Finite-difference methods for the advection equation

In this course note we study stability and convergence of various finite-difference schemes for simple hyperbolic PDEs (conservation laws) of the form

$$\frac{\partial U(x,t)}{\partial t} + \frac{\partial \left(F(U(x,t))\right)}{\partial x} = 0, \tag{1}$$

where F is a continuously differentiable nonlinear function. The material in this note is discussed in [3, Ch. 10]. More generally numerical methods nonlinear conservation laws, or systems of nonlinear conservation laws[1] are discussed in, e.g., in [1, 2]. Let us begin with the simple prototype linear initial/boundary value problem[2]

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

As is well-known, this PDE can be solved with the *method of characteristics*, by essentially transforming it into an ODE along the flow generated by the dynamical system (see Appendix at the end of this note)

$$\frac{dx(t)}{dt} = a \qquad x(0) = x_0. \tag{3}$$

In the case of (2) the ODE is $dz/dt = 0$, with initial condition $z(0) = U_0(x_0)$. This yields the analytical solution[3]

$$U(x,t) = U_0(x - at). \tag{4}$$

This is traveling wave moving with velocity a. If a is positive the wave moves to the right, while preserving entirely its structure. Once the wave reaches the periodic boundary, it comes back from the other side.

Finite-difference discretization. We discretize the IBVP (2) with second-order centered finite-differences. To this end, consider the following grid

$$x_j = j\Delta x, \qquad \Delta x = \frac{L}{N}, \qquad j = 0, \ldots, N \tag{5}$$

and approximate the first derivative $\partial U/\partial x$ as

$$\frac{\partial U(x_j,t)}{\partial x} \simeq \frac{U(x_{j+1},t) - U(x_{j-1},t)}{2\Delta x}. \tag{6}$$

A substitution of (6) into (2) yields the semi-discrete form

$$\frac{du_j(t)}{dt} = -a\frac{u_{j+1}(t) - u_{j-1}(t)}{2\Delta x} \qquad j = 0, \ldots, N-1 \tag{7}$$

with periodic boundary conditions

$$u_N(t) = u_0(t) \qquad u_{-1}(t) = u_{N-1}(t). \tag{8}$$

[1] A conservation law is an expression in mathematical terms of the balance within a physical system. It is a statement that the production of a physical quantity such as mass, energy or charge in a closed volume is exactly equal to the flux of that quantity across the boundary of that volume. Such conservation laws often take the form of partial differential equations with appropriate boundary conditions or equivalent integral forms.

[2] The IBVP is ill-posed if $a > 0$ and we set the boundary Dirichlet boundary condition $U(L,t) = g(t)$ where g is a continuous time-dependent function.

[3] To compute the solution of (2) we can of course also use other techniques such as Fourier series and Laplace transforms.

The system (7)-(8) can be written in a matrix-vector form as

$$\begin{cases} \dfrac{d\boldsymbol{u}}{dt} = -a\boldsymbol{D}^1_{\text{FD}}\boldsymbol{u} \\ \boldsymbol{u}(0) = \boldsymbol{U}_0 \end{cases} \tag{9}$$

where

$$\boldsymbol{D}^1_{\text{FD}} = \frac{1}{2\Delta x}\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & \cdots & -1 \\ -1 & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 0 & 1 & \cdots & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & -1 & 0 & 1 \\ 1 & \cdots & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix} u_0 \\ u_2 \\ u_3 \\ \vdots \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{bmatrix}, \tag{10}$$

The matrix $\boldsymbol{D}^1_{\text{FD}}$ is clearly skew-symmetric and therefore it has purely imaginary eigenvalues. It can be shown that the eigenvalues of $\boldsymbol{D}^1_{\text{FD}}$ are

$$\lambda_k = \frac{i}{\Delta x}\sin\left(\frac{2\pi}{L}k\Delta x\right) \qquad k = 1, \ldots, N. \tag{11}$$

Recall also that skew-symmetric matrices are normal. This implies that the spectral radius of the matrix $\boldsymbol{D}^1_{\text{FD}}$ coincides with its 2-norm, i.e., we have

$$\left\|\boldsymbol{D}^1_{\text{FD}}\right\|_2 = \rho\left(\boldsymbol{D}^1_{\text{FD}}\right). \tag{12}$$

Euler-forward time integration. Let us discretize the ODE system (7) in using the Euler-forward method. This yields the fully discrete scheme

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k - a\Delta t\boldsymbol{D}^1_{\text{FD}}\boldsymbol{u}^k, \tag{13}$$

where $\boldsymbol{u}^k$ denotes the approximation of the solution of (7) at time t_k. It is straightforward to show that the local truncation error (LTE) of (13) goes to zero linearly in Δt and quadratically in Δx. To this end, let us first write (13) component-by-component as

$$u^{k+1}_j = u^k_j - a\Delta t\frac{u^k_{j+1} - u^k_{j-1}}{2\Delta x}. \tag{14}$$

A substitution of the exact solution $U(x,t)$ of (2) into (13) gives the LTE

$$\tau^k_j = \frac{U^{k+1}_j - U^k_j}{\Delta t} + a\frac{U^k_{j+1} - U^k_{j-1}}{2\Delta x}, \tag{15}$$

where we denoted by U^k_j the exact solution evaluated at x_j and t_k, i.e., $U^k_j = U(x_j, t_k)$. Using Taylor series expansions yields

$$\tau^k_j = \frac{\Delta t}{2}\frac{d^2U^h_j}{dt^2} + \frac{a\Delta x^2}{12}\frac{d^3U^h_j}{dx^3} + \text{higher order terms}. \tag{16}$$

