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, or mixed boundary conditions. We assume that the IBVP (1) is well-posed, i.e., that the solution exists and is unique in a certain space of functions. 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 probability density function (PDF) 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.

- **Fokker-Planck equation:** The Fokker-Planck equation describes the evolution of the probability density function of the state vector solving the stochastic differential equation (SDE)

$$d\boldsymbol{y} = \boldsymbol{\mu}(\boldsymbol{y},t)dt + \sigma d\boldsymbol{W}(t). \tag{5}$$

  Here, $\boldsymbol{y}(t) \in \mathbb{R}^n$, $\boldsymbol{\mu}(\boldsymbol{y},t)$ is the n-dimensional drift, σ is a constant diffusion coefficient and $\boldsymbol{W}(t)$ is an n-dimensional Wiener random process. The Fokker-Planck equation that corresponds to (5) has the form

$$\frac{\partial p(\boldsymbol{y},t)}{\partial t} = -\nabla \cdot [\boldsymbol{\mu}(\boldsymbol{y},t)p(\boldsymbol{y},t)] + \frac{\sigma^2}{2}\nabla^2 p(\boldsymbol{y},t), \qquad p(\boldsymbol{y},0) = p_0(\boldsymbol{y}), \tag{6}$$

  where $p_0(\boldsymbol{y})$ is the PDF of the initial state $\boldsymbol{y}(0)$.

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

$$\frac{\partial^2 \psi(\boldsymbol{x},t)}{\partial t^2} = c^2 \nabla^2 \psi(\boldsymbol{x},t) \tag{7}$$

  and the equivalent system of two first-order PDEs as

$$\begin{cases} \dfrac{\partial \psi(\boldsymbol{x},t)}{\partial t} = \eta(\boldsymbol{x},t), \\ \dfrac{\partial \eta(\boldsymbol{x},t)}{\partial t} = c^2 \nabla^2 \psi(\boldsymbol{x},t). \end{cases} \tag{8}$$

Clearly, the system (8) can be written in the form (1) as

$$\frac{\partial}{\partial t}\underbrace{\begin{bmatrix}\psi(\boldsymbol{x},t)\\ \eta(\boldsymbol{x},t)\end{bmatrix}}_{\boldsymbol{U}(\boldsymbol{x},t)} = \underbrace{\begin{bmatrix}0 & 1\\ c^2\nabla^2 & 0\end{bmatrix}}_{\boldsymbol{L}(\boldsymbol{x},t)}\underbrace{\begin{bmatrix}\psi(\boldsymbol{x},t)\\ \eta(\boldsymbol{x},t)\end{bmatrix}}_{\boldsymbol{U}(\boldsymbol{x},t)} \tag{9}$$

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). For simplicity here 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

$$\boldsymbol{u}^{k+1} = \boldsymbol{B}\boldsymbol{u}^k + \boldsymbol{b}^k, \tag{10}$$

where $\boldsymbol{u}^k$ is the vector representing the approximation of the solution $\boldsymbol{U}(\boldsymbol{x},t)$ at all spatial grid points[1] and time t_k. More generally, $\boldsymbol{u}^k$ represents a vector representing the 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 functions and operators appearing in $\boldsymbol{L}(\boldsymbol{x})$. On the other hand, $\boldsymbol{b}^k$ takes cares of external forcing terms and/or boundary conditions. The vector $\boldsymbol{b}^k$ may also depend of Δt, Δx_1, Δx_2, etc. Hence, we shall write

$$\boldsymbol{B}(\Delta t, \Delta x_1, \Delta x_2, \ldots), \qquad \boldsymbol{b}^k(\Delta t, \Delta x_1, \Delta x_2, \ldots). \tag{11}$$

Example: Consider the one-dimensional heat-equation

$$\frac{\partial U}{\partial t} = \alpha\frac{\partial^2 U}{\partial x^2}, \tag{12}$$

with zero 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

$$\frac{d\boldsymbol{u}(t)}{dt} = \alpha\boldsymbol{D}^2_{\text{FD}}\boldsymbol{u}. \tag{13}$$

We have seen that we can discretize (13) in time using many different schemes,e.g.,

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

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

These schemes can be written in the form (10) provided we define

$$\boldsymbol{B} = \boldsymbol{I} + \alpha\Delta t\boldsymbol{D}^2_{\text{FD}} \qquad \text{(Euler forward)}, \tag{16}$$

$$\boldsymbol{B} = \left(\boldsymbol{I} - \frac{\alpha\Delta t}{2}\boldsymbol{D}^2_{\text{FD}}\right)^{-1}\left(\boldsymbol{I} + \frac{\alpha\Delta t}{2}\boldsymbol{D}^2_{\text{FD}}\right) \qquad \text{(Crank-Nicolson)}. \tag{17}$$

[1]The numerical solution $\boldsymbol{u}^k$ in (10) 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.

