Chapter 12

Method of Lines (MOL)

12.1 MOL for the heat equation

Definition 12.1. A second-order, constant-coefficient, linear partial differential equation (PDE) of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \qquad (12.1)$$

is called a *parabolic PDE* if its coefficients satisfy

$$B^2 - 4AC = 0. \qquad (12.2)$$

Remark 12.1. We say the equation (12.1) is elliptic or hyperbolic if $B^2 - 4AC$ is less than or greater than zero, respectively.

Remark 12.2. The name "parabolic" is used because the assumption on the coefficients is the same as the condition upon the equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

in defining a planar parabola in analytic geometry. This algebraic equation corresponds to the geometric view where, a conic section, i.e. the intersection of a plane to a circular cone, may be an ellipse, a hyperbola, a parabola, or other degenerate cases such as straight lines.

Remark 12.3. Usually the variable x in (12.1) represents one-dimensional position and y represents time, and the PDE is solved subject to prescribed initial and boundary conditions.

Definition 12.2. The *one-dimensional heat equation* is a parabolic PDE of the form

$$u_t = \nu u_{xx} \text{ in } \Omega := (0, 1) \times (0, T), \qquad (12.3)$$

where $x \in (0, 1)$ is the spatial location, $t \in (0, T)$ the time and $\nu > 0$ the dynamic viscosity; the equation has to be supplemented with an *initial condition*

$$u(x, 0) = \eta(x), \text{ on } (0, 1) \times \{0\} \qquad (12.4)$$

and appropriate boundary conditions at $\{0, 1\} \times (0, T)$.

Definition 12.3. An *initial-boundary value problem* is the problem of determining a solution to a differential equation with both initial conditions and boundary conditions.

Example 12.4. We derive the heat equation based on the principle of conservation. Let $u(x, t)$ be the *density* or *concentration* function (mass/length) for some substance, then the total mass within $[a, b]$ at time t is

$$M(t) := \int_a^b u(x, t)\mathrm{d}x.$$

Let $F(x, t)$ denote the *flux* of material across the point x at time t (mass/time) and $\psi(x, t)$ the *source* or *sink* of the substance (mass/length/time). Since the total mass within $[a, b]$ changes due to the flux at the endpoints and the source(or sink) within $[a, b]$, we have

$$\begin{aligned}\frac{\mathrm{d}M(t)}{\mathrm{d}t} &= F(a, t) - F(b, t) + \int_a^b \psi(x, t)\mathrm{d}x \\ &= -\kappa(a)\frac{\partial u(a, t)}{\partial x} + \kappa(b)\frac{\partial u(b, t)}{\partial x} + \int_a^b \psi(x, t)\mathrm{d}x \\ &= \int_a^b \frac{\partial}{\partial x}\left(\kappa(x)\frac{\partial u(x, t)}{\partial x}\right)\mathrm{d}x + \int_a^b \psi(x, t)\mathrm{d}x,\end{aligned}$$

where the second step follows from Fick's law of diffusion and the third step from the second fundamental theorem of calculus (Theorem C.74). By definition of the total mass, we have

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\int_a^b u(x, t)\mathrm{d}x = \int_a^b \frac{\partial u(x, t)}{\partial t}\mathrm{d}x,$$

and therefore

$$\int_a^b \left[\frac{\partial u(x, t)}{\partial t} - \frac{\partial}{\partial x}\left(\kappa(x)\frac{\partial u(x, t)}{\partial x}\right) - \psi(x, t)\right]\mathrm{d}x = 0.$$

Since the integral must be zero for all values of a and b, it follows that the integrand must be identically zero. This gives, finally, the differential equation

$$\frac{\partial u(x, t)}{\partial t} = \frac{\partial}{\partial x}\left(\kappa(x)\frac{\partial u(x, t)}{\partial x}\right) + \psi(x, t).$$

If the material is homogeneous, then $\kappa(x) = \kappa$ is independent of x and the above equation reduces to

$$\frac{\partial u(x, t)}{\partial t} = \kappa\frac{\partial^2 u(x, t)}{\partial x^2} + \psi(x, t).$$

Remark 12.4. The role of ν is to change the time scale since, by defining a new time variable $\tau = \nu t$, we can annihilate ν from the equation. This suggests that for large ν the physical processes are faster.

Remark 12.5. If L is an elliptic operator with a negative definite matrix, then the time-dependent equation

$$u_t = Lu - f$$

is well-posed and considered parabolic. In particular, if L is the Laplacian, then the above equation is the heat equation in multiple dimensions.

Remark 12.6. A system or a process is in a *steady state* if the state variables that define the behavior of the system or the process are unchanging in time, i.e. $\frac{\partial u}{\partial t} = 0$. The steady state of the heat equation is the Laplace equation or Poisson's equation.

Theorem 12.5. The exact solution to the heat equation (12.3) with Dirichlet conditions $g_0(t) = g_1(t) = 0$ is

$$u(x,t) = \sum_{j=0}^{\infty} \hat{u}_j(t)\sin(\pi j x), \tag{12.5}$$

where

$$\hat{u}_j(t) = \exp(-j^2\pi^2\nu t)\hat{u}_j(0), \tag{12.6}$$

and $\hat{u}_j(0)$ is the coefficient of the Fourier mode $\sin(\pi j x)$ in the initial data $u(x,0) = \sum_{j=0}^{\infty} \hat{u}_j(0)\sin(\pi j x)$, c.f. (12.5).

Proof. It is straightforward to verify that (12.5) is indeed the solution of (12.3). □

Remark 12.7. How is the solution (12.5) derived? We assume that the solution is of the form $u(x,t) = f(x)g(t)$, plug it into the equation (12.3), and obtain

$$\frac{g'(t)}{g(t)} = \frac{f''(x)}{f(x)}.$$

The LHS is independent of x and the RHS is independent of t, and thus both sides must be a constant, say $-k^2$. Then we have $u(x,t) = e^{-k^2 t}\sin kx$. Since the heat equation is known to have bounded solutions, we know that we have chosen the correct sign of the constant $-k^2$. Finally, the homogeneous boundary conditions are satisfied if k is restricted to $m\pi$ where m is a positive integer.

Remark 12.8. By Theorem 12.5, we have an infinite system of ODEs for Fourier modes of different wavenumbers,

$$\forall j = 1, 2, \ldots, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\hat{u}_j(t) = -j^2\pi^2\hat{u}_j(t).$$

As j increases, the exponential decay becomes faster and faster. For a general initial condition that consists of high oscillations, there would be a short transient period when high-frequency modes are quickly damped. In contrast, the long-time behavior of the heat equation is dominated by the slower decay rates. Therefore, the physical process of diffusion can be infinitely stiff since the wavenumber j can be any integer.

