Delta x^4}{24}\frac{\partial^4 U_j^k}{\partial x^4}+\cdots\right.\\&\left.\Delta x\frac{\partial U_j^k}{\partial x}+\frac{\Delta x^2}{2}\frac{\partial^2 U_j^k}{\partial x^2}+\frac{\Delta x^3}{6}\frac{\partial^3 U_j^k}{\partial x^3}+\frac{\Delta x^4}{24}\frac{\partial^4 U_j^k}{\partial x^4}+\cdots\right)\\=&\frac{\partial^2 U_j^k}{\partial x^2}+\frac{\Delta x^2}{12}\frac{\partial^4 U_j^k}{\partial x^4}+\cdots.\end{aligned} \tag{21}$$

Substituting (20)-(21) into (18), and using the PDE (1) yields

$$\begin{aligned}\tau(x_j,t_k)=&\underbrace{\frac{\partial U_j^k}{\partial t}-\alpha\frac{\partial^2 U_j^k}{\partial x^2}}_{=0}+\frac{\Delta t}{2}\frac{\partial^2 U_j^k}{\partial t^2}-\alpha\frac{\Delta x^2}{12}\frac{\partial^4 U_j^k}{\partial x^4}+\cdots\\=&\left(\alpha\frac{\Delta t}{2}-\frac{\Delta x^2}{12}\right)\alpha\frac{\partial^4 U_j^k}{\partial x^4}+\cdots\end{aligned} \tag{22}$$

where we replaced $\partial^2 U_j^k/\partial t^2$ with $\alpha^2\partial^4 U_j^k/\partial x^4$. This follows from the identity

$$\frac{\partial U}{\partial t}=\alpha\frac{\partial^2 U}{\partial x^2}\quad\Rightarrow\quad\frac{\partial U^2}{\partial t^2}=\alpha\frac{\partial^2}{\partial x^2}\left(\frac{\partial U}{\partial t}\right)=\alpha^2\frac{\partial^4 U}{\partial x^4}. \tag{23}$$

The local truncation error goes to zero linearly in Δt and quadratically in Δx. Therefore the "centered in space forward in time" scheme (17) is consistent with order one in Δt and with order two in Δx.

By following exactly the same steps it is possible to derive an expression for the local truncation error of finite-difference schemes involving different spatial and temporal discretizations. For example, we could have used a stencil with 5 points in space and the BDF3 method in time.

Absolute stability analysis. Consider the IBVP (1) with $q(x)=0$ and zero Dirichlet boundary conditions. The second-order finite-differences discretization of such problem is given by the system (12) with $\boldsymbol{h}(t)=\boldsymbol{0}$, i.e.,

$$\begin{cases}\dfrac{d\boldsymbol{u}}{dt}=\alpha\boldsymbol{D}_{\text{FD}}^2\boldsymbol{u}\\ \boldsymbol{u}(0)=\boldsymbol{U}_0\end{cases} \tag{24}$$

Recall that the matrix $\boldsymbol{D}_{\text{FD}}^2$ is negative definite with simple (real) eigenvalues

$$\lambda_k=\frac{2}{\Delta x^2}\left(\cos(k\pi\Delta x)-1\right)\qquad k=1,\ldots,N \tag{25}$$

Since $\lambda_k<0$ we have that the linear dynamical system (24) has a globally attracting stable node at the origin $\boldsymbol{u}=\boldsymbol{0}$. For small Δx (i.e., large number of spatial points) we obtain

$$\lambda_k\simeq\frac{2}{\Delta x^2}\left(1-\frac{1}{2}k^2\pi^2\Delta x^2+\cdots-1\right)\qquad k=1,\ldots,N \tag{26}$$

