Convergence analysis of finite difference methods for PDEs

Consider the following initial/boundary value problem for a system of linear PDEs

$$\begin{cases} \dfrac{\partial \boldsymbol{U}(\boldsymbol{x},t)}{\partial t} = \boldsymbol{L}(\boldsymbol{x},t)\boldsymbol{U}(\boldsymbol{x},t) + \boldsymbol{f}(\boldsymbol{x},t) \qquad t \geq 0 \qquad \boldsymbol{x} \in \Omega \\ \boldsymbol{S}\boldsymbol{U}(\boldsymbol{x},t) = 0 \qquad \boldsymbol{x} \in \partial\Omega \\ \boldsymbol{U}(\boldsymbol{x},0) = \boldsymbol{U}_0(\boldsymbol{x}) \end{cases} \tag{1}$$

Here $\boldsymbol{U}(\boldsymbol{x},t)$ denotes a vector field defined in a compact domain $\Omega \subseteq \mathbb{R}^d$, $\partial\Omega$ is the boundary of Ω, $\boldsymbol{L}$ is a linear operator that can depend on $\boldsymbol{x} = (x_1, \ldots, x_d)$ and t, and $\boldsymbol{S}$ is a (linear/affine) boundary operator enforcing Dirichlet, Neumann, Robin or mixed boundary conditions. We assume that the IBVP (1) is well-posed, i.e., that it admits a unique solution. Let us provide a few simple examples of PDEs that can be written in the form (1)

- **Liouville equation:** Consider a dynamical system

$$\begin{cases} \dfrac{d\boldsymbol{y}}{dt} = \boldsymbol{F}(\boldsymbol{y},t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases} \tag{2}$$

  evolving from a random initial state $\boldsymbol{y}_0$ with probability density function $p_0(\boldsymbol{y})$. The PDE governing the evolution equation of the joint probability density function of $\boldsymbol{y}(t)$ is

$$\frac{\partial p(\boldsymbol{y},t)}{\partial t} + \nabla \cdot [\boldsymbol{F}(\boldsymbol{y},t)p(\boldsymbol{y},t)] = 0. \tag{3}$$

  Clearly, this PDE can written in the form $\partial p/\partial t = L(\boldsymbol{x},t)p$, where

$$L(\boldsymbol{x},t)p = -\nabla \cdot \boldsymbol{F}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla p \tag{4}$$

  is a first-order differential operator that depends on the phase variables $\boldsymbol{y}$ as well as on time.

- **Wave equation:** Consider the wave equation

$$\frac{\partial^2 \psi(\boldsymbol{x},t)}{\partial t^2} = c^2 \nabla^2 \psi(\boldsymbol{x},t) \tag{5}$$

  and the equivalent system of two first-order PDEs as

$$\begin{cases} \dfrac{\partial \psi(\boldsymbol{x},t)}{\partial t} = \eta(\boldsymbol{x},t), \\ \dfrac{\partial \eta(\boldsymbol{x},t)}{\partial t} = c^2 \nabla^2 \psi(\boldsymbol{x},t). \end{cases} \tag{6}$$

  Clearly, the system (6) can be written in the form (1) as

$$\frac{\partial}{\partial t} \underbrace{\begin{bmatrix} \psi(\boldsymbol{x},t) \\ \eta(\boldsymbol{x},t) \end{bmatrix}}_{\boldsymbol{U}(\boldsymbol{x},t)} = \underbrace{\begin{bmatrix} 0 & 1 \\ c^2\nabla^2 & 0 \end{bmatrix}}_{\boldsymbol{L}(\boldsymbol{x},t)} \underbrace{\begin{bmatrix} \psi(\boldsymbol{x},t) \\ \eta(\boldsymbol{x},t) \end{bmatrix}}_{\boldsymbol{U}(\boldsymbol{x},t)} \tag{7}$$

  Note that the linear operator $\boldsymbol{L}(\boldsymbol{x},t)$ in this case does not depend on $\boldsymbol{x}$ and t.

Lax-Richtmyer stability theory. In this section we provide necessary and sufficient conditions for convergence of finite-difference schemes to approximate the solution of the IBVP (1). In the interest of simplicity we consider the case where the linear operator $\boldsymbol{L}(\boldsymbol{x},t)$ in (1) is time-independent, although all consideration in the present discussion apply as well when $\boldsymbol{L}$ is time-dependent. The fully discrete finite-difference form of the IBVP (1) with time-independent linear operator $\boldsymbol{L}$ can always be written in the form

$$\boldsymbol{u}^{k+1} = \boldsymbol{B}\boldsymbol{u}^k + \boldsymbol{b}^k, \tag{8}$$

where $\boldsymbol{u}^k$ is the vector representing the approximation of the solution $\boldsymbol{U}(x,t)$ at *all* grid points[1] and time t_k, or (more generally) a vector representing solution at all grid points and multiple time instants (see the AB2 method described below).