Although the stiffness of a discrete system is always limited by the highest wavenumber the spatial grid can represent, the potential maximum stiffness approaches infinity as $h \to 0$. See the discussions in Section 11.4.

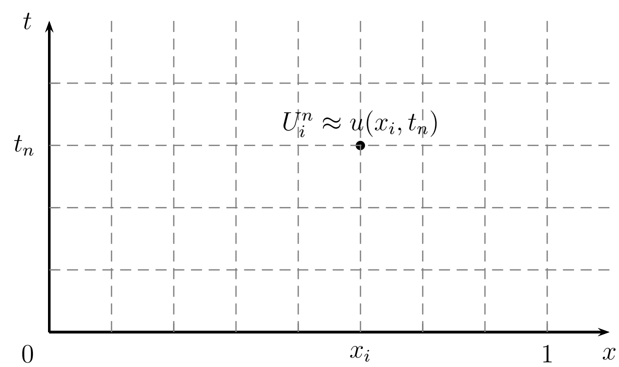
As for the long-time behavior of the heat equation with homogeneous Dirichlet conditions, Theorem 12.5 dictates that $\lim_{t\to+\infty} u(x,t) = 0$.

12.1.1 FTCS and Crank-Nicolson

Notation 12. The space-time domain of the PDE (12.3) can be discretized by the rectangular grids

$$x_i = ih, \quad t_n = nk, \tag{12.7}$$

$h = \frac{1}{m+1}$ is the uniform mesh spacing and $k = \Delta t$ is the uniform time-step size. The unknowns U_i^n are located at nodes (x_i, t_n).



Remark 12.9. To understand how stability theory for time-dependent PDEs relates to the stability theory for time-dependent ODEs, we consider the MOL for PDEs.

Definition 12.6. The *method of lines* (MOL) is a technique for solving PDEs via

(a) discretizing the spatial derivatives while leaving the time variable continuous;

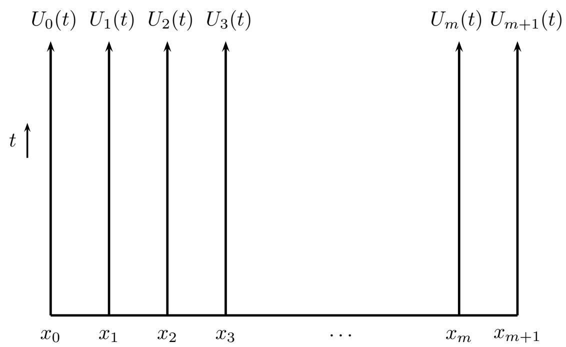
(b) solving the resulting ODEs with a numerical method designed for IVPs.

Remark 12.10. In Definition 12.6, the system of ODEs produced by step (a) is also called *semi-discrete*, since we have discretized in space but not in time yet.

Example 12.7. Discretize the heat equation (12.3) in space at grid point x_i by

$$U_i'(t) = \frac{\nu}{h^2}\Big(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)\Big), \tag{12.8}$$

where $U_i(t) \approx u(x_i, t)$ for $i = 1, 2, \cdots, m$.



For Dirichlet conditions

$$\begin{cases} u(0,t) = g_0(t), & \text{on } \{0\} \times (0,T); \\ u(1,t) = g_1(t), & \text{on } \{1\} \times (0,T), \end{cases} \tag{12.9}$$

this semi-discrete system (12.8) can be written as

$$\mathbf{U}'(t) = A\mathbf{U}(t) + \mathbf{g}(t), \tag{12.10}$$

where

$$A = \frac{\nu}{h^2}\begin{bmatrix} -2 & +1 & & & & \\ +1 & -2 & +1 & & & \\ & +1 & -2 & +1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & +1 & -2 & +1 \\ & & & & +1 & -2 \end{bmatrix}, \tag{12.11}$$

$$\mathbf{U}(t) := \begin{bmatrix} U_1(t) \\ U_2(t) \\ U_3(t) \\ \vdots \\ U_{m-1}(t) \\ U_m(t) \end{bmatrix}, \quad \mathbf{g}(t) = \frac{\nu}{h^2}\begin{bmatrix} g_0(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}. \tag{12.12}$$

By the negative definiteness of A, Definition 11.271, and Example 11.272, the ODE system (12.10) is dissipative.

Remark 12.11. If the boundary conditions of the heat equation is Dirichlet, the leftmost and rightmost lines in the plot in Example 12.7 are determined before the application of any time integrator to the semi-discrete system. However, this is not true for Neumann conditions, in which case the known vector in (12.12) would be augmented with $U_0(t)$ and $U_{m+1}(t)$ and the matrix in (12.11) with two more rows and columns accordingly; see Section 7.5.

Notation 13. The non-dimensional number

$$r := \frac{k\nu}{h^2} \tag{12.13}$$

is often used in numerically solving the heat equation.

Definition 12.8. The *FTCS (forward in time, centered in space) method* solves the heat equation (12.3) by

$$\frac{U_i^{n+1} - U_i^n}{k} = \frac{\nu}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n), \tag{12.14}$$

or, equivalently

$$U_i^{n+1} = U_i^n + r(U_{i-1}^n - 2U_i^n + U_{i+1}^n). \tag{12.15}$$

Remark 12.12. The FTCS method is based on central difference in space and the forward Euler method in time; it can also be considered as an MOL with the forward Euler as its time integrator.

Example 12.9. For homogeneous Dirichlet boundary conditions, the FTCS method can be written as

$$\mathbf{U}^{n+1} = (I + kA)\mathbf{U}^n, \tag{12.16}$$

where A is the matrix in (12.11) and

$$\mathbf{U}^n := \begin{bmatrix} U_1^n \\ U_2^n \\ \vdots \\ U_m^n \end{bmatrix}. \tag{12.17}$$

Definition 12.10. The *Crank-Nicolson method* solves the heat equation (12.3) by

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{k} &= \frac{1}{2}\Big(f(U_i^n, t_n) + f(U_i^{n+1}, t_{n+1})\Big) \\ =&\frac{\nu}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}), \end{aligned} \tag{12.18}$$

or, equivalently

$$\begin{aligned} &- rU_{i-1}^{n+1} + 2(1+r)U_i^{n+1} - rU_{i+1}^{n+1} \\ =&rU_{i-1}^n + 2(1-r)U_i^n + rU_{i+1}^n. \end{aligned} \tag{12.19}$$

Remark 12.13. The Crank-Nicolson method is an MOL with the trapezoidal rule as the time integrator.