Similarly, if we discretize (12) in time with the two-step Adams-Bashforth method we obtain

$$\boldsymbol{u}^{k+2} = \boldsymbol{u}^{k+1} + \alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 \left(\frac{3}{2}\boldsymbol{u}^{k+1} - \frac{1}{2}\boldsymbol{u}^{k}\right). \tag{18}$$

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

$$\boldsymbol{v}^{k+1} = \boldsymbol{u}^{k} \tag{19}$$

and rewrite (18) as

$$\begin{cases} \boldsymbol{u}^{k+2} = \boldsymbol{u}^{k+1} + \alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 \left(\frac{3}{2}\boldsymbol{u}^{k+1} - \frac{1}{2}\boldsymbol{v}^{k+1}\right) \\ \boldsymbol{v}^{k+2} = \boldsymbol{u}^{k+1} \end{cases} \tag{20}$$

i.e.,

$$\boldsymbol{z}^{k+2} = \boldsymbol{B}\boldsymbol{z}^{k+1} \tag{21}$$

where

$$\boldsymbol{z}^{k+2} = \begin{bmatrix} \boldsymbol{u}^{k+2} \\ \boldsymbol{v}^{k+2} \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{I} + \frac{3}{2}\alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 & -\frac{1}{2}\alpha\Delta t \boldsymbol{D}_{\text{FD}}^2 \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}. \tag{22}$$

Note that (21) is in the form (10). 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 (10) is said to be stable if there exists a constant C_T independent of k, Δt, Δx_1, Δx_2, etc. such that

$$\left\|\boldsymbol{B}^k\right\| \leq C_T \quad \text{for all } k \text{ such that} \quad k\Delta t \leq T. \tag{23}$$

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. Note that Lax-Richtmyer stability requirement is very similar to the condition we proved for zero-stability of numerical methods for ODEs, where we bounded the the matrix powers of the matrix that propagates the error (see Proposition 1 in Appendix of the course note 4).

Remark: The definition of Lax-Richtmyer stability follows from the following considerations. Consider a perturbation of the fully discrete scheme (10) of the form

$$\boldsymbol{z}^{k+1} = \boldsymbol{B}\boldsymbol{z}^k + \boldsymbol{b}^k + \boldsymbol{\delta}^k, \qquad \boldsymbol{z}^0 = \boldsymbol{u}^0 + \boldsymbol{\delta}^0, \tag{25}$$

where $\boldsymbol{\delta}^k$ is assumed to be small, i.e., $\left\|\boldsymbol{\delta}^k\right\| \leq \epsilon$. Define the error

$$\boldsymbol{e}^k = \boldsymbol{z}^k - \boldsymbol{u}^k, \tag{26}$$

where $\boldsymbol{u}^k$ is obtained by iterating (10). It is straightforward to show (see the proof of Theorem 1) that

$$\boldsymbol{e}^k \leq \left\|\boldsymbol{B}^k\right\| \left\|\boldsymbol{e}^0\right\| + \Delta t \sum_{j=1}^{k} \left\|\boldsymbol{B}^{k-j}\right\| \left\|\boldsymbol{\delta}^{j-1}\right\|. \tag{27}$$

[2]Uniformly bounded means that the bound

$$\left\|\boldsymbol{B}\left(\Delta t, \Delta x_1, \Delta x_2, \cdots\right)^k\right\| \leq C_T \tag{24}$$

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

Hence, if the Lax-Richtmyer stability condition (23) is satisfied we have that

$$\left\|\boldsymbol{z}^k - \boldsymbol{u}^k\right\| \leq C_T \left\|\boldsymbol{\delta}^0\right\| + C_T T \max_{j=1,\ldots,k} \left\|\boldsymbol{\delta}^{j-1}\right\| \leq \epsilon C_T(1+T) \qquad \text{for all} \qquad k \leq \frac{T}{\Delta t}. \tag{28}$$

In this form, we see Lax-Richtmyer stability (23) can be equivalently formulated as follows: the difference between the solution of the perturbed scheme (25) and the unperturbed one (10) stays bounded for all $k \leq T/\Delta t$, and goes to zero as we send the perturbation amplitude to zero.