The matrix $\boldsymbol{B}$ usually depends on Δt, Δx_1, Δx_2, etc., and also on the spatial discretization of the functions and operators appearing in $\boldsymbol{L}(\boldsymbol{x})$, while $\boldsymbol{b}^k$ takes cares of external forcing terms and/or the boundary conditions. The vector $\boldsymbol{b}^k$ may also depend of Δt, Δx_1, Δx_2, etc.

Example: Consider the one-dimensional heat-equation

$$\frac{\partial U}{\partial t} = \alpha \frac{\partial^2 U}{\partial x^2}, \tag{9}$$

with Dirichlet boundary conditions. Discretize the second derivative $\partial^2 U/\partial x^2$ by, e.g., second-order centered finite differences. This yields the semi-discrete form

$$\frac{d\boldsymbol{u}(t)}{dt} = \alpha \boldsymbol{D}^2_{\text{FD}} \boldsymbol{u}. \tag{10}$$

We have seen that we can discretize (10) in time using many different schemes,e.g.,

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \alpha \Delta t \boldsymbol{D}^2_{\text{FD}} \boldsymbol{u}^k \qquad \text{(Euler forward)}, \tag{11}$$

$$\left(\boldsymbol{I} - \frac{\alpha \Delta t}{2} \boldsymbol{D}^2_{\text{FD}}\right) \boldsymbol{u}^{k+1} = \left(\boldsymbol{I} + \frac{\alpha \Delta t}{2} \boldsymbol{D}^2_{\text{FD}}\right) \boldsymbol{u}^k \qquad \text{(Crank-Nicolson)}. \tag{12}$$

These schemes can be written in the form (8) provided we define

$$\boldsymbol{B} = \boldsymbol{I} + \alpha \Delta t \boldsymbol{D}^2_{\text{FD}} \qquad \text{(Euler forward)}, \tag{13}$$

$$\boldsymbol{B} = \left(\boldsymbol{I} - \frac{\alpha \Delta t}{2} \boldsymbol{D}^2_{\text{FD}}\right)^{-1} \left(\boldsymbol{I} + \frac{\alpha \Delta t}{2} \boldsymbol{D}^2_{\text{FD}}\right) \qquad \text{(Crank-Nicolson)}. \tag{14}$$

Similarly, if we discretize (9) in time with the two-step Adams-Bashforth method we obtain

$$\boldsymbol{u}^{k+2} = \boldsymbol{u}^{k+1} + \alpha \Delta t \boldsymbol{D}^2_{\text{FD}} \left(\frac{3}{2}\boldsymbol{u}^{k+1} - \frac{1}{2}\boldsymbol{u}^k\right). \tag{15}$$

We can always write a two-step method as a one-step method in a higher-dimensional space. To this end, define

$$\boldsymbol{v}^{k+1} = \boldsymbol{u}^k \tag{16}$$

[1]The numerical solution $\boldsymbol{u}^k$ in (8) can be arranged as a vector, a matrix, or a multi-dimensional array. Correspondingly, $\boldsymbol{B}$ can be a matrix, a tensor or a more general linear operator in the space in which $\boldsymbol{u}^k$ is defined.

and rewrite (15) as

$$\begin{cases} \boldsymbol{u}^{k+2} = \boldsymbol{u}^{k+1} + \alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 \left(\frac{3}{2}\boldsymbol{u}^{k+1} - \frac{1}{2}\boldsymbol{v}^{k+1} \right) \\ \boldsymbol{v}^{k+2} = \boldsymbol{u}^{k+1} \end{cases} \tag{17}$$

i.e.,

$$\boldsymbol{z}^{k+2} = \boldsymbol{B}\boldsymbol{z}^{k+1} \tag{18}$$

where

$$\boldsymbol{z}^{k+2} = \begin{bmatrix} \boldsymbol{u}^{k+2} \\ \boldsymbol{v}^{k+2} \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{I} + \frac{3}{2}\alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 & -\frac{1}{2}\alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}. \tag{19}$$

Note that (18) is in the form (8). However, in this case the vector of unknowns $\boldsymbol{z}^k$ is not just the solution $\boldsymbol{u}^k$ at time t_k but rather a concatenation of the solution at time t_k and t_{k-1}.

Definition 1 (Lax-Richtmyer stability)**.** The finite difference scheme (8) is *stable* if there exists a constant C_T independent of k, Δt, Δx_1, Δx_2 etc such that

$$\left\| \boldsymbol{B}^k \right\| \leq C_T \quad \text{for all } k \text{ such that} \quad k\Delta t \leq T. \tag{20}$$

Here, $\|\cdot\|$ denotes any matrix norm induced by a vector norm. In other words, we require that the matrix powers $\boldsymbol{B}$ are uniformly bounded[2] by a constant C_T for all $k \leq T/\Delta t$, where the integration period T is fixed and is chosen arbitrarily Δt.