Exercise 12.11. Show that the matrix form of the Crank-Nicolson method for solving the heat equation (12.3) with Dirichlet conditions is

$$\left(I - \frac{k}{2}A\right)\mathbf{U}^{n+1} = \left(I + \frac{k}{2}A\right)\mathbf{U}^n + \mathbf{b}^n, \tag{12.20}$$

where

$$\mathbf{b}^n = \frac{r}{2}\begin{bmatrix} g_0(t_n) + g_0(t_{n+1}) \\ 0 \\ \vdots \\ 0 \\ g_1(t_n) + g_1(t_{n+1}) \end{bmatrix}.$$

Definition 12.12. For $\theta \in [0,1]$, the *θ-method* solves the heat equation (12.3) by

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{k} = & \frac{\nu}{h^2}\big[\theta(U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}) \\ & +(1-\theta)(U_{i-1}^n - 2U_i^n + U_{i+1}^n)\big], \end{aligned} \tag{12.21}$$

or, equivalently

$$\begin{aligned} &-\theta rU_{i-1}^{n+1} + (1+2\theta r)U_i^{n+1} - \theta rU_{i+1}^{n+1} \\ &= (1-\theta)rU_{i-1}^n + [1 - 2(1-\theta)r]U_i^n + (1-\theta)rU_{i+1}^n. \end{aligned} \tag{12.22}$$

Example 12.13. The θ-method with $\theta = 0$ is the FTCS method, that with $\theta = \frac{1}{2}$ is the Crank-Nicolson method, and that with $\theta = 1$ is the *BTCS (backward in time and centered in space)* of which the ODE solver is the backward Euler method.

12.1.2 Accuracy and consistency

Remark 12.14. The Local truncation error of an MOL is defined in the same way as that of an IVP solver.

Definition 12.14. The *local truncation error (LTE) of an MOL* for solving a PDE is the error caused by replacing continuous derivatives with finite difference formulas.

Definition 12.15. An MOL is said to be *consistent* if

$$\lim_{k,h\to 0}\tau(x,t)=0. \tag{12.23}$$

Definition 12.16. An MOL is said to be *pth-order accurate in time* and *qth-order accurate in space* iff its LTE satisfies $\tau(x,t)=O(k^p+h^q)$.

Remark 12.15. In Definition 12.16 we use the big-O notation instead of the theta notation to comply with the convention in the community of numerical PDEs.

Example 12.17. The LTE of the FTCS method in Definition 12.8 is

$$\begin{aligned}\tau(x,t)= &\ \tfrac{u(x,t+k)-u(x,t)}{k}\\ &-\tfrac{\nu}{h^2}\Big(u(x-h,t)-2u(x,t)+u(x+h,t)\Big)\\ = &\ \Big(u_t+\tfrac12 ku_{tt}+\tfrac16 k^2u_{ttt}+\cdots\Big)\\ &-\nu\Big(u_{xx}+\tfrac{1}{12}h^2u_{xxxx}+\cdots\Big)\\ = &\ \Big(\tfrac12 k\nu^2-\tfrac{\nu}{12}h^2\Big)u_{xxxx}+O(k^2+h^4),\end{aligned}$$

where the first step follows from the Definition 11.59, the second from Taylor expansions and the last from $u_t=\nu u_{xx}$ and $u_{tt}=\nu u_{xxt}=\nu u_{txx}=\nu^2u_{xxxx}$. By Definition 12.16, the FTCS method is second-order accurate in space and first-order accurate in time.

Lemma 12.18. The LTE of the θ-method is

$$\begin{aligned}\tau_i^{n+\frac12} &=\left(\tfrac12-\theta\right)k\nu u_{xxt}-\tfrac{1}{12}h^2\nu u_{xxxx}\\ &+\tfrac{1}{12}\left(\tfrac12-\theta\right)kh^2\nu u_{xxxxt}+O(k^2+h^4).\end{aligned} \tag{12.24}$$

Proof. By Definition 12.14, we have

$$\begin{aligned}&\tau_i^{n+\frac12}=\tfrac{u(x_i,t_n+k)-u(x_i,t_n)}{k}\\ &-\tfrac{\nu}{h^2}\theta[u(x_i-h,t_n+k)-2u(x_i,t_n+k)+u(x_i+h,t_n+k)]\\ &-\tfrac{\nu}{h^2}(1-\theta)[u(x_i-h,t_n)-2u(x_i,t_n)+u(x_i+h,t_n)],\end{aligned}$$

which, together with Taylor expansions at $(x_i,t_{n+\frac12})$ and the equation $u_t=\nu u_{xx}$, yields (12.24). □

Corollary 12.19. The Crank–Nicolson method is second-order accurate both in space and in time.

Proof. This follows from setting $\theta=\frac12$ in Lemma 12.18. □

Example 12.20. Set $\theta=\frac12-\frac{h^2}{12k\nu}$ and we get a method whose LTE is $O(k^2+h^4)$. Note that this requires $h^2\le 6k\nu$ to ensure $\theta\ge 0$, c.f. Definition 12.12.

12.1.3 Absolute stability

Lemma 12.21. The eigenvalues λ_p and eigenvectors $\mathbf{w}^p$ of A in (12.11) are

$$\lambda_p=-\frac{4\nu}{h^2}\sin^2\left(\frac{p\pi h}{2}\right), \tag{12.25}$$

$$w_j^p=\sin(p\pi jh), \tag{12.26}$$

where $p,j=1,2,\cdots,m$ and $h=\frac{1}{m+1}$.

Proof. This follows directly from Lemma 7.25. □

Example 12.22. For the FTCS method (12.14) to be absolutely stable, we must have $|1+k\lambda|\le 1$ for each eigenvalue in (12.25), which implies $-2\le -4\nu k/h^2\le 0$ and thus limits the time-step size to

$$k\le\frac{h^2}{2\nu}, \tag{12.27}$$

which is equivalent to $r\le\frac12$.