When studying convergence of the finite-difference scheme (10) we are interested, in particular, in the behavior of $\left\|\boldsymbol{B}^k\right\|$ when Δt and Δx_i are sent to zero.

Theorem 1 (Lax-Richtmyer equivalence theorem [1])**.** Given a well-posed initial-boundary value problem (1) and a consistent[3] finite-difference approximation (10), stability is a necessary and sufficient condition for convergence. In other words,

$$\boxed{(\text{consistency} + \text{stability}) \Leftrightarrow \text{convergence.}} \tag{29}$$

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 also for multistep schemes. A substitution of the exact solution $\boldsymbol{U}(\boldsymbol{x},t)$ of the IBVP (1) into the fully discrete scheme (10) yields the local truncation error (LTE) $\boldsymbol{\tau}^k$, defined by the equation

$$\boldsymbol{U}^{k+1} = \boldsymbol{B}\boldsymbol{U}^k + \boldsymbol{b}^k + \Delta t \boldsymbol{\tau}^k. \tag{30}$$

Here, we denoted by

$$\boldsymbol{U}^k = \begin{bmatrix} U(\boldsymbol{x}_1,t_k) \\ U(\boldsymbol{x}_2,t_k) \\ \vdots \\ U(\boldsymbol{x}_N,t_k) \end{bmatrix} \tag{31}$$

where N denotes the total number of spatial grid points[4]. As an example, consider the local truncation error for the forward-in-time centered-in-space finite difference discretization of the heat equation

$$\tau_j^k = \frac{U_j^{k+1} - U_j^k}{\Delta t} - \frac{U_{j+1}^k - 2U_j^k + U_{j-1}^k}{\Delta x^2} \quad \Rightarrow \quad U_j^{k+1} = U_j^k + \frac{\Delta t}{\Delta x^2}\left(U_{j+1}^k - 2U_j^k + U_{j-1}^k\right) + \Delta t \tau_j^k. \tag{33}$$

Subtracting (10) from (30) yields

$$\boldsymbol{e}^{k+1} = \boldsymbol{B}\boldsymbol{e}^k + \Delta t \boldsymbol{\tau}^k, \tag{34}$$

where

$$\boldsymbol{e}^k = \boldsymbol{U}^k - \boldsymbol{u}^k \qquad \text{(error)} \tag{35}$$

[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.

[4] For example, in 3D we have

$$\boldsymbol{x}_i = (x_{l(i)}, y_{m(i)}, z_{q(i)}). \tag{32}$$

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 $\boldsymbol{U}^k$ represents the vector of unknown at different times. In other words, everything we saying applies also to multi-step time integration schemes, provided we replace $\boldsymbol{U}^k$ with $\boldsymbol{Z}^k = [\boldsymbol{U}^k \quad \boldsymbol{U}^{k-1} \quad \boldsymbol{U}^{k-2} \quad \cdots]^T$, i.e., a vector collecting the solution vector at times t_k, t_{k-1}, etc. (see, Eqs. (20)-(22)).

The recursion (34) can be iterated back to the error $\boldsymbol{e}^0$. To this end,

$$\begin{aligned}\boldsymbol{e}^k =& \boldsymbol{B}\boldsymbol{e}^{k-1} + \Delta t\boldsymbol{\tau}^{k-1}\\ =& \boldsymbol{B}\left(\boldsymbol{B}\boldsymbol{e}^{k-2} + \Delta t\boldsymbol{\tau}^{k-2}\right) + \Delta t\boldsymbol{\tau}^{k-1}\\ =& \boldsymbol{B}^2\boldsymbol{e}^{k-2} + \Delta t\boldsymbol{B}\boldsymbol{\tau}^{k-2} + \Delta t\boldsymbol{\tau}^{k-1}\\ &\vdots\\ =& \boldsymbol{B}^k\boldsymbol{e}^0 + \Delta t\sum_{j=1}^{k}\boldsymbol{B}^{k-j}\boldsymbol{\tau}^{j-1}.\end{aligned} \tag{36}$$