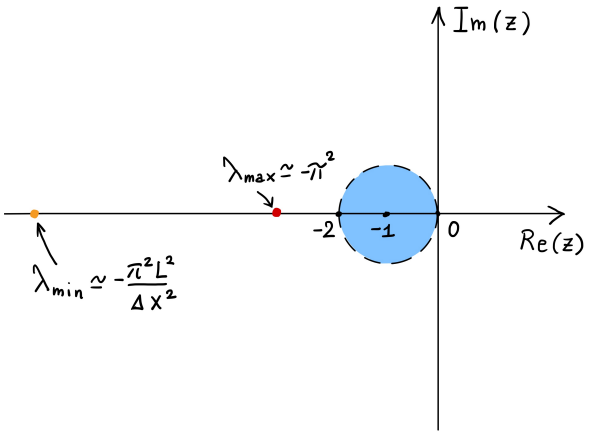

Figure 2: Absolute stability analysis of second-order finite-differences to solve the heat equation (1) with $q(x) = 0$ and zero Dirichlet boundary conditions. Shown are the smallest and largest eigenvalues of the second-order differentiation matrix defined in (13) and the region of absolute stability of the Euler forward method. The fully discrete form of the heat equation (30) is absolutely stable if and only if $\Delta t < 2\Delta x^2/(\alpha\pi^2 L^2)$.

Therefore, the smallest and largest eigenvalues of the matrix $\boldsymbol{D}^2_{\rm FD}$ for sufficiently small Δx are[4]

$$\lambda_{\rm min} = \lambda_L \simeq -\frac{\pi^2 L^2}{\Delta x^2}\left(\frac{N}{N+1}\right)^2 \simeq -\frac{\pi^2 L^2}{\Delta x^2}, \tag{28}$$

$$\lambda_{\rm max} = \lambda_1 = -\pi^2. \tag{29}$$

These equations show that as we increase the number of points in $[0, L]$ the system (24) becomes stiffer and stiffer, since there $\lambda_{\rm min} \to -\infty$ and $\lambda_{\rm max} \simeq -\pi^2$.

- **Euler forward time integration:** If we integrate the system (24) in time with the Euler Forward scheme we obtain the *fully discrete scheme*

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \alpha\Delta t \boldsymbol{D}^2_{\rm FD}\boldsymbol{u}^k, \tag{30}$$

where we denoted by $\boldsymbol{u}^k = \boldsymbol{u}(t_k)$. Clearly, the absolute stability condition for Euler forward is satisfied if (see Figure 2)

$$\lambda_{\rm min}\alpha\Delta t \geq -2 \qquad \text{i.e.} \qquad \Delta t \leq \frac{2\Delta x^2}{\alpha\pi^2 L^2} = \frac{2}{\alpha\pi^2(N+1)^2} \tag{31}$$

This result holds for a large number of points, i.e., for small Δx. For a small number of points we can still compute the smallest eigenvalue with (25) and use exactly the same reasoning. The condition

$$\Delta t \leq \frac{2}{\alpha\pi^2(N+1)^2} \tag{32}$$

clearly imposes *severe restrictions* on the largest time step we can use in (30). For instance, if $N = 2000$ and $\alpha = 1$ we have

$$\Delta t \leq 5.061 \times 10^{-8}. \tag{33}$$

[4]Recall that

$$\Delta x = \frac{L}{N+1}. \tag{27}$$

- **Three-step Adams-Bashforth time integration (AB3):** If we integrate the system (24) in time with the three-step Adams-Bashforth method we obtain the fully discrete scheme

$$\boldsymbol{u}^{k+3} = \boldsymbol{u}^{k+2} + \frac{\alpha\Delta t}{12}\boldsymbol{D}_{\text{FD}}^2\left(23\boldsymbol{u}^{k+2} - 16\boldsymbol{u}^{k+1} + 5\boldsymbol{u}^k\right). \tag{34}$$

  As we know, the region of absolute stability of AB3 intersects the real axis at $-6/11$. If the number of spatial points is large enough, then we obtain the absolute stability requirement

$$\Delta t \leq \frac{6\Delta x^2}{11\alpha\pi^2 L^2} = \frac{6}{11\alpha\pi^2(N+1)^2}, \tag{35}$$

  which is even more restrictive than the condition (32) we obtained for the Euler-forward time integrator.