When studying convergence of (8) we are interested, in particular, in the behavior of $\left\|B^k\right\|$ when Δt and Δx_i are sent to zero.

Theorem 1 (Lax-Richtmyer equivalence theorem [1])**.** *Given a properly posed initial-boundary value problem* (1) *and a consistent*[3] *finite-difference approximation* (8)*, stability is a necessary and sufficient condition for convergence.*

Proof. For simplicity we consider the case where time integration is defined by a one-step scheme although all consideration in the present proof apply as well for multistep schemes. A substitution of the exact solution $\boldsymbol{U}(\boldsymbol{x}, t)$ of the IBVP (1) into the fully discrete scheme (8) yields the local truncation error (LTE) $\boldsymbol{\tau}^k$, defined by the equation

$$\boldsymbol{U}^{k+1} = \boldsymbol{B}\boldsymbol{U}^k + \boldsymbol{b}^k + \Delta t \boldsymbol{\tau}^k. \tag{22}$$

Here, we denoted by

$$\boldsymbol{U}^k = \begin{bmatrix} U(\boldsymbol{x}_1, t_k) \\ U(\boldsymbol{x}_2, t_k) \\ \vdots \\ U(\boldsymbol{x}_N, t_k) \end{bmatrix} \tag{23}$$

where N denotes the total number of spatial grid points. For example, in 3D we have

$$\boldsymbol{x}_i = (x_{l(i)}, y_{m(i)}, z_{n(i)}). \tag{24}$$

[2]Uniformly bounded means that the bound

$$\left\| \boldsymbol{B}\left(\Delta t, \Delta x_1, \Delta x_2, \cdots\right)^k \right\| \leq C_T \tag{21}$$

holds for every Δt, Δx_1, Δx_2, etc., and every $k \leq T\Delta t$ (fixed T, and any chosen Δt).

[3]Recall that a finite-difference approximation is said to be consistent if the local truncation error goes to zero as we send Δt and Δx_i $(i = 1, \ldots, d)$ to zero.

i.e, we have $N = n^3$ points, where n is the number of points in each variable z, y or z. Note that for multi-step time integration methods we just need to replace $\boldsymbol{U}^k$ by $\boldsymbol{Z}^k = [U^k, \quad U^{k-1}, \quad U^{k-2}, \ldots]^T$, i.e., a vector collecting the solution vector at times t_k, t_{k-1}, etc. (see, e.g., Eqs. (17)-(19)). Subtracting (8) from (22) yields

$$\boldsymbol{e}^{k+1} = \boldsymbol{B}\boldsymbol{e}^k + \Delta t \boldsymbol{\tau}^k, \tag{25}$$

where

$$\boldsymbol{e}^k = \boldsymbol{U}^k - \boldsymbol{u}^k \qquad \text{(error)} \tag{26}$$

The recursion (25) can be iterated back to the error $\boldsymbol{e}^0$. To this end,

$$\begin{aligned}
\boldsymbol{e}^k =& \boldsymbol{B}\boldsymbol{e}^{k-1} + \Delta t \boldsymbol{\tau}^{k-1} \\
=& \boldsymbol{B}\left(\boldsymbol{B}\boldsymbol{e}^{k-2} + \Delta t \boldsymbol{\tau}^{k-2}\right) + \Delta t \boldsymbol{\tau}^{k-1} \\
=& \boldsymbol{B}^2\boldsymbol{e}^{k-2} + \Delta t \boldsymbol{B}\boldsymbol{\tau}^{k-2} + \Delta t \boldsymbol{\tau}^{k-1} \\
& \vdots \\
=& \boldsymbol{B}^k\boldsymbol{e}^0 + \Delta t \sum_{j=1}^{k} \boldsymbol{B}^{k-j}\boldsymbol{\tau}^{j-1}.
\end{aligned} \tag{27}$$

At this point we take any vector norm of $\boldsymbol{e}^k$ (and corresponding induced matrix norm), and use the stability assumption (20) to obtain

$$\begin{aligned}
\left\|\boldsymbol{e}^k\right\| =& \left\| \boldsymbol{B}^k\boldsymbol{e}^0 + \Delta t \sum_{j=1}^{k} \boldsymbol{B}^{k-j}\boldsymbol{\tau}^{j-1} \right\| \\
\leq& \left\|\boldsymbol{B}^k\right\| \left\|\boldsymbol{e}^0\right\| + \Delta t \sum_{j=1}^{k} \left\|\boldsymbol{B}^{k-j}\right\| \left\|\boldsymbol{\tau}^{j-1}\right\| \\
\leq& C_T \left\|\boldsymbol{e}^0\right\| + k\Delta t C_T \max_{j=1,\ldots,k} \left\|\boldsymbol{\tau}^{j-1}\right\| \\
\leq& C_T \left\|\boldsymbol{e}^0\right\| + T C_T \max_{j=1,\ldots,k} \left\|\boldsymbol{\tau}^{j-1}\right\|,
\end{aligned} \tag{28}$$

where T is period of integration, and C_T is the the uniform bound in (20). The upper bound in (28) goes to zero if the method is consistent, i.e., if

$$\max_{j=1,\ldots,k} \left\|\boldsymbol{\tau}^{j-1}\right\| \to 0 \quad \text{for} \quad \Delta t, \Delta x_i \to 0, \tag{29}$$

and if the error at initial time $\left\|\boldsymbol{e}^0\right\|$ is either zero or goes to zero as we send Δt and Δx_1, Δx_2, etc., to zero. This proves that consistency plus Lax-Richtmyer stability implies convergence.