At this point we take any vector norm of $\boldsymbol{e}^k$ (and corresponding induced matrix norm), and use the stability assumption (23) to obtain

$$\begin{aligned}\left\|\boldsymbol{e}^k\right\| =& \left\|\boldsymbol{B}^k\boldsymbol{e}^0 + \Delta t\sum_{j=1}^{k}\boldsymbol{B}^{k-j}\boldsymbol{\tau}^{j-1}\right\|\\ \leq& \left\|\boldsymbol{B}^k\right\|\left\|\boldsymbol{e}^0\right\| + \Delta t\sum_{j=1}^{k}\left\|\boldsymbol{B}^{k-j}\right\|\left\|\boldsymbol{\tau}^{j-1}\right\|\\ \leq& C_T\left\|\boldsymbol{e}^0\right\| + C_T\underbrace{k\Delta t}_{\leq T}\max_{j=1,\dots,k}\left\|\boldsymbol{\tau}^{j-1}\right\|\\ \leq& C_T\left\|\boldsymbol{e}^0\right\| + TC_T\max_{j=1,\dots,k}\left\|\boldsymbol{\tau}^{j-1}\right\|,\end{aligned} \tag{37}$$

where T is period of integration, and C_T is the the uniform bound in (23). The upper bound in (37) goes to zero if the method is consistent, i.e., if

$$\max_{j=1,\dots,k}\left\|\boldsymbol{\tau}^{j-1}\right\| \to 0 \quad \text{for} \quad \Delta t, \Delta x_1, \Delta x_2, \dots \to 0, \tag{38}$$

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.

□

Necessary and sufficient conditions for stability. Recall that for any matrix $\boldsymbol{B}$, any matrix norm $\|\cdot\|$, and any integer k we have

$$\left\|\boldsymbol{B}^k\right\| \leq \|\boldsymbol{B}\|^k. \tag{39}$$

Therefore, to prove stability of (10) it is sufficient to show that for sufficiently small Δt

$$\|\boldsymbol{B}\| \leq 1 + \beta\Delta t \quad \text{for some } \beta \in \mathbb{R}. \tag{40}$$

In fact,

$$\left\|\boldsymbol{B}^k\right\| \leq \|\boldsymbol{B}\|^k \leq (1+\beta\Delta t)^k \leq e^{k\Delta t\beta} \leq e^{T\beta}. \tag{41}$$

The last two inequalities follow from the basic inequality

$$\log(1+x) \leq x \quad \Leftrightarrow \quad \log\left[(1+x)^k\right] = k\log(1+x) \leq kx \quad \Leftrightarrow \quad (1+x)^k \leq e^{kx}. \tag{42}$$

Of course, if the matrix $\boldsymbol{B}$ is a contraction in some norm, i.e., if

$$\|\boldsymbol{B}\| \leq 1, \tag{43}$$

then the finite-difference scheme (10) is Lax-Richtmyer stable. Next, recall that the spectral radius of a matrix

$$\rho(\boldsymbol{B}) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}. \tag{44}$$

represents a lower bound for any natural matrix norm[5], i.e.,

$$\rho(\boldsymbol{B}) \leq \|\boldsymbol{B}\| \quad \Rightarrow \quad \rho(\boldsymbol{B})^k \leq \left\|\boldsymbol{B}^k\right\| \leq C_T \quad \Leftrightarrow \quad \rho(\boldsymbol{B})^k \leq C_T. \tag{46}$$

From this equation it follows that

$$\rho(\boldsymbol{B}) \leq 1 + \beta\Delta t \tag{47}$$

is necessary for stability. Hence, the modulus of the eigenvalues of the matrix $\boldsymbol{B}$ must always be smaller than $1 + \beta\Delta t$. 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.

$$\rho(\boldsymbol{B}) = \|\boldsymbol{B}\|_2 \tag{48}$$

and (47) is *necessary and sufficient for stability*.

Remark: The norm of the matrix $\boldsymbol{B}$ depends on the discretization scheme, on Δx_i ($i = 1, 2, \ldots$), and on Δt. Hence, condition (40) effectively sets a condition on all these variables for the scheme to be stable. Hereafter we provide some examples.