- **Crank-Nicolson time integration:** If we discretize the system (24) in time using the Crank-Nicolson method or any other A-stable time stepping scheme then we do not have any time step restrictions. As is well-known the Crank-Nicolson method

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \frac{\alpha\Delta t}{2}\boldsymbol{D}_{\text{FD}}^2\left(\boldsymbol{u}^{k+1} + \boldsymbol{u}^k\right). \tag{36}$$

  can be conveniently written as

$$\left(\boldsymbol{I} - \frac{\alpha\Delta t}{2}\boldsymbol{D}_{\text{FD}}^2\right)\boldsymbol{u}^{k+1} = \left(\boldsymbol{I} + \frac{\alpha\Delta t}{2}\boldsymbol{D}_{\text{FD}}^2\right)\boldsymbol{u}^k. \tag{37}$$

  The matrix

$$\boldsymbol{K} = \boldsymbol{I} - \frac{\alpha\Delta t}{2}\boldsymbol{D}_{\text{FD}}^2 \tag{38}$$

  is symmetric and positive-definite[5]. Therefore we can perform a Cholesky decomposition $\boldsymbol{K} = \boldsymbol{R}^T\boldsymbol{R}$, where $\boldsymbol{R}$ upper-triangular, in a *pre-processing stage* and write the system (37) as

$$\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{u}^{k+1} = \left(\boldsymbol{I} + \frac{\alpha\Delta t}{2}\boldsymbol{D}_{\text{FD}}^2\right)\boldsymbol{u}^k. \tag{39}$$

  This system can be decomposed as a hierarchy of two triangular systems

$$\begin{cases} \boldsymbol{R}^T\boldsymbol{q}^{k+1} = \left(\boldsymbol{I} + \frac{\alpha\Delta t}{2}\boldsymbol{D}_{\text{FD}}^2\right)\boldsymbol{u}^k & \text{(lower triangular system)} \\ \boldsymbol{R}\boldsymbol{u}^{k+1} = \boldsymbol{q}^{k+1} & \text{(upper triangular system)} \end{cases} \tag{40}$$

  which can be solved by using forward/backward substitution at a cost of $O(N^2)$ operations.

Absolute stability analysis can be generalized to higher-order finite difference schemes and other time integrators, e.g., RK or BDF methods.

Remark: The time step restrictions imposed by absolute stability requirement in explicit methods (e.g., (33)) have nothing to do with accuracy. In fact, it can be shown that the local truncation error of the Crank Nicolson (CN) method (36) is second-order in time and second order in space. However, the CN method does not suffer from absolute stability requirements (it is unconditionally stable). Hence we are allowed to set any Δt we like. In particular, if we set $\Delta t = 10^{-4}$ we get a truncation error of the same order as Euler forward with $\Delta t = 10^{-8}$ for the same spatial grid. Solving the linear system (40) once is likely to be less expensive than performing 1000 time steps on a grid with $N = 2000$ spatial points.

[5]For second-order finite differences the matrix $\boldsymbol{K}$ is actually tridiagonal. This means that it can be inverted at a linear cost in N using Thomas' algorithm [3, p. 93].