□

At this point a few remarks are in order.

- **Sufficient condition for stability:** Recall that for any matrix norm and any $k \in \mathbb{N}$

$$\left\|\boldsymbol{B}^k\right\| \leq \|\boldsymbol{B}\|^k. \tag{30}$$

  Therefore, to prove stability it is sufficient to show that for sufficiently small Δt

$$\|\boldsymbol{B}\| \leq 1 + \beta \Delta t \quad \text{for some } \beta \in \mathbb{R}. \tag{31}$$

In fact[4],

$$\left\|\boldsymbol{B}^k\right\| \leq \|\boldsymbol{B}\|^k \leq (1+\beta\Delta t)^k \leq e^{k\Delta t\beta} \leq e^{T\beta}. \tag{33}$$

- **Necessary and sufficient conditions for stability:** The spectral radius of a matrix is a lower bound for any matrix sub-multiplicative matrix norm. This implies that

$$\rho(\boldsymbol{B}) \leq \|\boldsymbol{B}\| \quad \Rightarrow \quad \rho(\boldsymbol{B})^k \leq \left\|\boldsymbol{B}^k\right\| \leq C_T \quad \Leftrightarrow \quad \rho(\boldsymbol{B})^k \leq C_T. \tag{34}$$

From this equation it follows that

$$\rho(\boldsymbol{B}) \leq 1+\beta\Delta t \tag{35}$$

is necessary for stability. If the matrix $\boldsymbol{B}$ is normal (i.e. $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{B}^T\boldsymbol{B}$) then the 2-norm coincides with the spectral radius, i.e.

$$\rho(\boldsymbol{B}) = \|\boldsymbol{B}\|_2 \tag{36}$$

and (35) is *necessary and sufficient for stability.*

Stability analysis of forward-in-time centered-in-space scheme for the heat equation: It is important to remark that the stability condition may depend on the way we send Δt and Δx_i to zero. To show this, consider the one-dimensional heat equation (9) with zero Dirichlet boundary conditions, and the matrix $\boldsymbol{B}$ corresponding to the centered-in-space forward-in-time finite-difference discretization (13). The matrix $\boldsymbol{B}$ is symmetric, and therefore the 2-norm coincides with the spectral radius. This yields[5],

$$\|\boldsymbol{B}\|_2 = \max_{i=1,\dots,N} |\alpha\Delta t\lambda_i + 1|. \tag{37}$$

At this point we recall that, for large N (number of spatial grid points), we have

$$\min_{i=1,\dots,N} \lambda_i = \lambda_N = \frac{2}{\Delta x^2}\left(\cos(N\pi\Delta x) - 1\right) \simeq -\frac{4}{\Delta x^2}. \tag{38}$$

Hence,

$$\|\boldsymbol{B}\|_2 \simeq \left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \tag{39}$$

Recalling equation (31), we conclude that a necessary and sufficient condition for stability is

$$\left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \leq 1+\beta\Delta t. \tag{40}$$

This equation defines a *stability region* in the $(\Delta t, \Delta x)$-plane for each β (see Figure 20). Such a stability region can be computed analytically, although the computation is a bit cumbersome (except for the case $\beta = 0$). In fact, define

$$\eta = \frac{\Delta t}{\Delta x^2}. \tag{41}$$

The inequality (40) can be split into the following two inequalities

$$\begin{cases} 4\alpha\eta - \beta\eta\Delta x^2 \leq 2 & \Leftrightarrow \quad \eta\left(4\alpha - \beta\Delta x^2\right) \leq 2 \quad \text{for} \quad \Delta x^2 \leq 4\alpha\Delta t, \\ -4\alpha\eta - \beta\eta\Delta x^2 \leq 0 & \Leftrightarrow \quad 4\alpha + \beta\Delta x^2 \geq 0 \quad\quad\quad\; \text{for} \quad \Delta x^2 \geq 4\alpha\Delta t. \end{cases} \tag{42}$$

[4]The inequalities (33) follow from the basic inequality

$$\log(1+x) \leq x \quad \Leftrightarrow \quad \log(1+x)^k = k\log(1+x) \leq kx \quad \Leftrightarrow \quad (1+x)^k \leq e^{kx} \tag{32}$$

[5]The eigenvalues of the matrix $\boldsymbol{B} = \boldsymbol{I} + \alpha\Delta t\boldsymbol{D}_{\text{FD}}^2$ are $1+\alpha\Delta t\lambda_i$, where λ_i are the eigenvalues of $\boldsymbol{D}_{\text{FD}}^2$.