An example of Lax-Richtmyer stability analysis

Consider the one-dimensional heat equation (12) with zero Dirichlet boundary conditions, and the matrix $\boldsymbol{B}$ corresponding to the forward-in-time centered-in-space finite-difference discretization (16). The matrix $\boldsymbol{B}$ is symmetric, and therefore the 2-norm coincides with the spectral radius $\rho(\boldsymbol{B})$. This yields[6],

$$\|\boldsymbol{B}\|_2 = \max_{i=1,\ldots,N} |\alpha\Delta t\lambda_i + 1| \tag{49}$$

where λ_i are the the eigenvalues of the differentiation matrix $\boldsymbol{D}^2_{\text{FD}}$. At this point we recall that all the eigenvalues λ_i are negative, and that

$$\min_{i=1,\ldots,N} \lambda_i = \lambda_N = \frac{2}{\Delta x^2}\left(\cos(N\pi\Delta x) - 1\right) \geq -\frac{4}{\Delta x^2}. \tag{50}$$

Hence,

$$\|\boldsymbol{B}\|_2 \leq \left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right|. \tag{51}$$

Recalling equation (40), we conclude that a necessary and sufficient condition for stability is

$$\left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \leq 1 + \beta\Delta t. \tag{52}$$

This equation defines a *stability region* in the $(\Delta t, \Delta x)$-plane for each β (see Figure 23). Such a stability region can be computed analytically, although the computation is a bit cumbersome, except for the case $\beta = 0$. In the latter case we have

$$\left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \leq 1 \quad \Leftrightarrow \quad \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2\alpha} \qquad \text{(Lax-Richtmyer stability condition)}. \tag{53}$$

[5] A natural matrix norm is a matrix norm induced by vector norm as

$$\|\boldsymbol{B}\| = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}. \tag{45}$$

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

Condition (53) 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 (53) to hold in the limits $\Delta t, \Delta x \to 0$.

Lax-Richtmyer stability analysis can be applied 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 (17)). This makes the stability analysis of implicit schemes not straightforward nor practical using the matrix $\boldsymbol{B}$. We shall see hereafter that this issue can be mitigated (at least for linear PDEs) by using discrete Fourier series (for periodic boundary conditions), which are at the basis of the *Von-Neumann stability analysis*.

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

$$\sum_{k=0}^{q} \alpha_q u_j^{k+q} = \Delta t \Phi_j \left(\boldsymbol{u}^{k+q}, \ldots, \boldsymbol{u}^k, \Delta t, \Delta x \right). \tag{54}$$

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

$$u_j^{k+1} - u_j^k = \Delta t \left(-u_j^k \frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} - \frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{\Delta x^2} - \frac{u_{j-2}^k - 3u_{j-1}^k + 6u_j^k - 3u_{j+1}^k + u_{j+2}^k}{\Delta x^4} \right), \tag{55}$$

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 appendix of course note 4. To use such a proof in the context of finite-difference (or other) approximations of PDEs, we need to make sure that the Lipshitz constant of $\boldsymbol{\Phi}$ in (54) 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 the 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 spatial 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". The idea is to write the finite difference form of the PDE as a recurrence in Fourier space, i.e., as

$$\boldsymbol{c}^k = \boldsymbol{G}\boldsymbol{c}^{k-1} + \boldsymbol{v}^{k-1}, \tag{56}$$

where $\boldsymbol{c}^k$ are the Fourier coefficient of the finite-difference numerical solution, and $\boldsymbol{v}^{k-1}$ are Fourier coefficient of some external forcing term. With this form available, we can apply the same reasoning of Theorem 1 to conclude that the scheme is convergent if and only if it is stable and consistent. Stability here refers to the norm of the matrix $\boldsymbol{G}$ which defines the discrete Fourier transform of the finite-difference scheme. As we shall see hereafter there is a one-to-one correspondence between the matrix $\boldsymbol{G}$ and the matrix $\boldsymbol{B}$ in the previous section. In particular, $\boldsymbol{G}$ and $\boldsymbol{F}$ are unitarily equivalent via the Fourier tranformation matrix $\boldsymbol{F}$, i.e.,

$$\boldsymbol{G} = \boldsymbol{F}^{-1}\boldsymbol{B}\boldsymbol{F}. \tag{57}$$

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.