Finite-difference methods for nonlinear PDEs. Consider the following initial-boundary value problem for a fourth-order nonlinear PDE (Kuramoto-Sivashinsky equation)

$$\begin{cases} \dfrac{\partial U}{\partial t} + U\dfrac{\partial U}{\partial x} + \dfrac{\partial^2 U}{\partial x^2} + \dfrac{\partial^4 U}{\partial x^4} = 0 \qquad t \geq 0 \qquad x \in [-L, L] \\ U(x,0) = U_0(x) \\ \text{Periodic B.C.} \end{cases} \tag{41}$$

The Kuramoto-Sivashinsky equation models the diffusive instabilities in a laminar flame front. Its solution can exhibit chaotic space-time dynamics. We discretize the IBVP with a second-order (in space) finite-difference method. To this end, we first approximate the derivatives $\partial U/\partial x$, $\partial^2 U/\partial x^2$ and $\partial^4 U/\partial x^4$ with fourth-order centered finite formulas on the grid

$$x_j = j\Delta x - L \qquad \Delta x = \frac{2L}{N} \qquad j = 0, \ldots, N. \tag{42}$$

Upon definition of $U_j(t) = U(x_j, t)$ such derivatives can be expressed as

$$\frac{\partial U(x_j,t)}{\partial x} \simeq \frac{U_{j+1}(t) - U_{j-1}(t)}{2\Delta x}, \tag{43}$$

$$\frac{\partial^2 U(x_j,t)}{\partial x^2} \simeq \frac{U_{j-1}(t) - 2U_j(t) + U_{j+1}(t)}{\Delta x^2}, \tag{44}$$

$$\frac{\partial^4 U(x_j,t)}{\partial x^4} \simeq \frac{U_{j-2}(t) - 4U_{j-1}(t) + 6U_j(t) - 4U_{j+1}(t) + U_{j+2}(t)}{\Delta x^4}. \tag{45}$$

A substitution of (43)-(45) into (41) yields the semi-discrete form

$$\frac{du_j}{dt} = -\underbrace{u_j \frac{u_{j+1} - u_{j-1}}{2\Delta x}}_{\text{nonlinear term}} - \frac{u_{j-1} - 2u_j + u_{j+1}}{\Delta x^2} - \frac{u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j+2}}{\Delta x^4}, \tag{46}$$

for $j = 0, \ldots, N-1$. Here $u_j(t)$ denotes the finite-difference approximation of the solution to (41). The system (46) is supplemented with the periodic conditions

$$u_{j+N}(t) = u_j(t) \quad \text{for all } j \tag{47}$$

and with the initial condition

$$u_j(0) = U_0(x_j) \quad \text{for all} \quad j = 0, \ldots, N-1. \tag{48}$$

Note that the second-order discretization (46) involves stencils with different number of points, i.e., three points for the first- and the second-order derivatives, and five points for the fourth-order derivative. The system (46) can be discretized in time with any time-stepping, e.g., with the AB2 method.

Remark: The stability of the fully discrete scheme may depend on the PDE being discretized and on the type of spatial and temporal discretization, in particular for hyperbolic IBVP problems.

Finite difference methods in two-dimensional spatial domains. Consider the following initial-boundary value problem for the two-dimensional heat equation

$$\begin{cases} \dfrac{\partial U}{\partial t} = \alpha\left(\dfrac{\partial^2 U}{\partial x^2} + \dfrac{\partial^2 U}{\partial y^2}\right) + q(x,y) \qquad t \geq 0 \qquad (x,y) \in \Omega \\ U(x,y,0) = U_0(x,y) \quad (x,y) \in \Omega \\ \text{Periodic B.C.} \end{cases} \tag{49}$$