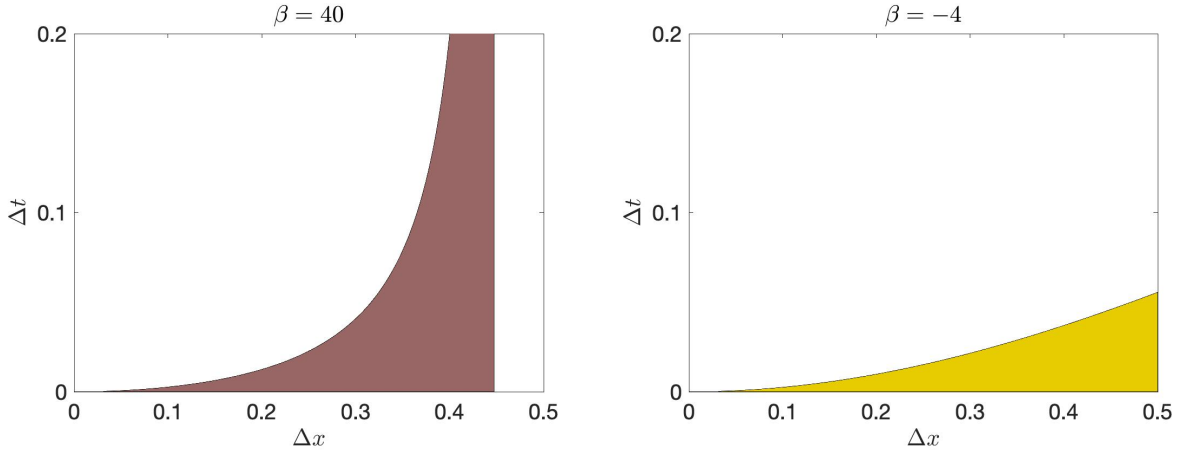

Figure 1: Lax-Richtmyer stability regions for the forward-in-time (Euler) centered-in-space (second-order) discretization of the heat equation (9) with $\alpha = 2$ and zero Dirichlet boundary conditions. The regions of stability are computed numerically using (40). Note the vertical asymptote for $\beta = 40$ at $\Delta x = 2\sqrt{\alpha/\beta} = 0.4472$ (see Eq. (43))

The first one can be written as

$$\frac{\Delta t}{\Delta x^2} \leq \frac{2}{4\alpha - \beta \Delta x^2}. \tag{43}$$

For $\beta > 0$ this yields the additional condition $\Delta x \leq 2\sqrt{\alpha/\beta}$ (see Figure 20). On the other hand, for $\beta = 0$ everything simplifies substantially. In particular, the second inequality in (42) yields the trivial condition $\alpha \geq 0$, while the first inequality yields

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2\alpha}. \tag{44}$$

The condition (44) also shows that Δt and Δx cannot be sent to zero at arbitrary rates. Indeed, we must have $\Delta t \sim \kappa \Delta x^2$ for (44) (or (43)) to hold in the limits $\Delta t, \Delta x \to 0$.

Lax-Richtmyer stability analysis applies to both implicit and explicit temporal integration schemes. However, for implicit schemes the matrix $\boldsymbol{B}$ involves the inverse of some other matrix (see, e.g., equation (14)). This makes the stability analysis of implicit schemes not straightforward nor practical using the matrix $\boldsymbol{B}$. We will see hereafter that this issue can be mitigated (at least for linear PDEs) by using discrete Fourier series.

Convergence analysis for nonlinear PDEs. The fully discrete finite-difference formulation of a one-dimensional nonlinear PDE can be written as

$$\sum_{k=0}^{q} \alpha_q u_j^{k+q} = \Delta t \Phi_j \left(\boldsymbol{u}^{k+q}, \ldots, \boldsymbol{u}^k, \Delta t, \Delta x \right). \tag{45}$$

For instance, the second-order central finite-difference discretization of the Kuramoto-Sivashinsky equation with Euler forward time stepping can be written as

$$u_j^{k+1} - u_j^k = \Delta t \left(-u_j^k \frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} - \frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{\Delta x^2} - \frac{u_{j-2}^k - 3u_{j-1}^k + 6u_j^k - 3u_{j+1}^k + u_{j+2}^k}{\Delta x^4} \right), \tag{46}$$

To study convergence of this scheme for Δt and Δx going to zero, we can use methods similar to the those we used in the convergence analysis of numerical schemes for ODEs, in particular the convergence proof

in the course note 4. To use such a proof in the context of finite-difference approximations of PDEs, we need to make sure that the Lipshitz constant of $\mathbf{\Phi}$ in (45) can be bounded by some constant when we send Δt and Δx to zero at an appropriate rate. Under this assumption, it is rather straightforward to show that method is convergent (provided the method it is consistent). To this end, just follow the proof in the appendix of the course note 4.