Example: To illustrate how the discrete Fourier transform method works, consider the prototype IBVP

$$\begin{cases} \dfrac{\partial U}{\partial t} = \alpha \dfrac{\partial^2 U}{\partial x^2} \qquad t \geq 0 \qquad x \in [0, L] \\ U(x,0) = U_0(x) \\ \text{Periodic B.C.} \end{cases} \tag{58}$$

We have seen that (58) can be discretized with second-order finite-differences in space and Euler-forward time integration as

$$u_j^{k+1} = u_j^k + \frac{\alpha \Delta t}{\Delta x^2} \left(u_{j-1}^k - 2u_j^k + u_{j+1}^k \right) \qquad j = 0, \ldots, N-1, \tag{59}$$

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

$$x_j = j \frac{L}{N} \qquad j = 0, \ldots, N. \tag{60}$$

The scheme (59) is supplemented with periodic boundary conditions

$$u_j^k = u_{j+N}^k \quad \text{for all } j \in \mathbb{Z}. \tag{61}$$

and the initial condition

$$u_j^0 = U_0(x_j) \qquad j = 0, \ldots, N-1. \tag{62}$$

Let us now expand the numerical solution u_j^k in a discrete Fourier series[7]

$$u_j^k = \frac{1}{N} \sum_{p=0}^{N-1} c_p^k e^{ipj\xi}, \qquad \text{where} \qquad \xi = \frac{2\pi \Delta x}{L} \quad \text{and} \quad c_p^k \in \mathbb{C}. \tag{67}$$

A substitution on (67) into (59) yields

$$\begin{aligned} \sum_{p=0}^{N-1} c_p^{k+1} e^{ipj\xi} &= \sum_{p=0}^{N-1} c_p^k e^{ipj\xi} \left[1 + \frac{\alpha \Delta t}{\Delta x^2} \left(e^{-ip\xi} - 2 + e^{ip\xi} \right) \right] \\ &= \sum_{p=0}^{N-1} c_p^k e^{ipj\xi} \left[1 + \frac{\alpha \Delta t}{\Delta x^2} \left(2\cos(p\xi) - 2 \right) \right], \end{aligned} \tag{68}$$

[7] As is well known, the solution to (58) can be expanded in a Fourier series as

$$U(x,t) = \sum_{k=-\infty}^{\infty} C_k(t) e^{2\pi i k x / L} \tag{63}$$

Evaluating $U(x,t)$ on the grid (60) yields the discrete Fourier series

$$U(x_j, t) = \sum_{k=-\infty}^{\infty} C_k(t) e^{2\pi i k x_j / L} = \sum_{k=-\infty}^{\infty} C_k(t) e^{2\pi i k j / N} \qquad j = 0, \ldots, N-1. \tag{64}$$

Moreover,

$$U(x_j, t) = \sum_{k=-\infty}^{\infty} C_k(t) e^{2ikj\pi/N} = \sum_{k=0}^{N-1} \sum_{p=-\infty}^{\infty} C_{k+pN}(t) e^{2\pi i (k+pN) j/N} = \sum_{k=0}^{N-1} e^{2\pi i k j/N} \underbrace{\sum_{p=-\infty}^{\infty} C_{k+pN}(t)}_{c_k(t)/N}. \tag{65}$$

This can be written as

$$U(x_j, t) = \frac{1}{N} \sum_{k=0}^{N-1} c_k(t) e^{ikj\xi}, \qquad \xi = \frac{2\pi \Delta x}{L} = \frac{2\pi}{N} \tag{66}$$

i.e.,

$$c_p^{k+1} = c_p^k \underbrace{\left[1 + \frac{2\alpha\Delta t}{\Delta x^2}\left(\cos\left(2\pi p\frac{\Delta x}{L}\right) - 1\right)\right]}_{\text{“amplification factor” } G_p(\Delta t, \Delta x)}. \tag{69}$$

Upon definition of

$$\boldsymbol{c}^k = \begin{bmatrix} c_0^k \\ c_0^k \\ \vdots \\ c_{N-1}^k \end{bmatrix}, \qquad \boldsymbol{G}(\Delta t, \Delta x) = \begin{bmatrix} G_0(\Delta t, \Delta x) & 0 & \cdots & 0 \\ 0 & G_1(\Delta t, \Delta x) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & G_{N-1}(\Delta t, \Delta x) \end{bmatrix} \tag{70}$$

we can write (69) as

$$\boldsymbol{c}^{k+1} = \boldsymbol{G}(\Delta t, \Delta x)\boldsymbol{c}^k. \tag{71}$$

Fron this equation we see that the matrix $\boldsymbol{G}(\Delta t, \Delta x)$ in (71) plays the same role in Fourier space as the matrix $\boldsymbol{B}$ in (10) does in physical space. Of course, any finite-difference scheme for an arbitrary linear PDE can be written in the form (71). This prompts the following definition.