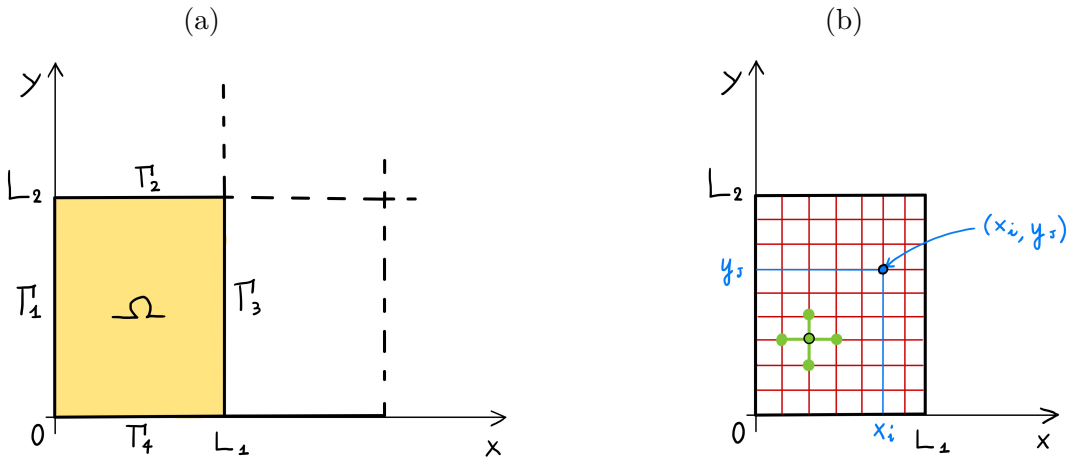

Figure 3: (a) Sketch of the spatial domain for the IBVP (49). The boundary of the domain Ω is the union between Γ_1, Γ_2, Γ_3 and Γ_4. The solution is assumed to be periodic in x and y. (b) Two-dimensional grid and stencil (green cross) used to approximate the Laplacian $\nabla^2 U = U_{xx} + U_{yy}$.
.

where Ω is a spatial domain defined as the Cartesian product of two intervals $[0, L_1]$ and $[0, L_2]$, i.e.,

$$\Omega = [0, L_1] \times [0, L_2]. \tag{50}$$

Periodic boundary conditions are set as

$$U(0, y) = U(L_1, y), \qquad \frac{\partial U(0, y)}{\partial x} = \frac{\partial U(L_1, y)}{\partial x}, \tag{51}$$

$$U(x, 0) = U(x, L_2), \qquad \frac{\partial U(x, 0)}{\partial y} = \frac{\partial U(x, L_2)}{\partial x}. \tag{52}$$

We discretize Ω in terms of the dimensional grid (see Figure 3(b))

$$(x_i, y_j) = \begin{cases} x_i = i\Delta x \qquad \Delta x = \dfrac{L_1}{N} \qquad i = 0, \ldots, N, \\ \\ y_j = j\Delta y \qquad \Delta y = \dfrac{L_2}{M} \qquad j = 0, \ldots, M. \end{cases} \tag{53}$$

By using second-order (in space) centered finite differences, we approximate the partial derivatives $\partial^2 U/\partial x^2$ and $\partial^2 U/\partial y^2$ at (x_i, y_j) as

$$\frac{\partial^2 U(x_i, y_j, t)}{\partial x^2} \simeq \frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{\Delta x^2} \tag{54}$$

$$\frac{\partial^2 U(x_i, y_j, t)}{\partial x^2} \simeq \frac{U_{i,j-1} - 2U_{i,j} + U_{i,j+1}}{\Delta y^2}, \tag{55}$$

where we denoted by $U_{i,j}(t) = U(x_i, y_j, t)$. A substitution of (54)-(55) into (49) yields

$$\frac{du_{i,j}(t)}{dt} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{\Delta x^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{\Delta y^2} + q(x_i, y_j), \tag{56}$$

with boundary conditions

$$u_{i+N,j}(t) = u_{i,j}(t), \qquad u_{i,j+M}(t) = u_{i,j}(t), \tag{57}$$

and initial condition

$$u_{i,j}(0) = U_0(x_i, y_j). \tag{58}$$

The system (56) can written in terms of differentiation matrices applied to the solution matrix $u_{i,j}(t)$. Alternatively, we can reshape the solution matrix into a column vector and construct appropriate differentiation matrices. The third option is to just write a function that takes in the matrix $u_{i,j}(t)$ and returns the right hand side of the system (56) at each time. This is usually the best option for practical implementation, especially for nonlinear system, or systems with space-dependent coefficients.