Von-Neumann stability theory. Stability analysis of finite-difference schemes can be simplified substantially if the PDE is defined in a periodic domain. The key idea is to use discrete Fourier series applied to the finite-difference discretization of the PDE and determine under which conditions on Δt, Δx_1, Δx_2, etc., the scheme is stable. One reason for the Fourier series analysis is that it allows us to determine stability conditions for both implicit and explicit schemes is a rather straightforward way. To illustrate the method, let us consider the prototype IBVP

$$\begin{cases} \dfrac{\partial U(x,t)}{\partial t} = \alpha \dfrac{\partial^2 U}{\partial x^2} \qquad t \geq 0 \qquad x \in [0, L] \\ U(x,0) = U_0(x) \\ \text{Periodic B.C.} \end{cases} \tag{47}$$

We have seen that (47) can be discretized with second-order finite-differences in space and Euler-forward time integration as

$$u_j^{k+1} = u_j^k + \frac{\alpha \Delta t}{\Delta x^2}\left(u_{j-1}^k - 2u_j^k + u_{j+1}^k\right) \qquad j = 0, \ldots, N-1, \tag{48}$$

where u_j^k is the approximation of $U(x_j, t_k)$, and

$$x_j = j\frac{L}{N} \qquad j = 0, \ldots, N. \tag{49}$$

The scheme (48) is supplemented with periodic boundary conditions

$$u_j^k = u_{j+N}^k \quad \text{for all } j \in \mathbb{Z}. \tag{50}$$

and the initial condition

$$u_j^0 = U_0(x_j) \qquad j = 0, \ldots, N-1. \tag{51}$$

Let us now expand the numerical solution u_j^k in a discrete Fourier series[6]

[6] As is well known, the solution to (47) can be expanded in a Fourier series as

$$U(x,t) = \sum_{k=-\infty}^{\infty} C_k(t) e^{2\pi i k x/L} \tag{52}$$

Evaluating $U(x,t)$ on the grid (49) yields the discrete Fourier series

$$U(x_j,t) = \sum_{k=-\infty}^{\infty} C_k(t) e^{2\pi i k x_j/L} = \sum_{k=-\infty}^{\infty} C_k(t) e^{2\pi i k j/N} \qquad j = 0, \ldots, N-1. \tag{53}$$

Moreover,

$$U(x_j,t) = \sum_{k=-\infty}^{\infty} C_k(t) e^{2ik j\pi/N} = \sum_{k=0}^{N-1} \sum_{p=-\infty}^{\infty} C_{k+pN}(t) e^{2\pi i (k+pN) j/N} = \sum_{k=0}^{N-1} e^{2\pi i k j/N} \underbrace{\sum_{p=-\infty}^{\infty} C_{k+pN}(t)}_{c_k(t)/N}. \tag{54}$$

This can be written as

$$U(x_j,t) = \frac{1}{N}\sum_{k=0}^{N-1} c_k(t) e^{2\pi i k j/N} = \frac{\Delta x}{L}\sum_{k=0}^{N-1} c_k(t) e^{ikj\xi}, \qquad \xi = \frac{2\pi \Delta x}{L}. \tag{55}$$

$$u_j^k = \frac{1}{N}\sum_{p=0}^{N-1} c_p^k e^{ipj\xi}, \qquad \text{where} \qquad \xi = \frac{2\pi\Delta x}{L}, \tag{56}$$

and substitute it into (48) to obtain

$$\begin{aligned}\sum_{p=0}^{N-1} c_p^{k+1} e^{ipj\xi} &= \sum_{p=0}^{N-1} c_p^k e^{ipj\xi}\left[1+\frac{\alpha\Delta t}{\Delta x^2}\left(e^{-ip\xi}-2+e^{ip\xi}\right)\right] \\ &= \sum_{p=0}^{N-1} c_p^k e^{ipj\xi}\left[1+\frac{\alpha\Delta t}{\Delta x^2}\left(2\cos\left(p\xi\right)-2\right)\right],\end{aligned} \tag{57}$$

i.e.,

$$c_p^{k+1} = \underbrace{c_p^k\left[1+\frac{2\alpha\Delta t}{\Delta x^2}\left(\cos\left(2\pi p\frac{\Delta x}{L}\right)-1\right)\right]}_{\text{amplification factor } G_p(\Delta t,\Delta x)}. \tag{58}$$