Definition 2 (Von-Neumann stability)**.** The finite difference scheme is said to be stable if there exists a constant H_T independent of k, Δt, Δx_1, Δx_2, etc. such that

$$\left\|\boldsymbol{G}^k\right\| \leq H_T \quad \text{for all } k \text{ such that} \quad k\Delta t \leq T, \tag{72}$$

where $\boldsymbol{G}$ is the Von-Neumann matrix appearing in (71) or (56).

Equivalence between Von-Neumann and Lax-Richtmyer stability conditions. In this section we show that the definitions of Von-Neumann stability (Eq. (72)) and Lax-Richtmyer stability (Eq. (23)) are equivalent for PDEs with periodic boundary conditions.

To this end, we notice that if we compute the inverse Fourier transform of (71) we obtain

$$\boldsymbol{u}^{k+1} = \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{-1}\boldsymbol{u}^k, \tag{73}$$

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

$$\boldsymbol{u}^k = \frac{1}{\sqrt{N}}\boldsymbol{F}\boldsymbol{c}^k, \qquad \boldsymbol{F} = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{i\xi} & e^{2i\xi} & \cdots & e^{i(N-1)\xi} \\ 1 & e^{2i\xi} & e^{4i\xi} & \cdots & e^{2i(N-1)\xi} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & e^{(N-1)i\xi} & e^{2(N-1)\xi} & \cdots & e^{i(N-1)^2\xi}. \end{bmatrix} \tag{74}$$

A comparison between (73) and (10) shows that

$$\boldsymbol{B} = \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{-1}. \tag{75}$$

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

$$\left\|\boldsymbol{B}^k\right\| = \left\|\boldsymbol{F}\boldsymbol{G}^k\boldsymbol{F}^{-1}\right\|. \tag{76}$$

The Fourier transform matrix $\boldsymbol{F}$ plays no role in the stability properties of the scheme. In fact, the DFT matrix $\boldsymbol{F}$ is *unitary*, i.e.,

$$\boldsymbol{F}\boldsymbol{F}^* = \boldsymbol{F}^*\boldsymbol{F} = \boldsymbol{I} \qquad (\boldsymbol{F}^* \text{ is the conjugate transpose of } \boldsymbol{F}). \tag{77}$$

This can be verified directly using the definition (74). We know that the singular values and the eigenvalues of a matrix do not change under unitary transformations. This implies that $\boldsymbol{B}^k$ and $\boldsymbol{G}^k$ have exactly the same eigenvalues and the same singular values. Hence the matrices $\boldsymbol{B}^k$ and $\boldsymbol{G}^k$ have the same spectral radius and the same 2-norm

$$\left\|\boldsymbol{B}^k\right\|_2 = \left\|\boldsymbol{G}^k\right\|_2. \tag{78}$$

Example: The matrix $\boldsymbol{G}$ that corresponds to the finite difference scheme

Necessary and sufficient conditions for Von-Neumann stability

Let us recall that the spectral radius of a matrix $\boldsymbol{G}$, i.e.,

$$\rho(\boldsymbol{G}) = \max_i |\lambda_i| \tag{79}$$

where λ_i are the eigenvalues of $\boldsymbol{G}$, is a lower bound for any natural matrix norm[8] of $\boldsymbol{G}$, i.e.,

$$\rho(\boldsymbol{G}) \leq \|\boldsymbol{G}\| \quad \text{for every natural matrix norm } \|\cdot\|. \tag{83}$$

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 (72) we obtain

$$\rho(\boldsymbol{G})^k \leq \left\|\boldsymbol{G}^k\right\| \leq H_T \tag{84}$$

i.e.,

$$\rho(\boldsymbol{G})^k \leq H_T. \tag{85}$$

As before, this implies that for sufficiently small Δt the spectral radius of $\boldsymbol{G}$ must satisfy (see Eq. (40))

$$\rho(\boldsymbol{G}) \leq 1 + \gamma \Delta t \qquad \gamma \in \mathbb{R}. \tag{86}$$

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,

$$\rho(\boldsymbol{G})^k \leq H_T \quad \text{does not imply} \quad \left\|\boldsymbol{G}^k\right\| \leq H_T. \tag{87}$$

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 (86) is sufficient.