Hence, the method is consistent with order one in time and order two in space. Regarding stability, let us write the scheme (13) as

$$\boldsymbol{u}^{k+1} = \boldsymbol{B}\boldsymbol{u}^k \tag{17}$$

where

$$\boldsymbol{B} = \boldsymbol{I} - a\Delta t \boldsymbol{D}^1_{\text{FD}} \tag{18}$$

and $\boldsymbol{D}^1_{\text{FD}}$ is defined in (10). Recall that for normal matrices $\boldsymbol{B}$, a necessary and sufficient condition for Lax-Richtmyer stability[4] is

$$\rho(\boldsymbol{B}) \leq 1 + \beta\Delta t. \tag{20}$$

The spectral radius of the matrix $\boldsymbol{B}$ is easily obtained by shifting and rescaling the eigenvalues of $\boldsymbol{D}^1_{\text{FD}}$, i.e.,

$$\begin{aligned}
\rho(\boldsymbol{B}) &= \max_{p=1,\ldots,N} \left| 1 - i\frac{a\Delta t}{\Delta x} \sin\left(\frac{2\pi}{L} p\Delta x\right)\right| \\
&= \max_{p=1,\ldots,N} \sqrt{1 + \frac{a^2\Delta t^2}{\Delta x^2} \sin\left(\frac{2\pi}{L} p\Delta x\right)^2} \qquad (21)\\
&\leq \sqrt{1 + \frac{a^2\Delta t^2}{\Delta x^2}}. \qquad (22)
\end{aligned}$$

Taking the k-th power yields

$$\left\|\boldsymbol{B}^k\right\| = \rho(\boldsymbol{B})^k \leq \left(1 + \frac{a^2\Delta t^2}{\Delta x^2}\right)^{k/2} \leq \exp\left(a^2 \frac{k\Delta t^2}{2\Delta x^2}\right) \leq \exp\left(a^2 \frac{T\Delta t}{2\Delta x^2}\right). \tag{23}$$

If we choose Δt and Δx such that

$$\frac{\Delta t}{\Delta x^2} \leq b \tag{24}$$

for arbitrary (finite) b, then we see that the scheme (13) is Lax-Richtmyer stable. In fact, substituting (24) into (23) yields

$$\left\|\boldsymbol{B}^k\right\|_2 \leq \exp\left(a^2 \frac{Tb}{2}\right) \quad \text{for all } k \text{ such that } k\Delta t \leq T. \tag{25}$$

By using the Lax equivalence theorem we conclude that the method is convergent, since it is consistent and stable (under the condition (24))

The stability analysis that lead us to (23) is based on the knowledge of the spectral radius of the matrix $\boldsymbol{B}$ which, in turn, is based on the knowledge of the eigenvalues of $\boldsymbol{D}^1_{\text{FD}}$. A more direct method to obtain a stability inequality is based on Von-Neumann analysis. To this end, we study the dynamics of an arbitrary Fourier mode[5]

$$\widehat{u}^k_p = c^k_p e^{ijp\xi} \quad \text{where} \quad \xi = \frac{2\pi\Delta x}{L}. \tag{26}$$

Substituting (26) into (14) yields

$$\begin{aligned}
c^k_p &= c^k_p \left[1 - \frac{a\Delta t}{2\Delta x}\left(e^{ip\xi} - e^{-ip\xi}\right)\right] \\
&= \underbrace{c^k_p \left[1 - i\frac{a\Delta t}{\Delta x}\sin(\xi p)\right]}_{G_p(\Delta t,\Delta x)}. \qquad (27)
\end{aligned}$$

[4]The 2-norm of a normal matrix $\boldsymbol{B}$ coincides with the spectral radius $\rho(\boldsymbol{B})$, i.e.,

$$\rho(\boldsymbol{B}) = \|\boldsymbol{B}\|_2. \tag{19}$$

[5]Since the PDE (2) is linear with constant coefficients it is sufficient to consider one Fourier mode to perform stability analysis.

The amplification matrix $\boldsymbol{G}$ is diagonal

$$\boldsymbol{G}(\Delta t,\Delta x)=\begin{bmatrix} G_0(\Delta t,\Delta x) & 0 & \cdots & 0 \\ 0 & G_1(\Delta t,\Delta x) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & G_{N-1}(\Delta t,\Delta x) \end{bmatrix} \tag{28}$$

Since $\boldsymbol{G}$ is normal, we have that the following Von-Neumann condition

$$\|\boldsymbol{G}\|_2=\rho(\boldsymbol{G})\leq 1+\gamma\Delta t \tag{29}$$

is necessary and sufficient for stability. The spectral radius of $\boldsymbol{G}$ is the same as the spectral radius of the matrix $\boldsymbol{B}$, i.e.,

$$\begin{aligned}\rho(\boldsymbol{G})&=\max_{p=0,\ldots,N-1}|G_p(\Delta t,\Delta x)|\\ &=\max_{p=0,\ldots,N-1}\left|1-i\frac{a\Delta t}{\Delta x}\sin\left(\frac{2\pi}{L}p\Delta x\right)\right|\\ &=\max_{p=0,\ldots,N-1}\sqrt{1+\frac{a^2\Delta t^2}{\Delta x^2}\sin\left(\frac{2\pi}{L}p\Delta x\right)^2}.\end{aligned} \tag{30}$$

As before (see Eq. (23)),

$$\rho(\boldsymbol{G})^k\leq\left(1+\frac{a^2\Delta t^2}{\Delta x^2}\right)^{k/2}\leq e^{Tba^2/2} \tag{31}$$

provided we select $\Delta t\leq b\Delta x^2$, for any finite $b>0$. Under this condition we have that the scheme (14) is Lax-Richtmyer stable, and therefore convergent.