Upon definition of

$$\boldsymbol{c}^k = \begin{bmatrix} c_0^k \\ c_0^k \\ \vdots \\ c_{N-1}^k \end{bmatrix}, \qquad \boldsymbol{G}(\Delta t,\Delta x) = \begin{bmatrix} G_0(\Delta t,\Delta x) & 0 & \cdots & 0 \\ 0 & G_1(\Delta t,\Delta x) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & G_{N-1}(\Delta t,\Delta x) \end{bmatrix} \tag{59}$$

we can write (58) as

$$\boldsymbol{c}^{k+1} = \boldsymbol{G}(\Delta t,\Delta x)\boldsymbol{c}^k. \tag{60}$$

The matrix $\boldsymbol{G}(\Delta t,\Delta x)$ in (60) plays the same role in Fourier space as the matrix $\boldsymbol{B}$ in (8) does in physical space. In other words, for the scheme (48) to be stable we must have

$$\left\|\boldsymbol{G}^k\right\| \leq H_T \quad \text{for all } k \text{ such that } k\Delta t \leq T. \tag{61}$$

where H_T is a constant that does not depend on Δx or on Δt. In (61) $\|\cdot\|$ denotes any matrix norm induced by a vector norm.

Remark: Clearly, if we compute the inverse Fourier transform of (60) we obtain

$$\boldsymbol{u}^{k+1} = \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{-1}\boldsymbol{u}^k, \tag{62}$$

where $\boldsymbol{F}$ is the Fourier transform matrix such that

$$\boldsymbol{u}^k = \boldsymbol{F}\boldsymbol{c}^k, \qquad \boldsymbol{F} = \frac{1}{N}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{i\xi} & e^{2i\xi} & \cdots & e^{i(N-1)\xi} \\ 1 & e^{2i\xi} & e^{4i\xi} & \cdots & e^{2i(N-1)\xi} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & e^{(N-1)i\xi} & e^{2(N-1)\xi} & \cdots & e^{i(N-1)^2\xi}. \end{bmatrix} \tag{63}$$

A comparison between (62) and (8) shows that

$$\boldsymbol{B} = \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{-1}. \tag{64}$$

This equation justifies why stability can be equivalently studied in Fourier space by studying the norm of $\boldsymbol{G}^k$. In fact,

$$\left\|\boldsymbol{B}^k\right\| = \left\|\boldsymbol{F}\boldsymbol{G}^k\boldsymbol{F}^{-1}\right\|. \tag{65}$$

The Fourier transform matrix $\boldsymbol{F}$ plays no role in the stability properties of the scheme.

Necessary and sufficient conditions for Von-Neumann stability. Let us recall that the spectral radius of a matrix $\boldsymbol{G}$, i.e.,

$$\rho(\boldsymbol{G}) = \max_i |\lambda_i| \tag{66}$$

where λ_i are the eigenvalues of $\boldsymbol{G}$, is a lower bound for any (sub-multiplicative) matrix norm[7] of $\boldsymbol{G}$, i.e.,

$$\rho(\boldsymbol{G}) \leq \|\boldsymbol{G}\| \quad \text{for every sub-multiplicative matrix norm } \|\cdot\|. \tag{69}$$

Moreover, the spectral radius of the matrix power $\boldsymbol{G}^k$ is equal to $\rho(\boldsymbol{G})^k$ (recall that the eigenvalues of $\boldsymbol{G}^k$ are λ_i^k). By using the stability condition (61) we obtain

$$\rho(\boldsymbol{G})^k \leq \left\|\boldsymbol{G}^k\right\| \leq H_T \tag{70}$$

i.e.,

$$\rho(\boldsymbol{G})^k \leq H_T. \tag{71}$$

As before, this implies that for sufficiently small Δt the spectral radius of $\boldsymbol{G}$ must satisfy (see Eq. (31))

$$\rho(\boldsymbol{G}) \leq 1 + \gamma\Delta t \tag{72}$$

This is a *necessary condition*, not a sufficient condition. In fact, it is possible that $\rho(\boldsymbol{G})^k \leq H_T$ even though $\left\|\boldsymbol{G}^k\right\|$ grows unboundedly as we send Δt, Δx_1, Δx_2, etc., to zero. In other words,

$$\rho(\boldsymbol{G})^k \leq H_T \quad \text{does not imply} \quad \left\|\boldsymbol{G}^k\right\| \leq H_T. \tag{73}$$

However, if the matrix $\boldsymbol{G}$ is normal, i.e., if $\boldsymbol{G}\boldsymbol{G}^* = \boldsymbol{G}\boldsymbol{G}^*$ (where * denotes the conjugate transpose) then it is easy to show that Von-Neumann stability condition (72) is sufficient.

Lemma 1. The Von-Neumann stability condition (72) is sufficient if the amplification matrix $\boldsymbol{G}$ is normal.