Galerkin and collocation methods

In this section we briefly review Galerkin and collocation methods for the one-dimensional diffusion equation with Dirichlet boundary conditions. To this end, consider the problem

$$\begin{cases} \dfrac{\partial U}{\partial t} = \alpha \dfrac{\partial^2 U}{\partial x^2} + q(x) \qquad t \geq 0 \qquad x \in [0, L] \\ U(x, 0) = U_0(x) \\ U(0, t) = g_0 \\ U(L, t) = g_L \end{cases} \tag{59}$$

Galerkin method. To solve the IBVP with the Galerkin method, we consider the function space

$$V = \{v \in L^2 \quad \text{such that} \quad \frac{\partial v}{\partial x} \in L^2, \quad v(0,t) = g_0 \quad \text{and} \quad v(L,t) = g_L\}. \tag{60}$$

where L^2 is the space of square integrable functions in $\Omega = [0, L]$. The function space V can be approximated by the finite-dimensional space

$$V_N = \text{span}\{\varphi_0, \ldots, \varphi_{N+1}\} \tag{61}$$

where $\varphi_k(x)$ can be, e.g., Lagrange characteristic polynomials associated with a set of Gauss-Lobatto nodes in $[0, L]$, e.g., Gauss-Lobatto-Legendre nodes. Alternatively, φ_0 and φ_{N+1} can be linear boundary modes, i.e.,

$$\varphi_0(x) = 1 - \frac{x}{L} \qquad \varphi_{N+1}(x) = \frac{x}{L} \tag{62}$$

while φ_k can be eigenfunctions of a Sturm-Liouville problem with zero boundary conditions, i.e.,

$$\varphi_k(x) = \sin\left(\frac{k\pi}{L} x\right), \qquad k = 1, 2, \ldots, N, \tag{63}$$

or shifted Chebyshev polynomials

$$\varphi_k(x) = x(L - x) T_{k-1}\left(\frac{2}{L} x - 1\right), \qquad k = 1, 2, \ldots, N. \tag{64}$$

In any case, a representation of the solution $U(x, t)$ in V_N takes the form

$$U_N(x, t) = \underbrace{g_0 \varphi_0(x) + g_L \varphi_{N+1}(x)}_{\text{boundary modes}} + \sum_{k=1}^{N} a_k(t) \varphi_k(x). \tag{65}$$

Substituting (65) into (1) and projecting the resulting equation onto $\varphi_j(x)$ ($j=1,\ldots,N$) yields

$$\sum_{k=1}^{N}\frac{da_k(t)}{dt}\int_0^L \varphi_j(x)\varphi_k(x)dx = \alpha g_0\int_0^L \frac{d^2\varphi_0(x)}{dx^2}\varphi_j(x)dx + \alpha g_L\int_0^L \frac{d^2\varphi_{N+1}(x)}{dx^2}\varphi_j(x)dx + \alpha\sum_{k=1}^{N}a_k(t)\int_0^L \frac{d^2\varphi_k(x)}{dx^2}\varphi_j(x)dx + \int_0^L q(x)\varphi_j(x)dx + \int_0^L R_N(x)\varphi_j(x)dx. \tag{66}$$

By integrating by parts the terms at the right hand side involving second derivatives and imposing that the residual $R_N(x)$ is orthogonal to the span of $\{\varphi_1,\ldots,\varphi_N\}$ (Galerkin method) we obtain

$$\sum_{k=1}^{N}M_{jk}\frac{da_k(t)}{dt} = -\alpha g_0 S_{0j} - \alpha g_L S_{N+1j} - \alpha\sum_{k=1}^{N}S_{jk}a_k(t) + \int_0^L q(x)\varphi_j(x)dx \qquad j=1,\ldots,N \tag{67}$$

where we defined

$$M_{jk} = \int_0^L \varphi_j(x)\varphi_k(x)dx \qquad \text{(mass matrix)}, \tag{68}$$

$$S_{jk} = \int_0^L \frac{d\varphi_j(x)}{dx}\frac{d\varphi_k(x)}{dx}dx \qquad \text{(stiffness matrix)}. \tag{69}$$