- **Remark:** Although the scheme (14) is provably convergent (it is consistent and Lax-Richtmyer stable) it is easy to see that the method is in practice "unstable" for every finite Δt and Δx. In fact, by taking the modulus of (27) we obtain

$$\begin{aligned}\left|c_p^{k+1}\right|&=\left|c_p^k\right|\left|1-i\frac{a\Delta t}{\Delta x}\sin\left(\frac{2\pi}{L}p\Delta x\right)\right|\\ &=\left|c_p^k\right|\sqrt{1+\frac{a^2\Delta t^2}{\Delta x^2}\sin\left(\frac{2\pi}{L}p\Delta x\right)^2}\\ &\geq\left|c_p^k\right|.\end{aligned} \tag{32}$$

  This shows that the amplitude of each discrete Fourier mode is *always amplified as time integration proceeds*, no matter how we pick Δt and Δx.

  On the other hand, the analytical solution of (2) in Fourier space suggests that

$$c_p(t)=e^{-iat}c_p(0)\quad\Rightarrow\quad|c_p(t)|=\left|e^{-2\pi iapt/L}\right||c_p(0)|\quad\Rightarrow\quad|c_p(t)|=|c_p(0)|, \tag{33}$$

  i.e., the amplitude of each Fourier mode must be preserved. That's why the scheme (14) is often designated as "unstable". The situation here has similarities to the one we have seen when studying convergence of the leapfrog method applied to ODEs. In fact, such method is zero stable and consistent, and therefore convergent. However, the method is unconditionally absolutely unstable. Note, that here the solution doesn't really go to zero though.

Leapfrog time integration. Let us discretize (7) in time with the leapfrog method

$$u_j^{k+2} = u_j^k - a\frac{\Delta t}{\Delta x}\left(u_{j+1}^{k+1} - u_{j-1}^{k+1}\right). \tag{34}$$

We know that (34) is consistent with order two in both Δx and Δt. Let us perform a Von-Neumann stability analysis. To this end, we first substitute (26) into (34) to obtain

$$c_p^{k+2} = c_p^k - a\frac{\Delta t}{\Delta x}\left(e^{ip\xi} - e^{-ip\xi}\right)c_p^{k+1}. \tag{35}$$

At this point, we write the two-step method (35) as a one step method

$$\begin{bmatrix} c_p^{k+2} \\ d_p^{k+2} \end{bmatrix} = \underbrace{\begin{bmatrix} -2ai\Delta t\sin(p\xi)/\Delta x & 1 \\ 1 & 0 \end{bmatrix}}_{\boldsymbol{G}_p} \begin{bmatrix} c_p^{k+1} \\ d_p^{k+1} \end{bmatrix}. \tag{36}$$

In this case, the amplification matrix $\boldsymbol{G}$ is block-diagonal and symmetric, hence normal. The eigenvalues of $\boldsymbol{G}$ are easily obtained by computing the eigenvalues of each block. The characteristic polynomial corresponding to $\boldsymbol{G}_p$ is

$$\lambda^2 + \frac{2ia\Delta t\sin(\xi p)}{\Delta x}\lambda - 1 = 0. \tag{37}$$

The eigenvalues are

$$\lambda_{1,2}(p) = -\frac{ia\Delta t\sin(\xi p)}{\Delta x} \pm \sqrt{1 - \frac{a^2\Delta t^2\sin(\xi p)^2}{\Delta x^2}}. \tag{38}$$

At this point we notice that for all Δt and Δx such that

$$\left|a\frac{\Delta t}{\Delta x}\right| \leq 1 \tag{39}$$

we have that quantity within the square root in (38) is real. In this assumption, the modulus of the eigenvalues can be computed as

$$|\lambda_{1,2}(p)|^2 = \frac{a^2\Delta t^2\sin(\xi p)^2}{\Delta x^2} + 1 - \frac{a^2\Delta t^2\sin(\xi p)^2}{\Delta x^2} = 1. \tag{40}$$

This implies that the spectral radius of all $\boldsymbol{G}_p$ is equal to one, and therefore

$$\left\|\boldsymbol{G}^k\right\|_2 = \rho(\boldsymbol{G})^k = 1. \tag{41}$$

This proves that the leapfrog method is Lax-Richtmyer stable (provided (39) is satisfied), and therefore convergent. The condition (39) is called Courant-Friedrichs-Levy (CFL) condition, and it is described in more detail hereafter.

Remark: The calculation of the spectral radius of $\boldsymbol{G}$ is more involved when $|a\Delta t/\Delta x| \geq 1$. In fact, in this case we have that the square root in (38) is imaginary.