Remark 12.16. From Lemma 12.21, we have

$$\lambda_1\approx-\frac{4\nu}{h^2}\frac{\pi^2h^2}{2^2}=-\nu\pi^2,\quad \lambda_m\approx-\frac{4\nu}{h^2},$$

and hence the stiffness ratio in Definition 11.146 is

$$\frac{|\lambda|_{\max}}{|\lambda|_{\min}}=\frac{4}{\pi^2h^2},$$

which could be arbitrarily large as $h\to 0$. Consequently, the heat equation is stiff for an ODE solver with bounded RAS such as the FTCS method. As a typical symptom of stiffness, (12.27) is a severe restriction on the time-step size as we refine the grid: $k=O(h^2)$, which can certainly be regarded as "excessively small."

For better efficiency, we would like to use a time-step size $k=O(h)$. This holds for pure diffusion equation even when ν is very small such as $\nu=10^{-6}$. However, in the case of a mixed PDE such as the advection-diffusion equation where there are other time scales determined by other processes, we may be able to use an explicit ODE solver for the case of ν being sufficiently small.

Definition 12.23. An MOL is said to be *unconditionally stable (in the sense of absolute stability)* for a PDE if in solving the semi-discrete system of the PDE its ODE solver is absolutely stable for any $k>0$.

Lemma 12.24. Suppose the ODE solver of the MOL is $A(\alpha)$-stable for the semi-discrete system that results from spatially discretizing the heat equation. Then the MOL is unconditionally stable for the heat equation.

Proof. The RAS of an $A(\alpha)$-stable method contains the negative real axis. All eigenvalues of the heat equations are negative real numbers, hence $k\lambda$ is in the RAS for any $k>0$. □

Lemma 12.25. The θ-method (12.21) is unconditionally stable for $\theta\in[\frac12,1]$. For $\theta\in[0,\frac12)$, the time step size must satisfy $k\le\frac{h^2}{2(1-2\theta)\nu}$ for the θ-method to be stable.

Exercise 12.26. Prove Lemma 12.25 via the stability function of one-step methods.

Corollary 12.27. The Crank-Nicolson method (12.19) is unconditionally stable for the heat equation.

Proof. This follows directly from Lemma 12.25. □

Remark 12.17. In comparison to the FTCS method, the Crank-Nicolson method allows us to determine the value of $k > 0$ from the desired accuracy. For efficiency, we usually choose $k = O(h)$, a benefit from the unconditional stability.

12.1.4 Lax-Richtmyer stability

Definition 12.28. A linear MOL of the form

$$\mathbf{U}^{n+1} = B_h(k)\mathbf{U}^n + \mathbf{b}_h^n(k) \tag{12.28}$$

is *Lax-Richtmyer stable* iff

$$\begin{array}{ll} \forall T > 0, & \exists h_0, k_0, C_T > 0, \text{ s.t.} \\ & \forall k \in (0, k_0], \forall h \in (0, h_0], \forall n \in \mathbb{N}^+, \\ & nk \le T \implies \|B_h(k)^n\| \le C_T, \end{array} \tag{12.29}$$

where B_h is the MOL iteration matrix for the grid with size h and the constant C_T depends neither on k nor on h.

More specifically, the MOL (12.28) is *Lax-Richtmyer stable under the constraint* $\mathbf{g}(k, h) \le \mathbf{0}$ iff (12.29) holds with its third line replaced with

$$\mathbf{g}(k, h) \le \mathbf{0},\ nk \le T \implies \|B_h(k)^n\| \le C_T,$$

where $\mathbf{g}(k, h) \le \mathbf{0}$ means

$$\forall i = 1, 2, \ldots, m, \quad g_i(k, h) \le 0 \text{ or } g_i(k, h) < 0$$

and each $g_i(k, h)$ is an analytic function.

Remark 12.18. The constraint $\mathbf{g}(k, h) \le \mathbf{0}$ leads to a semi-analytic set in the k-h plane. Note that an equation on k and h such as $k = h^2$ can be described as $g_1 = k - h^2 \le 0$ and $g_2 = h^2 - k \le 0$. Sometimes this subset is referred to as the *stability region*.

Definition 12.29. A linear MOL (12.28) is said to have *strong stability* if

$$\forall h \in \mathbb{R}^+, \quad \|B_h\|_2 \le 1. \tag{12.30}$$

Corollary 12.30. The Crank-Nicolson method has strong stability with

$$B = \Big(I - \frac{k}{2}A\Big)^{-1}\Big(I + \frac{k}{2}A\Big). \tag{12.31}$$

Proof. (12.31) follows directly from Exercise 12.11. The symmetry of A implies the symmetry of B and thus the 2-norm of B equals its spectral radius:

$$\|B\|_2 = \rho(B) = \max\left|\frac{1 + k\lambda_p/2}{1 - k\lambda_p/2}\right| \le 1.$$

Then the proof is completed by Definition 12.29. □

Remark 12.19. Strong stability implies Lax-Richtmyer stability, but the converse is false. If there is a constant α so that a bound of the form

$$\|B_h(k)\| \le 1 + \alpha k,$$

then we will have Lax-Richtmyer stability in this norm, since

$$\|B_h^n(k)\| \le (1 + \alpha k)^n \le e^{\alpha T}$$

for $nk < T$. This is similar to the fact that A-stability is sufficient but not necessary for the convergence of a consistent IVP solver.

12.1.5 Convergence

Definition 12.31. The *solution error* of an MOL is

$$E_i^n = U_i^n - u(x_i, t_n), \tag{12.32}$$

where $u(x_i, t_n)$ is the exact solution of the PDE at the grid point (x_i, t_n).

Remark 12.20. We have chosen to define the LTE in a way such that the solution error and the LTE are of the same order. For a given desired accuracy ϵ, we often determine k and h from ϵ and the LTE. For example, it is natural to have $k = O(\epsilon^{\frac{1}{2}})$ and $h = O(\epsilon^{\frac{1}{2}})$ for the Crank-Nicolson method. On the other hand, $k = O(\epsilon)$ has to be used for the FTCS method. Since its temporal accuracy and its spatial accuracy are of different orders, a method like FTCS is not a good choice for practical use.

Definition 12.32. An MOL is *convergent* iff

$$\forall T > 0, \qquad \lim_{k\to 0, h\to 0; kN=T} \left\|E^N\right\| = 0. \tag{12.33}$$

Remark 12.21. Lemma 12.33 is a stepping stone for proving the necessity part of the Lax equivalence theorem 12.34.