The system (67) can be written as

$$\boldsymbol{M}\frac{d\boldsymbol{a}(t)}{dt} = -\alpha\boldsymbol{S}\boldsymbol{a} + \boldsymbol{q}, \tag{70}$$

where

$$\boldsymbol{q} = \begin{bmatrix} \int_0^L q(x)\varphi_1(x)dx - \alpha\left(g_0 S_{01} + g_L S_{01}\right) \\ \vdots \\ \int_0^L q(x)\varphi_N(x)dx - \alpha\left(g_0 S_{0N} + g_L S_{N+1,N}\right) \end{bmatrix}. \tag{71}$$

If we use the interior modes (63) (basis functions) then the mass matrix and the stiffness matrix are both diagonal matrices. In particular,

$$M_{ij} = \frac{L}{2}\delta_{ij} \qquad \text{and} \qquad S_{ij} = \frac{\pi^2 j^2}{2L}\delta_{ij}. \tag{72}$$

This implies that the initial condition for the ODE (70) is

$$a_k(0) = \frac{1}{\|\varphi_k\|_{L^2}^2}\int_0^L U_0(x)\varphi_k(x)dx = \frac{2}{L}\int_0^L U_0(x)\varphi_k(x)dx \tag{73}$$

To study absolute stability of the Galerkin method, let us set $\boldsymbol{q}=\boldsymbol{0}$ in (70). In this way the solution certainly decays to zero. By using the matrices (72), we rewrite the system (70) as

$$\frac{da_k(t)}{dt} = -\frac{\alpha\pi^2 k^2}{L^2}a_k(t). \tag{74}$$

If we use the Euler forward time integration scheme is we obtain the absolute stability condition

$$\Delta t \leq -\frac{2}{\lambda_N} = \frac{2L^2}{\alpha\pi^2 N^2}. \tag{75}$$

This implies that as we add more and more modes the Galerkin system becomes stiffer and stiffer, which result is a smaller and smaller Δt if we use an explicit method.

Collocation method. In the Gauss-Legendre-Lobatto collocation method [2, p.132] we seek solutions to (59) in the form

$$U_N(x,t) = \sum_{k=0}^{N} U_N(x_j,t) l_j(x), \tag{76}$$

where $l_j(x)$ are the Lagrange characteristic polynomials corresponding to the Legendre-Gauss-Lobatto quadrature points. A substitution of (76) into (59) yields

$$\frac{\partial U_N}{\partial t} = \alpha \frac{\partial^2 U_N}{\partial x^2} + q(x) + R_N(x,t). \tag{77}$$

By requiring that the residual $R_N(x,t)$ vanish at the interior points yields the $N-1$ equations

$$\frac{dU_N(x_j,t)}{dt} = \alpha \sum_{k=0}^{N} D_{jk}^2 U_N(x_k,t) + q(x_j) \qquad j = 1,\dots,N-1. \tag{78}$$

Here, D_{ij}^2 is the second-order differentiation matrix corresponding to the Gauss-Legendre-Lobatto quadrature points (see [2, §5.4.1]). We close the system by using the boundary conditions

$$U_N(0,t) = g_0, \quad U_N(L,t) = g_L. \tag{79}$$

Of course, we can replace the Gauss-Legendre-Lobatto expansion with the Gauss-Chebyshev-Lobatto expansion described at the end of Chapter 7 of the course notes (see also [2, §5.4.2]). This yields easily computable collocation points and differentiation matrices.

References

[1] D. W. Hahn and M. N. Özisik. *Heat Conduction*. Wiley, third edition, 2012.

[2] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral methods for time-dependent problems*, volume 21 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007.

[3] A. Quarteroni, R. Sacco, and F. Salieri. *Numerical mathematics*. Springer, 2007.