Lax-Friedrichs method. The Lax-Friedrichs scheme is obtained by replacing u_j^k in the Euler-Forward method (14) with the average over neighboring nodes, i.e.,

$$u_j^{k+1} = \frac{u_{j-1}^k + u_{j+1}^k}{2} - a\Delta t\frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x}. \tag{42}$$

The reason for the modification can be explained as follows. By adding and subtracting u_j^k from the right hand side of (42) we can write the scheme as

$$u_j^{k+1} = u_j^k - a\Delta t \frac{u_{j+1}^k - u_{j-1}^k}{2\Delta x} + \frac{\Delta x^2}{2}\left(\frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{\Delta x^2}\right). \tag{43}$$

In this form, we recognize that the scheme introduces a numerical diffusion term with amplitude proportional to the square of the grid spacing. The diffusion term is meant to counterbalance the numerical amplification of Fourier modes induced by the Euler-Forward scheme, see Eq. (32). It is straightforward to show that the Lax-Friedrichs method is consistent with order one in Δt and order two in Δx. Let us now perform a Von-Neumann stability analysis of the scheme (43) (assuming we are considering periodic boundary conditions in $[0, L]$). To this end, we substitute (26) into (43) to obtain

$$\begin{aligned} c_p^{k+1} =& c_p^k \left[1 - \frac{a\Delta t}{2\Delta x}\left(e^{ip\xi} - e^{-ip\xi}\right) + \frac{1}{2}\left(e^{ip\xi} + e^{-ip\xi} - 2\right)\right] \\ =& c_p^k \left[1 - i\frac{a\Delta t}{\Delta x}\sin(p\xi) + \cos(p\xi) - 1\right] \\ =& c_p^k \left[\cos(p\xi) - i\frac{a\Delta t}{\Delta x}\sin(p\xi)\right]. \end{aligned} \tag{44}$$

Again, we have a diagonal amplification matrix $\boldsymbol{G}$ with diagonal entries

$$G_p(\Delta t, \Delta x) = \cos(p\xi) - i\frac{a\Delta t}{\Delta x}\sin(p\xi). \tag{45}$$

Since $\boldsymbol{G}$ is diagonal, the Von-Neumann condition

$$\rho(\boldsymbol{G}) \leq 1 + \gamma\Delta t \tag{46}$$

is necessary and sufficient for stability. We have

$$\begin{aligned} \rho(\boldsymbol{G}) =& \max_{p=0,\ldots,N-1} |G_p(\Delta t, \Delta x)| \\ =& \max_{p=0,\ldots,N-1} \sqrt{\cos(p\xi)^2 + \frac{a^2\Delta t^2}{\Delta x^2}\sin(p\xi)^2}. \end{aligned} \tag{47}$$

Clearly, if

$$\left|a\frac{\Delta t}{\Delta x}\right| \leq 1 \tag{48}$$

then $\rho(\boldsymbol{G}) \leq 1$, and the Lax-Friedrichs method is stable. The condition (48) is known as *Courant-Friedrichs-Levy (CFL) condition*, and the number

$$\nu = |a|\,\frac{\Delta t}{\Delta x} \tag{49}$$

is known as *Courant number*.

CFL condition: The CFL condition is a necessary condition for convergence of a numerical scheme based on general statement that the *domain of dependence of the numerical scheme* must contain the *domain of dependence of the physical problem* (see Figure 1) at least for small Δx and Δt. For the particular case of a the linear PDEs we are studying in this section, the physical domain of dependence is a point (the root of the characteristic curve), while the domain of dependence of the numerical scheme is an interval. For the heat equation, the physical domain of dependence is the whole spatial domain.

Remark: Is there a CLF condition for the heat equation discretized with Euler forward and centered finite differences? To answer this question, recall that the Lax-Richtmyer stability condition is

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2\alpha} \tag{50}$$

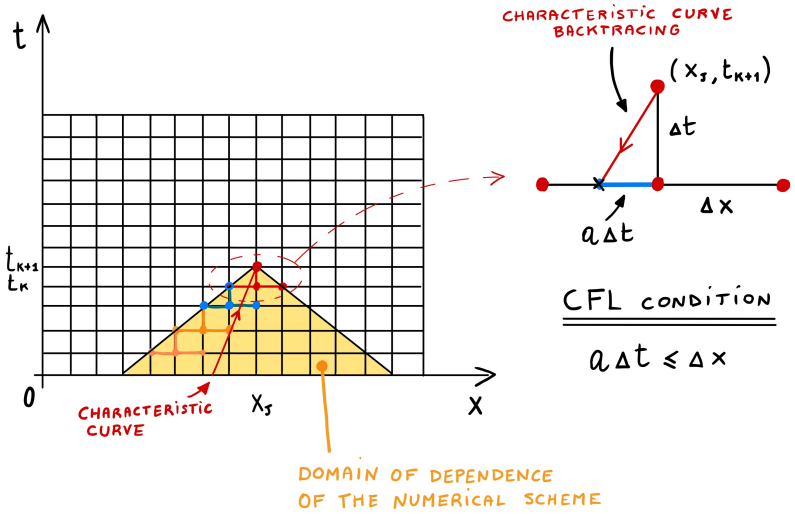

Figure 1: Illustration of the CFL condition, highlighting that the domain of dependence of the numerical scheme must contain that of the physical problem.

Hence, if $\Delta x \to 0$ we have that the domain of dependence of the numerical scheme becomes larger and larger and tends to be whole domain. In fact, Δt must shrinks as $\Delta x^2/(2\alpha)$ for the scheme to be stable. This implies that for $\Delta x \to 0$ the domain of numerical dependence is the whole domain.