Lemma 1. The Von-Neumann stability condition (86) is sufficient if the amplification matrix $\boldsymbol{G}$ is normal.

[8] All natural matrix norms, i.e., matrix norms defined in terms of vector norms as

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \tag{80}$$

are sub-multiplicative. This means that they

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\| \tag{81}$$

for all matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. To prove (83) we leverage the sub-multiplicative property of natural matrix norm. To this end, 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

$$\|\boldsymbol{G}\boldsymbol{V}\| = |\lambda_i|\,\|\boldsymbol{V}\| \leq \|\boldsymbol{G}\|\,\|\boldsymbol{V}\| \quad \Rightarrow \quad \|\boldsymbol{G}\| \geq \max_i |\lambda_i| = \rho(\boldsymbol{G}). \tag{82}$$

the matrix

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

$$\rho(\boldsymbol{G}) = \sqrt{\rho(\boldsymbol{G}\boldsymbol{G}^*)} = \|\boldsymbol{G}\|_2 . \tag{88}$$

This allows us to write (84) as

$$\rho(\boldsymbol{G})^k = \left\|\boldsymbol{G}^k\right\|_2 \leq H_T. \tag{89}$$

Hence, for normal matrices $\boldsymbol{G}$ we have that (86) implies

$$\left\|\boldsymbol{G}^k\right\|_2 \leq H_T \tag{90}$$

i.e., that the scheme is stable. □

For normal matrices $\boldsymbol{G}$ a simplified condition that grants us stability of the scheme is

$$\rho(\boldsymbol{G}) \leq 1. \tag{91}$$

Examples of Von-Neumann stability analysis

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 mode. In practice, we can simply substitute

$$u_j^k = c_p^k e^{ijp\xi} \quad \text{where} \quad \xi = \frac{2\pi\Delta x}{L} \tag{92}$$

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.

Heat equation with Euler-forward time integration. We have seen that the Fourier transform of finite-difference scheme (59) yields the diagonal matrix of amplification factors defined in (70). The diagonal entries of $\boldsymbol{G}$ are the eigenvalues of $\boldsymbol{G}$. Hence, the spectral radius of $\boldsymbol{G}$ is

$$\rho(\boldsymbol{G}) = \max_{p=0,\ldots,N-1} \left|1 + \frac{2\alpha\Delta t}{\Delta x^2}\left(\cos\left(2\pi p\frac{\Delta x}{L}\right) - 1\right)\right| \leq \left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right|. \tag{93}$$

By using the Von-Neumann condition (86) we conclude that the scheme (59) is stable if and only if

$$\left|\frac{4\alpha\Delta t}{\Delta x^2} - 1\right| \leq 1 + \gamma\Delta t. \tag{94}$$

This is exactly the same condition we obtained in (52) (see the discussion thereafter).

Heat equation with Crank-Nicolson time integration. Consider the fully discrete finite-difference scheme

$$u_j^{k+1} - \frac{\alpha\Delta t}{2\Delta x^2}\left(u_{j-1}^{k+1} - 2u_j^{k+1} + u_{j+1}^{k+1}\right) = u_j^k + \frac{\alpha\Delta t}{2\Delta x^2}\left(u_{j-1}^k - 2u_j^k + u_{j+1}^k\right). \tag{95}$$

Substitute (92) into (95) to obtain

$$c_p^{k+1}\left[1 - \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right] = c_p^k\left[1 + \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right], \tag{96}$$

i.e.,

$$c_p^{k+1} = \frac{1 + \dfrac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)}{1 - \dfrac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)} c_p^k. \tag{97}$$

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

$$\rho(\boldsymbol{G}) = \max_{p=0,\ldots,N-1} \left| \frac{1 + \dfrac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)}{1 - \dfrac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right)} \right|. \tag{98}$$

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

$$\left| 1 + \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right) \right| \leq \left| 1 - \frac{\alpha\Delta t}{\Delta x^2}\left(\cos(p\xi) - 1\right) \right| \tag{99}$$

and

$$\rho(\boldsymbol{G}) \leq 1. \tag{100}$$

Recalling the Von-Neumann stability condition (86) 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.

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.