Lemma 12.33. Let $\{E_i\}_{i=1}^{\infty}$ be a sequence of Banach spaces. Define a new space E, a projection $P_j : E \to E_j$, and an embedding operator $I_j : E_j \to E$ as follows,

$$E := \bigoplus_{j=1}^{\infty} E_j = \left\{ (x_i)_{i\in\mathbb{N}^+} : x_i \in E_i,\ \sum_{i=1}^{\infty} \|x_i\| < +\infty \right\}, \tag{12.34}$$

$$\begin{cases} \forall (x_i) \in E, & P_j((x_i)) = x_j, \\ \forall x \in E_j, & I_j(x) = \overline{x_j} := (0, \cdots, 0, \underset{j}{x}, 0, \cdots). \end{cases} \tag{12.35}$$

Then we have

(a) $(E, \|\cdot\|)$ is a Banach space where

$$\forall (x_i)_{i\in\mathbb{N}^+} \in E, \quad \|(x_i)_{i\in\mathbb{N}^+}\| := \sum_{i=1}^{\infty} \|x_i\|; \tag{12.36}$$

(b) $\forall j \in \mathbb{N}^+$, both P_j and E_j are continuous linear maps, i.e., $P_j \in \mathcal{CL}(E, E_j)$, $I_j \in \mathcal{CL}(E_j, E)$, c.f. Section E.2.1;

(c) $\|P_j\| = \|I_j\| = 1$ and $P_j I_j$ is the identity map on E_j;

(d) $\forall T \in \mathcal{CL}(E_j, E_j)$, the linear operator $\overline{T} := I_j T P_j$ is a norm-preserving extension of T, i.e., $\overline{T} \in \mathcal{CL}(E, E)$ and

$$\forall n \in \mathbb{N}^+, \quad \left\|\overline{T}^n\right\| = \|T^n\|. \tag{12.37}$$

Proof. (a) follows from the condition of each E_j being a Banach space, the construction (12.34), and Definition E.82.

(b) follows from (12.35) and Theorem E.96.

(c) holds trivially from (12.35).

For (d), we consider an arbitrary element $(x_i) \in E$,

$$(*): \quad \begin{aligned} &\left\|\overline{T}^n((x_i))\right\| = \|I_jT^nP_j((x_i))\| = \|I_jT^n(x_j)\| \\ &= \|(0,\ldots,0,T^n(x_j),0,\ldots,0)\| = \|T^n(x_j)\|, \end{aligned}$$

where the first equality follows from (c), the second and third from (12.35), and the fourth from (12.36). In $(*)$, the range of x_j covers E_j while that of (x_i) covers E. Then the proof is completed by taking supremum of $(*)$ and applying Lemma E.109. □

Remark 12.22. Caution: I_jP_j is not an identity map.

Theorem 12.34 (Lax equivalence theorem)**.** A consistent linear MOL (12.28) is convergent if and only if it is Lax-Richtmyer stable.

Proof. For the sufficiency, we apply the numerical method (12.28) to the exact solution $\hat{\mathbf{U}}^n$ to obtain

$$(*) \qquad \hat{\mathbf{U}}^{n+1} = B_h\hat{\mathbf{U}}^n + \mathbf{b}^n + k\tau^n,$$

where the dependence on k has been suppressed for clarity and where

$$\hat{\mathbf{U}}^n := \begin{bmatrix} u(x_1,t_n) \\ u(x_2,t_n) \\ \vdots \\ u(x_m,t_n) \end{bmatrix}, \quad \tau^n := \begin{bmatrix} \tau(x_1,t_n) \\ \tau(x_2,t_n) \\ \vdots \\ \tau(x_m,t_n) \end{bmatrix}.$$

Subtracting $(*)$ from (12.28) gives the difference equation for the global error $E^n = \mathbf{U}^n - \hat{\mathbf{U}}^n$:

$$E^{n+1} = B_hE^n - k\tau^n,$$

and hence, by induction,

$$E^N = B_h^NE^0 - k\sum_{n=1}^{N} B_h^{N-n}\tau^{n-1},$$

from which we have

$$(**): \quad \left\|E^N\right\| \le \left\|B_h^N\right\|\left\|E^0\right\| + k\sum_{n=1}^{N}\left\|B_h^{N-n}\right\|\left\|\tau^{n-1}\right\|.$$

If the MOL is Lax-Richtmyer stable, we have, for $Nk \le T$,

$$\left\|E^N\right\| \le C_T\left\|E^0\right\| + kN\cdot C_T\max_{1\le n\le N}\left\|\tau^{n-1}\right\|,$$

the RHS goes to 0 as $k \to 0$ and $h \to 0$.

As for the necessity, first we know from Definition 12.28 that it suffices to prove, for all sufficiently small $k, h \in \mathbb{R}^+$, the power boundedness of $B_h(k)$ for the homogeneous case, i.e., $\mathbf{b}_h(k) = \mathbf{0}$ in (12.28), since we have the same $B_h(k)$ for both homogeneous and non-homogeneous PDEs.

Second, we construct a space of all grid functions,

$$\mathcal{B} := \bigoplus_{m=1}^{+\infty}\mathbb{G}_{\frac{1}{m}}, \tag{12.38}$$

where "$\bigoplus$" is given by (12.34) and $\mathbb{G}_{\frac{1}{m}}$ is the Banach space of grid functions on the grid $\mathbf{X}$ with size $h = \frac{1}{m}$,

$$\mathbb{G}_h := \{U : \mathbf{X} \to \mathbb{R} \mid \|U\|_q < +\infty\}, \tag{12.39}$$

and $\|\cdot\|_q$ is a q-norm in Definition 7.13. By Lemma 12.33, $(\mathcal{B}, \|\cdot\|)$ with $\|\cdot\|$ given in (12.36) is a Banach space. Since the PDE is linear, we can assume WLOG that the spatial domain is unit and each h equals $\frac{1}{m}$ for some $m \in \mathbb{N}^+$.

Note that our main strategy is to deduce stability from convergence via the principle of uniform boundedness (Theorem E.148). More precisely, for the family of linear maps

$$\begin{aligned} &\mathcal{T} := \left\{\overline{B_h(k)}^n : (h,k,n) \in I_{\mathcal{T}}\right\}, \\ &I_{\mathcal{T}} := \{(h,k,n) : kn \le T, h \in (0,h_0), k \in (0,k_0)\}, \end{aligned}$$

where $\overline{B_h(k)} : \mathcal{B} \to \mathcal{B}$ is the extension of $B_h(k)$ as defined in Lemma 12.33(d) and the constants h_0, k_0 come from Definition 12.32 (see below), we want to show

$$(\square): \ \forall \eta \in \mathcal{B},\ \exists M_\eta > 0 \text{ s.t. } \forall \overline{B_h(k)}^n \in \mathcal{T},\ \left\|\overline{B_h(k)}^n(\eta)\right\| \le M_\eta.$$

Then the principle of uniform boundedness and Lemma 12.33(d) would imply stability.