- **Analysis of the Lax-Friedrichs scheme with the method of modified equations:** To analyze the scheme (42) it is possible to use another method based on the so-called "modified equation". Such equation represents the PDE that governs a smooth function $v(x,t)$ that satisfies the numerical scheme (42) exactly, i.e.,

$$v(x_j, t_{k+1}) = \frac{v(x_{j+1}, t_k) + v(x_{j-1}, t_k)}{2} - a\Delta t \frac{v(x_{j+1}, t_k) - v(x_{j-1}, t_k)}{2\Delta x}. \tag{51}$$

By expanding (51) in a Taylor series at $(x_j, t_k$ we obtain

$$v(x_j, t_k) + \frac{\partial v(x_j, t_k)}{\partial t}\Delta t + \frac{1}{2}\frac{\partial^2 v(x_j, t_k)}{\partial t^2}\Delta t^2 = v(x_j, t_k) + \frac{\partial^2 v(x_j, t_k)}{\partial x^2}\Delta x^2 - a\Delta t \frac{\partial v(x_j, t_k)}{\partial x} + \cdots \tag{52}$$

i.e.,

$$\frac{\partial v}{\partial t} + a\frac{\partial v}{\partial x} = \frac{\Delta t}{2}\left(\frac{\Delta x^2}{\Delta t^2}\frac{\partial^2 v}{\partial x^2} - \frac{\partial v^2}{\partial t^2}\right) + O(\Delta t^2). \tag{53}$$

This equation can be written as[6]

$$\begin{aligned}\frac{\partial v}{\partial t} + a\frac{\partial v}{\partial x} =& \frac{\Delta t}{2}\left(\frac{\Delta x^2}{\Delta t^2} - a^2\right)\frac{\partial^2 v}{\partial x^2} + O(\Delta t^2) \\ =& \frac{a\Delta x}{2\nu}\left(1 - \nu^2\right)\frac{\partial^2 v}{\partial x^2} + O(\Delta t^2).\end{aligned} \tag{55}$$

[6]To obtain (55) we first differentiate (53) with respect to time

$$\frac{\partial v}{\partial t} = -a\frac{\partial v}{\partial x} + O(\Delta t) \quad \Rightarrow \quad \frac{\partial^2 v}{\partial t^2} = -a\frac{\partial^2 v}{\partial x \partial t} + O(\Delta t) = a^2\frac{\partial^2 v}{\partial x^2} + O(\Delta t) \tag{54}$$

.

Hence, if the Courant number (49) satisfies $\nu \leq 1$ then we see that the numerical solution satisfies an advection-diffusion equation, which is known to have smooth solutions. On the other hand, if $\nu > 1$ the modified equation has "negative diffusion", which is an ill-posed problem.

Remark: Note that if

$$\nu = a\frac{\Delta t}{\Delta x} = 1 \tag{56}$$

then the amplitude of the Fourier modes c_p^k in (44) is preserved in time, i.e., we have

$$\left|c_p^{k+1}\right| = \left|c_p^k\right| \quad \text{for all} \quad p = 0, \ldots, N-1. \tag{57}$$

It is easy to see that with condition (56) the Lax-Friedrichs scheme (43) is actually exact for the linear advection equation (2). In fact, if we substitute (56) into (43) we obtain

$$u_j^{k+1} = u_j^k, \tag{58}$$

which is what the exact solution does along the characteristics curves evaluated on the grid. In other words, (56) sets up the (space-time) grid in a way that a characteristic passing through one point x_j lands at another grid point (either x_{j+1} or x_{j-1} after Δt time units).

Lax-Wendroff method. Consider the formal solution of the semi-discrete scheme (9)

$$\boldsymbol{u}(t_k + \Delta t) = e^{-a\Delta t \boldsymbol{D}_{\text{FD}}^1}\boldsymbol{u}(t_k) \tag{59}$$

and expand it to second-order in Δt. This yields

$$\boldsymbol{u}(t_k + \Delta t) \simeq \left(\boldsymbol{U} - a\Delta t \boldsymbol{D}_{\text{FD}}^1 + \frac{1}{2}a^2\Delta t^2 \boldsymbol{D}_{\text{FD}}^1\boldsymbol{D}_{\text{FD}}^1\right)\boldsymbol{u}(t_k). \tag{60}$$

Replacing $\boldsymbol{D}_{\text{FD}}^1\boldsymbol{D}_{\text{FD}}^1$ with the second-order differentiation matrix based on a stencil with three points, yields the *Lax-Wendroff method*

$$u_j^{k+1} = u_j^k - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^k - u_{j-1}^k\right) + \frac{a^2\Delta t^2}{2\Delta x^2}\left(u_{j-1}^k - 2u_j^k + u_{j+1}^k\right). \tag{61}$$

It is straightforward to show that the method is consistent with order two in both Δt and Δx. Regarding stability, a substitution of (26) into (61) yields the following equation for the amplification factors of the discrete Fourier modes

$$c_p^{k+1} = c_p^k \underbrace{\left[1 - i\frac{a\Delta t}{\Delta x}\sin(p\xi) + \frac{a^2\Delta t^2}{\Delta x^2}\left(\cos(p\xi) - 1\right)\right]}_{G_p(\Delta t, \Delta x)}. \tag{62}$$

As before the amplification matrix $\boldsymbol{G}$ is diagonal, with diagonal entries G_p. By using the trigonometric identities

$$\cos(p\xi) - 1 = -2\sin^2\left(\frac{p\xi}{2}\right), \qquad \sin(p\xi) = 2\sin\left(\frac{p\xi}{2}\right)\sin\left(\frac{p\xi}{2}\right) \tag{63}$$

we can rewrite G_p in (62) as

$$G_p(\Delta t, \Delta x) = 1 - 2i\frac{a\Delta t}{\Delta x}\sin\left(\frac{p\xi}{2}\right)\cos\left(\frac{p\xi}{2}\right) - 2\frac{a^2\Delta t^2}{\Delta x^2}\sin^2\left(\frac{p\xi}{2}\right). \tag{64}$$