Proof. The spectral radius of normal matrices is equal to the matrix 2-norm

$$\rho(\boldsymbol{G}) = \sqrt{\rho(\boldsymbol{G}\boldsymbol{G}^*)} = \|\boldsymbol{G}\|_2. \tag{74}$$

This allows us to write (70) as

$$\rho(\boldsymbol{G})^k = \left\|\boldsymbol{G}^k\right\|_2 \leq H_T. \tag{75}$$

Hence, for normal matrices $\boldsymbol{G}$ we have that (72) implies

$$\left\|\boldsymbol{G}^k\right\|_2 \leq H_T \tag{76}$$

[7] A sub-multiplicative matrix norm is a norm satisfying

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\| \tag{67}$$

for all matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. All matrix norms induces by vector norms are sub-multiplicative. To prove (69) it is sufficient to consider one eigenvalue λ_i of $\boldsymbol{G}$ and the corresponding eigenvector $\boldsymbol{v}$. Construct the matrix $\boldsymbol{V} = [\boldsymbol{v} \quad \cdots \quad \boldsymbol{v}]$, and note that

$$\|\boldsymbol{G}\boldsymbol{V}\| = |\lambda_i|\,\|\boldsymbol{V}\| \leq \|\boldsymbol{G}\|\,\|\boldsymbol{V}\| \quad \Rightarrow \quad \|\boldsymbol{G}\| \geq \max_i |\lambda_i| = \rho(\boldsymbol{G}). \tag{68}$$

the matrix

i.e., that the scheme is stable. Recall that stability in one norm imples stability in any other norm.

□

Fast computation of the amplification factors. The Fourier series of the solution of linear PDEs with constant coefficients can be always decoupled into a system of equations involving one Fourier mode at a time. Hence, to determine the amplification factors of the Fourier coefficients it is sufficient to consider only one wave number. In practice, we can simply substitute

$$u_j^k = c_p^k e^{ijp\xi} \quad \text{where} \quad \xi = \frac{2\pi\Delta x}{L} \tag{77}$$

into the numerical scheme and compute the amplification factors for the p-th mode. Let us show how to perform this calculation for second-order centered finite-difference discretization of the heat equation with Crank-Nicolson time-integration.

- **Von-Neumann stability analysis of the heat equation (Euler-forward time integration).** We have seen that the Fourier transform of finite-difference scheme (48) yields the diagonal matrix of amplification factors defined in (59). The diagonal entries of $\boldsymbol{G}$ are the eigenvalues of $\boldsymbol{G}$. Hence, the spectral radius of $\boldsymbol{G}$ is

$$\rho(\boldsymbol{G}) = \max_{p=0,\dots,N-1} \left| 1 + \frac{2\alpha\Delta t}{\Delta x^2}\left(\cos\left(2\pi p\frac{\Delta x}{L}\right) - 1\right)\right| \simeq \left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \qquad \text{(for large } N\text{)}. \tag{78}$$

  By using the Von-Neumann condition (72) we conclude that the scheme (48) is stable if and only if

$$\left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \leq 1 + \gamma\Delta t. \tag{79}$$

  This is exactly the same condition we obtained in (40) (see the discussion thereafter).

- **Von-Neumann stability analysis of the heat equation (Crank-Nicolson time integration).** Consider the fully discrete finite-difference scheme

$$u_j^{k+1} - \frac{\alpha\Delta t}{2\Delta x^2}\left(u_{j-1}^{k+1} - 2u_j^{k+1} + u_{j+1}^{k+1}\right) = u_j^k + \frac{\alpha\Delta t}{2\Delta x^2}\left(u_{j-1}^k - 2u_j^k + u_{j+1}^k\right). \tag{80}$$

  Substitute (79) into (80) to obtain

$$c_p^{k+1}\left[1 - \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right] = c_p^k\left[1 + \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right], \tag{81}$$

  i.e.,

$$c_p^{k+1} = \frac{1 + \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)}{1 - \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)} c_p^k. \tag{82}$$

  Again, the amplification matrix $\boldsymbol{G}$ is diagonal with spectral radius

$$\rho(\boldsymbol{G}) = \max_{p=0,\dots,N-1} \left|\frac{1 + \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)}{1 - \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)}\right|. \tag{83}$$

  At this point we notice that $\cos(p\xi) - 1 \leq 0$ for any p. This implies that

$$\left|1 + \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right| \leq \left|1 - \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right| \tag{84}$$

and

$$\rho(\boldsymbol{G}) \leq 1. \tag{85}$$

Recalling that the Von-Neumann stability condition (72) we conclude that the second-order centered finite-difference scheme with Crank-Nicolson time integration is unconditionally stable. Moreover, the scheme is consistent, and therefore convergent.

Clearly, by following the same steps that lead us to (31) we see that (61) yields the following *sufficient condition* for stability

$$\|\boldsymbol{G}\| \leq 1 + \delta\Delta t \quad \text{as} \quad \Delta t \to 0, \tag{86}$$

where $\|\boldsymbol{G}\|$ denotes any matrix norm compatible with a vector norm.

References

[1] P. D. Lax and R. D. Richtmyer. Survey of the stability of linear finite-difference equations. *Communications on Pure and Applied Mathematics*, 9:267–293, 1956.