Third, the convergence of the MOL and Definition 12.32 imply that we can pick some fixed $\epsilon > 0$ such that

$$\begin{aligned} &\exists k_0, h_0 > 0 \text{ s.t. } \forall k \in (0,k_0], \forall h \in (0,h_0], \forall n \in \mathbb{N}^+, \\ &t := kn \in [0,T] \implies \|E(t)\| = \|E^n\| < \epsilon. \end{aligned}$$

Finally, we prove $(\square)$ by contradiction. The negation of $(\square)$ yields

$$(\triangle): \quad \exists \eta \in \mathcal{B}, \quad (h_j)_{j\in\mathbb{N}}, (k_j)_{j\in\mathbb{N}}, (n_j)_{j\in\mathbb{N}} \text{ s.t.} \\ \begin{cases} \lim_{j\to\infty} n_jk_j = t' \in [0,T]; \\ \lim_{j\to\infty} \left\|\overline{B_{h_j}(k_j)}^{n_j}(\eta)\right\| = +\infty. \end{cases}$$

By construction of $\overline{B_h(k)}$, we have

$$h_1 \ne h_2 \implies \overline{B_{h_1}(k)}\left(\overline{U^0_{h_2}}\right) = \mathbf{0} := (0,0,\ldots).$$

Hence $\mathcal{B}\backslash\overline{\mathbb{G}_{h_1}}$ is a subset of the null space of $\overline{B_{h_1}(k)}$, leading to $\overline{B_{h_j}(k_j)}^{n_j}(\eta) = \overline{B_{h_j}(k_j)^{n_j}U^0_{h_j}}$ where $U^0_{h_j}$ is the component of η for grid size h_j. Let $u(t')$ be the exact solution at time $t = t'$. Then we have

$$\begin{aligned} \forall j \in \mathbb{N}^+, \quad &\left\|\overline{B_{h_j}(k_j)}^{n_j}(\eta)\right\| = \left\|\overline{B_{h_j}(k_j)^{n_j}U^0_{h_j}}\right\| \\ \le \quad &\|u(t')\| + \epsilon \le C\|U^0_{h_j}\| + \epsilon \le C\|\eta\| + \epsilon, \end{aligned}$$

where the second step follows from $\|E(t')\| < \epsilon$, the third from Duhamel's principle (Theorem 11.51), and the last from (12.36). This contradicts $(\triangle)$ and thus $(\square)$ holds. □

Remark 12.23. The principle of uniform boundedness is mainly used to remove the dependence of the bound of the operator norms on h. This actually points to a main difference between the stability analysis of PDEs and that of ODEs; see Remark 12.24.

Corollary 12.35. In solving the heat equation (12.3), the θ-method is convergent for any $\theta \in [0,1]$.

Proof. For the θ-method, we have

$$\begin{aligned} B &= (I - \theta kA)^{-1}[I + (1-\theta)kA], \\ \rho(B) &= \max \left| \tfrac{1+(1-\theta)k\lambda_p}{1-\theta k\lambda_p} \right| \le 1, \end{aligned}$$

where the inequality holds for k sufficiently small. The rest follows from Theorem 12.34 and Lemma 12.18. □

Remark 12.24. There are a number of differences in the stability theories for PDE and ODE. First, it is more difficult to prove the absolute stability of an MOL than to obtain that of an ODE solver. For any given ODE, the number of equations in the system is fixed, hence λ is fixed and the correct behavior of an ODE solver at $k\lambda \to 0$ is all we need for convergence. In comparison, 0 being in the RAS is not enough for MOLs in solving a PDE, due to the fact that, as $h \to 0$, both the spectral radius the dimension of the semi-discrete system go to infinity. (This is one reason that we need functional analysis to deal with infinite-dimensional linear spaces!) In Example 12.7, some eigenvalues of A satisfies $\lambda = O(\frac{1}{h^2})$. As $h \to 0$, we have $\lambda \to -\infty$, and the choice of k for FTCS must be small enough so that $k\lambda$ stay in the RAS of the Euler method.

Second, for solving an ODE we often assume Lipschitz continuity of the RHS function f in terms of the evolutionary variable. This Lipschitz continuity can rarely be satisfied in the case of MOLs. For the heat equation, the semi-discrete system is an ODE of the form $f(U) = AU$, but Remark 12.16 implies that the Lipschitz constant is $\|A\| = O(\frac{1}{h^2})$ and grows to infinity as $h \to 0$.

Finally, the stability theory of ODE applies to both linear and nonlinear IVPs while the Lax equivalence theorem is much more limited: it holds only for linear PDEs, i.e., those for which the semi-discrete system of MOL is of the form (12.28). The core difficulty of a general stability theory for nonlinear PDEs is the lack of universal bounds to prove convergence. Consequently, the stability theory for nonlinear PDEs can only be on a case-by-case basis.

12.1.6 Discrete maximum principle

Theorem 12.36 (Maximum principle of the heat equation)**.** Suppose $u(x,t)$ satisfies the heat equation (12.3) in $\Omega := (0,1) \times (0,T)$ and is continuous in $\overline{\Omega} := [0,1] \times [0,T]$. Then both the maximum and the minimum of $u(x,t)$ over $\overline{\Omega}$ are assumed either initially at $t = 0$ or on the boundary $x = 0$ or $x = 1$. More precisely, define

$$\Gamma_\Omega := \left\{(x,t) \in \overline{\Omega} : \ t = 0 \ \text{ or } \ x = 0 \ \text{ or } \ x = 1\right\} \tag{12.40}$$

and we have

$$\max_{(x,t)\in\overline{\Omega}} \{u(x,t)\} = \max_{(x,t)\in\Gamma_\Omega} \{u(x,t)\}, \tag{12.41a}$$

$$\min_{(x,t)\in\overline{\Omega}} \{u(x,t)\} = \min_{(x,t)\in\Gamma_\Omega} \{u(x,t)\}. \tag{12.41b}$$

Remark 12.25. Sometimes it is important to maintain the maximum principle in numerical results. For example, it looks nonphysical when the computed solutions for the concentration of some pollutant contain some negative numbers. The following theorem states that a θ-method can fulfill the maximum principle in the discrete sense.