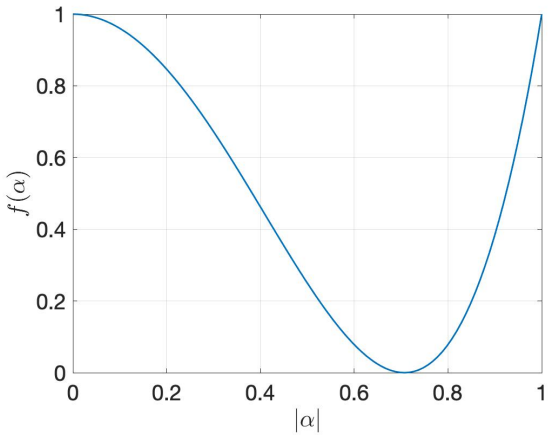

Figure 2: Graph of the function (67) characterizing the square of spectral radius of the amplification matrix $\boldsymbol{G}$ for the Lax-Wendroff method. It is seen that for $|\alpha| \leq 1$ the method is stable.

Taking the modulus yields

$$\begin{aligned}
|G_p(\Delta t, \Delta x)|^2 =& \left[1 - 2\frac{a^2\Delta t^2}{\Delta x^2}\sin^2\left(\frac{p\xi}{2}\right)\right]^2 + 4\frac{a^2\Delta t^2}{\Delta x^2}\sin^2\left(\frac{p\xi}{2}\right)\cos^2\left(\frac{p\xi}{2}\right) \\
=& 1 + 4\frac{a^4\Delta t^4}{\Delta x^4}\sin^4\left(\frac{p\xi}{2}\right) - 4\frac{a^2\Delta t^2}{\Delta x^2}\sin^2\left(\frac{p\xi}{2}\right) + 4\frac{a^2\Delta t^2}{\Delta x^2}\sin^2\left(\frac{p\xi}{2}\right)\cos^2\left(\frac{p\xi}{2}\right) \\
=& 1 + 4\frac{a^4\Delta t^4}{\Delta x^4}\sin^4\left(\frac{p\xi}{2}\right) - 4\frac{a^2\Delta t^2}{\Delta x^2}\sin^4\left(\frac{p\xi}{2}\right) \\
=& 1 - 4\frac{a^2\Delta t^2}{\Delta x^2}\left(1 - \frac{a^2\Delta t^2}{\Delta x^2}\right)\sin^4\left(\frac{p\xi}{2}\right).
\end{aligned} \tag{65}$$

With this expression, we can easily bound the spectral radius of the amplification matrix $\boldsymbol{G}$ as

$$\begin{aligned}
\rho(\boldsymbol{G})^2 =& \max_p |G_p(\Delta t, \Delta x)|^2 \\
\leq& 1 - 4\frac{a^2\Delta t^2}{\Delta x^2}\left(1 - \frac{a^2\Delta t^2}{\Delta x^2}\right).
\end{aligned} \tag{66}$$

In Figure 2 we plot the function

$$f(\alpha) = 1 - 4\alpha^2(1-\alpha^2) \quad \text{versus} \quad |\alpha| = |a|\frac{\Delta t}{\Delta x}. \tag{67}$$

It is seen that $f(\beta) \leq 1$ for all $|\beta| \leq 1$. This allows us to conclude that the Lax-Wendroff method is stable (and therefore convergent) if

$$\left|a\frac{\Delta t}{\Delta x}\right| \leq 1. \tag{68}$$

From equation (65) we also see that if $|a\Delta t/\Delta x| = 1$ then the amplitude of each discrete Fourier mode is preserved. As before, the condition $|a\Delta t/\Delta x| = 1$ implies that we are working with a space-time grid that is defined by the discrete characteristic curves of (2).

- **Analysis of the Lax-Wendoff scheme with the method of modified equations:** As before, consider a smooth function $v(x,t)$ that satisfies the numerical scheme (61) exactly, i.e.,

$$v(x_j, t_{k+1}) = v(x_j, t_k) - \frac{a\Delta t}{2\Delta x}\left(v(x_{j+1}, t_k) - v(x_{j-1}, t_k)\right) + \frac{a^2\Delta t^2}{\Delta x^2}\left(v(x_{j+1}, t_k) - 2v(x_j, t_k) + v(x_{j-1}, t_k)\right). \tag{69}$$

By expanding v in a Taylor series at (x_j, t_k) we obtain

$$\frac{\partial v}{\partial t} + a\frac{\partial v}{\partial x} = -\frac{a\Delta x^2}{6}\left(1-\nu^2\right)\frac{\partial^3 v}{\partial x^3} + O(\Delta t^2). \tag{70}$$

This equation is fundamentally different from (55) since the right-hand side contains a third order term, i.e., it is a dispersive term for which there is no guarantee of smooth decay. Furthermore, the dispersive nature of the term will cause solutions with different slopes to propagate at different speeds, typically resulting in a train of waves.