Theorem 12.37 (Discrete maximum principle)**.** For the heat equation (12.3) in Ω, a θ-method with $\theta \in [0,1]$ and $2r(1-\theta) \le 1$ satisfies

$$\min \mathcal{U}_\Gamma \le U_j^n \le \max \mathcal{U}_\Gamma \tag{12.42}$$

where U_j^n is a solution of the θ-method at $(x_j, t_n) \in \Omega$ and the set of values of U at the boundary points is

$$\begin{aligned} \mathcal{U}_\Gamma := \quad & \{U_0^n : n = 0,1,\cdots,N\} \cup \left\{U_{m+1}^n : n = 0,1,\cdots,N\right\} \\ & \cup \left\{U_j^0 : j = 0,1,\cdots,m+1\right\}. \end{aligned}$$

Proof. Suppose there exists an internal point $U_j^{n+1} \notin \Gamma_\Omega$, c.f. (12.40), such that U_j^{n+1} is a *global* maximum and satisfies $U_j^{n+1} > \max \mathcal{U}_\Gamma$. Then for U_j^{n+1} we write (12.22) as

$$\begin{aligned} (1+2\theta r)U_j^{n+1} = {} & \theta r(U_{j-1}^{n+1} + U_{j+1}^{n+1}) \\ & + (1-\theta)r(U_{j-1}^n + U_{j+1}^n) \\ & + [1 - 2(1-\theta)r]U_j^n, \end{aligned} \tag{12.43}$$

where all coefficients of the five values of U on the RHS are nonnegative and sum to $1 + 2\theta r$. Let U^* be the *local* maximum of the five values of U in the stencil of (x_j, t_{n+1}) on the RHS of (12.43). Then we have

$$\begin{aligned} (1+2\theta r)U_j^{n+1} \le \quad & \theta r(U^* + U^*) + (1-\theta)r(U^* + U^*) \\ & + [1 - 2(1-\theta)r]U^* = (1+2\theta r)U^*, \end{aligned}$$

i.e., $U_j^{n+1} \le U^*$. But since U_j^{n+1} is the global maximum, we also have $U_j^{n+1} \ge U^*$, and thus $U^* = U_j^{n+1}$. However, in order for the inequality "$\le$" to be "$=$" so that $U^* = U_j^{n+1}$ holds, we must have

$$U_{j-1}^{n+1} = U_{j+1}^{n+1} = U_{j-1}^n = U_{j+1}^n = U_j^n = U^* = U_j^{n+1}.$$

Now that each of the five values of U is also the global maximum, repeat the above arguments and we know that all points in the stencils of the five locations must also have the global maximum of U. Since the initial condition is always a Dirichlet condition and the discrete grid is always connected along the positive direction of the time axis, the values of U must be the same for all interior points and all boundary points. This contradicts the starting assumption.

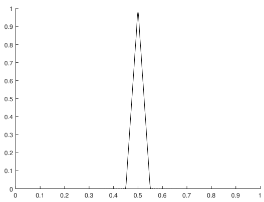
The first inequality in (12.42) can be proven by arguments similar to those in the above paragraphs. □

Remark 12.26. The proof of Theorem 12.37 is very similar to that of Lemma 7.57. One common key point is the condition of all coefficients being nonnegative.

Remark 12.27. The condition for the discrete maximum principle, $r(1-\theta) \le \frac{1}{2}$, is more restrictive than that needed in stability, $r(1-2\theta) \le \frac{1}{2}$. For example, the Crank-Nicolson method always satisfies the stability condition, but only if $r \le 1$ does it satisfy the condition for the discrete maximum principle.

Example 12.38. Consider the model problem (12.3) with $\nu = 1$, the boundary conditions $u(0,t) = u(1,t) = 0$, and the initial condition

$$u(x,0) = \varphi(x) = \begin{cases} 20(x - \frac{9}{20}), & \frac{9}{20} \le x < \frac{1}{2}, \\ -20(x - \frac{11}{20}), & \frac{1}{2} \le x < \frac{11}{20}, \\ 0, & \text{otherwise.} \end{cases}$$



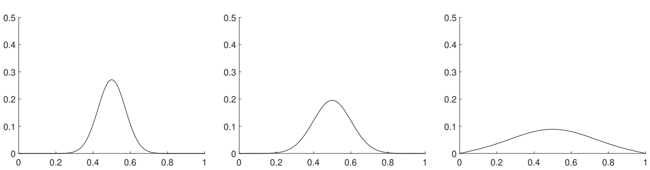
The method of separation of variables yields

$$u(t,x) = \sum_{k=1}^{\infty} A_k e^{-k^2\pi^2 t} \sin(k\pi x) \tag{12.44}$$

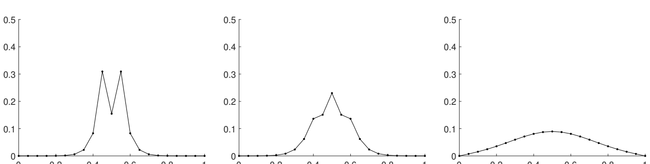
where

$$\begin{aligned} A_k &= 2\int_0^1 \varphi(\xi)\sin(k\pi\xi)\mathrm{d}\xi \\ &= \frac{40}{k^2\pi^2}\left[-\sin\left(\frac{9}{20}k\pi\right) + 2\sin\left(\frac{1}{2}k\pi\right) - \sin\left(\frac{11}{20}k\pi\right)\right]. \end{aligned}$$

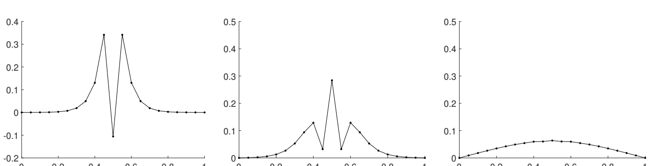
In the rest of this example we fix $h = \frac{1}{20}$. The following plots represent the exact solution (12.44) with $r = 1$ at $t = k$, $t = 2k$ and $t = 10k$, respectively.



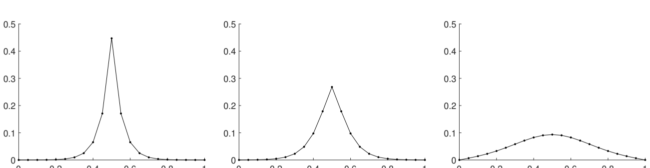
After 1, 2, and 10 time steps, the Crank-Nicolson method with $r = 1$ gives results as follows.