Appendix A: The method of characteristics

Consider the semi-linear scalar first-order PDE

$$\begin{cases} \dfrac{\partial U(\boldsymbol{x},t)}{\partial t} + \boldsymbol{a}(\boldsymbol{x},t)\cdot\nabla U(\boldsymbol{x},t) = f(\boldsymbol{x},t,U(\boldsymbol{x},t)) \qquad \boldsymbol{x}\in\mathbb{R}^d \qquad t\geq 0 \\ U(\boldsymbol{x},0) = U_0(\boldsymbol{x}) \end{cases} \tag{71}$$

This equation can be transformed into an ODE on the flow generated by the nonlinear dynamical system

$$\frac{\boldsymbol{X}(t,\boldsymbol{x}_0)}{dt} = \boldsymbol{a}(\boldsymbol{X}(t,\boldsymbol{x}_0),t) \qquad \boldsymbol{X}(0,\boldsymbol{x}_0) = \boldsymbol{x}_0. \tag{72}$$

The aforementioned ODE is[7]

$$\frac{dz}{dt} = f(\boldsymbol{X}(t,\boldsymbol{x}_0),t,z(t)) \qquad z(0) = U_0(\boldsymbol{x}_0) \tag{75}$$

The meaning of $\boldsymbol{X}(t,\boldsymbol{x}_0)$ and $z(t)$ is summarized in Figure 3.

If we are interested in the solution of (71) at a particular point in space, say $\boldsymbol{x}^*$ (e.g., a point on a grid) and a particular time say t^* then we proceed as follows:

1. Integrate (72) backward in time from $t = t^*$ to $t = 0$ with initial condition $\boldsymbol{x}^*$. That gives us the point $\boldsymbol{x}_0^*$ shown in Figure (4)
2. with $\boldsymbol{x}_0^*$ available, we integrate (75) forward in time from $t = 0$ to $t = t^*$.

We can use this method to compute the solution of (71) at time $t = t^*$ at all spatial grid points of a given grid. To do so, we simply need to solve (72) backward and (75) forward at each grid point.

Geometric aspects of first-order PDEs. Consider the case where the PDE (71) is defined in a one-dimensional spatial domain, i.e.,

$$\frac{\partial U(x,t)}{\partial t} + a(x,t)\frac{\partial U(x,t)}{\partial x} = f(x,t,U(x,t)) \qquad x\in\mathbb{R} \qquad t\geq 0. \tag{76}$$

[7]Equation (75) is easily derived by defining

$$z(t) = U(\boldsymbol{X}(t,\boldsymbol{x}_0),t) \qquad \text{(solution along the flow)} \tag{73}$$

and noting that

$$\frac{dz(t)}{dt} = \frac{dU(\boldsymbol{X}(t,\boldsymbol{x}_0),t)}{dt} = \frac{\partial U(\boldsymbol{X}(t,\boldsymbol{x}_0),t)}{\partial t} + \boldsymbol{a}(\boldsymbol{x},t)\cdot\nabla U(\boldsymbol{X}(t,\boldsymbol{x}_0),t) = f(\boldsymbol{X}(t,\boldsymbol{x}_0),t,z(t)). \tag{74}$$

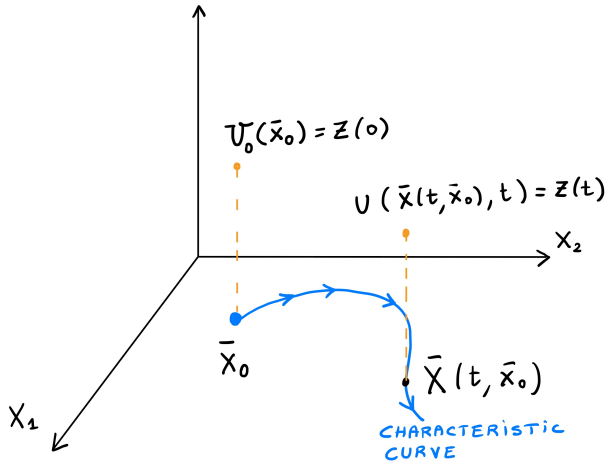

Figure 3: Sketch of the method of characteristics.

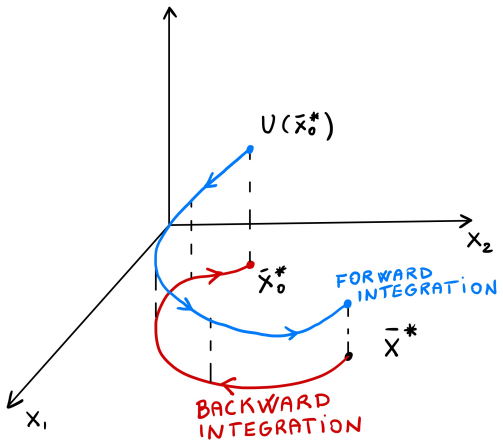

Figure 4: Sketch of the process used to compute the solution of the PDE (71) at a particular point $\boldsymbol{x}^*$ and particular time t^*. Essentially, we integrate the characteristic system (72) backward in time from $t = t^*$ and position $\boldsymbol{x}^*$ to $t = 0$. This gives us the point $\boldsymbol{x}_0^*$. Next we integrate (75) forward in time with initial condition $z(0) = U(\boldsymbol{x}_0^*)$ along the same characteristic curve.

The solution to this PDE (if it exists) defines a surface $\mathbb{R}^3$

$$(x,t) \mapsto U(x,t) \tag{77}$$

Equation (76) can be written as

$$\begin{bmatrix} 1 & a(x,t) & f(x,t,U) \end{bmatrix} \cdot \underbrace{\begin{bmatrix} \dfrac{\partial U}{\partial t} & \dfrac{\partial U}{\partial x} & -1 \end{bmatrix}}_{\perp \text{ to the surface } U(x,t)} = 0 \tag{78}$$

Recall, in fact, that if we are given a surface in $\mathbb{R}^3$ defined as $z = U(x,t)$, i.e., $F(t,x,z) = U(x,t) - z = 0$ then the gradient

$$\begin{bmatrix} \dfrac{\partial F}{\partial t} & \dfrac{\partial F}{\partial x} & \dfrac{\partial F}{\partial z} \end{bmatrix} \tag{79}$$

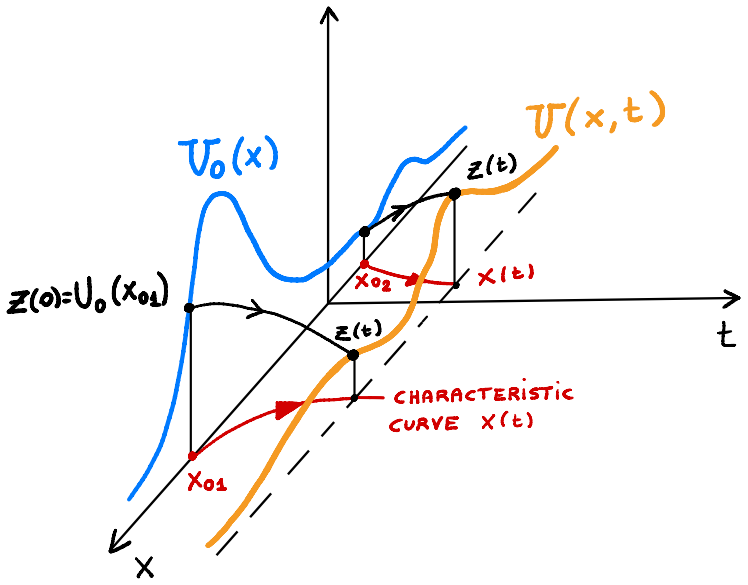

Figure 5: Geometric interpretation of the process used to compute the solution of the PDE (71) using the characteristic system (81).

is orthogonal to F at each point $(x,t,U(x,t))$. This means that the vector

$$\boldsymbol{t}(x,t) = \begin{bmatrix} 1 & a(x,t) & f(x,t,U(x,t)) \end{bmatrix} \tag{80}$$

belongs to the plane that is tangent to of $U(x,t)$ at each point $(x,t,U(x,t))$. Therefore, the solution of the PDE (76) can be constructed as infinite union of curves obtained by integrating the vector field $\boldsymbol{t}(x,t)$ relative to some parameter s, i.e.,

$$\begin{cases} \dfrac{dt(s)}{ds} = 1 \\ \dfrac{dx(s)}{ds} = a(x(s),t(s)) \\ \dfrac{dz(s)}{ds} = f(x(s),t(s),z(s)) \\ t(0) = 0 \\ x(0) = x_0 \\ z(0) = U_0(x_0) \end{cases} \tag{81}$$

In Figure 5 we sketch the meaning of the characteristic system (81). Based on such sketch it is clear that we need a little bit careful when we set boundary conditions for $U(x,t)$. In fact, depending on the direction of the characteristic curves we may end up setting boundary conditions that are incompatible with the solution. As an example, it is straightforward to see that the initial/boundary value problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial u}{\partial x} = 0 \\ (U(x,0) = \sin(x) \\ U(2\pi,t) = \sin(2t) \end{cases} \tag{82}$$

has no solution, i.e., it is ill-posed! In fact the initial condition $U_0(x)$ is advected to the right with velocity 1 towards the boundary at $x = 2\pi$, on which we set a condition is incompatible with the solution along the characteristic $U(2\pi, t) = sin(2\pi - t)$.

Nonlinear first-order PDEs. More generally, the method of characteristics can be applied to solve first-order nonlinear PDEs of the form

$$\frac{\partial U(\boldsymbol{x},t)}{\partial t} + \boldsymbol{a}(\boldsymbol{x},t,U(\boldsymbol{x},t)) \cdot \nabla U(\boldsymbol{x},t) = f(\boldsymbol{x},t,U(\boldsymbol{x},t)). \tag{83}$$

In this case the characteristic system is

$$\begin{cases} \dfrac{\boldsymbol{X}(t)}{dt} = \boldsymbol{a}(\boldsymbol{X}(t),t,z(t)), & \boldsymbol{X}(0,\boldsymbol{x}_0) = \boldsymbol{x}_0, \\ \dfrac{z(t)}{dt} = f(\boldsymbol{X}(t),t,z(t)), & z(0) = U_0(\boldsymbol{x}_0). \end{cases} \tag{84}$$

Note that, in this case,computing the solution at a specific point in space and time is not as easy as before since (84) is coupled. In other words, when we integrate (84) backward in time we need to guess $z(t)$. Long story short, to compute the solution of (84) at a specific point in space and time, we could use the shooting method, the control variable being $z(t)$ at $\boldsymbol{x}^*$.

References

[1] J. S. Hesthaven. *Numerical methods for conservation laws: from analysis to algorithms.* SIAM, 2018.

[2] R. J. LeVeque. *Numerical methods for conservation laws.* Birkhäuser, 1992.

[3] R. J. LeVeque. *Finite difference methods for ordinary and partial differential equations.* SIAM, 2007.