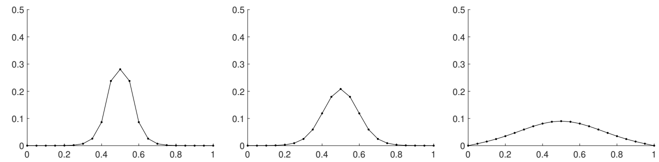
Crank-Nicolson with $r = 2$ gives results as follows.



The BTCS with $r = 1$ gives results as follows.



The collocation method in Example 11.258 with $r = 1$ gives results as follows.



For $r = 1$, Crank-Nicolson preserves the discrete maximum principle of the heat equation; but for $r = 2$, results of Crank-Nicolson violates the discrete maximum principle. This illustrates Theorem 12.37. The oscillation in the results of Crank-Nicolson after one time step is probably due to the fact that it is not L-stable. When we switch to the L-stable methods, either the BTCS or the collocation method in Example 11.258, the oscillations disappear and the monotonicity is preserved.

Remark 12.28. In general, the B-stability discussed in Section 11.6.9 does not help the discrete maximum principle that is related to the max-norm, because the B-stability hinges on a norm deduced from an inner product. By Theorem B.165 and Exercise B.164, the max-norm cannot be induced from an inner product.

However, we can deduce the second inequality in (12.42) for the special case in Example 12.38 simply because the initial condition contains only a single non-zero data in the middle of the domain, in which case the 2-norm of the initial condition is the maximum value 1.

12.1.7 Von Neumann stability

Remark 12.29. The MOL formulation relates stability theory for PDEs to that for ODEs. In comparison, the von Neumann analysis sheds physical insights on the proper stability restrictions in a way that is much more straightforward. In the rest of this chapter we shall assume that the reader is familiar with the materials presented in Appendix F.

Definition 12.39. An MOL for a first-order equation (F.43) is *von Neumann stable* iff

$$\begin{aligned} &\forall T > 0,\ \exists h_0, k_0, C_T > 0,\ \exists S \in \mathbb{N} \text{ s.t.} \\ &\forall k \in (0, k_0], \forall h \in (0, h_0], \forall n \in \mathbb{N}^+, \\ &nk \le T \implies \|\mathbf{U}^n\|_{h,2}^2 \le C_T \textstyle\sum_{i=0}^{S} \|\mathbf{U}^i\|_{h,2}^2, \end{aligned} \tag{12.45}$$

where C_T is a constant depending only on T and $\|\mathbf{U}^n\|_{h,2}$ is the 2-norm of the grid function $\mathbf{U}^n$ in Definition 7.13.

More specifically, the MOL is *von Neumann stable under the constraint* $\mathbf{g}(k,h) \le \mathbf{0}$ iff (12.45) holds with its third line replaced with

$$\mathbf{g}(k,h) \le \mathbf{0},\ nk \le T \implies \|\mathbf{U}^n\|_{h,2}^2 \le C_T \textstyle\sum_{i=0}^{S} \|\mathbf{U}^i\|_{h,2}^2,$$

where $\mathbf{g}(k,h) \le \mathbf{0}$ means

$$\forall i = 1, 2, \ldots, m, \quad g_i(k,h) \le 0 \text{ or } g_i(k,h) < 0$$

and each $g_i(k,h)$ is an analytic function.

Remark 12.30. Essentially, Definition 12.39 identifies stability with boundedness. It suffices to have $S=0$ for MOLs with one-step time integration while it requires $S>0$ for MOLs with LMMs.

For $S=0$, Definition 12.39 can be regarded as a specialized version of the Lax-Richtmyer stability in Definition 12.28 in the case of 2-norms so that the techniques of Fourier analysis can be applied. For the comments on the contraints, see Remark 12.18.

Lemma 12.40. Define $h\mathbb{Z} := \{hj : j\in\mathbb{Z}\}$ and

$$L^2(h\mathbb{Z}) := \left\{ g : h\mathbb{Z}\to\mathbb{R} \;\middle|\; \sum_{i\in\mathbb{Z}} |g(ih)|^2 < \infty \right\}.$$

The Fourier transform of a function $U\in L^2(h\mathbb{Z})$ is the continuous function $\hat{U} : [-\frac{\pi}{h}, \frac{\pi}{h}] \to \mathbb{C}$ given by

$$\hat{U}(\xi) = (\mathcal{F}U)(\xi) := \frac{1}{\sqrt{2\pi}} \sum_{m\in\mathbb{Z}} e^{-\mathbf{i}mh\xi} U_m h \tag{12.46}$$

while its inverse Fourier transform is given by

$$U_m = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{+\frac{\pi}{h}} e^{\mathbf{i}mh\xi} \hat{U}(\xi)\mathrm{d}\xi. \tag{12.47}$$

Proof. This follows from Definition F.12 and a change of variable. □

Remark 12.31. By extending to $h\mathbb{Z}$ the domain of a grid function with uniform grid size h, we can regard the grid function as special elements in $L^2(h\mathbb{Z})$ with only a finite number of nonzero values. Thus the space $L^2(h\mathbb{Z})$ can be regarded as a natural generalization of grid functions; this generalization facilitates Fourier analysis of grid functions.

Exercise 12.41. Show that the grid function $\mathbf{U}\in L^2(h\mathbb{Z})$ is recovered by a Fourier transform followed by an inverse Fourier transform.

Theorem 12.42. Parseval's equality holds for grid functions in terms of $\|\cdot\|_{h,2}$, the 2-norm in Definition 7.13, i.e.,

$$\forall \mathbf{U}\in L^2(h\mathbb{Z}), \qquad \|\mathbf{U}\|_{h,2} = \left\|\hat{U}\right\|_2. \tag{12.48}$$

Proof. Lemma 12.40 yields

$$\begin{aligned}
&\left\|\hat{U}(\xi)\right\|_2^2 = \textstyle\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\hat{U}(\xi)|^2 \mathrm{d}\xi \\
=\ & \textstyle\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \overline{\hat{U}(\xi)} \frac{1}{\sqrt{2\pi}} \sum_{m\in\mathbb{Z}} e^{-\mathbf{i}mh\xi} U_m h \mathrm{d}\xi \\
=\ & \frac{h}{\sqrt{2\pi}} \overline{\textstyle\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \hat{U}(\xi) \sum_{m\in\mathbb{Z}} e^{\mathbf{i}mh\xi} U_m \mathrm{d}\xi} \\
=\ & \frac{h}{\sqrt{2\pi}} \textstyle\sum_{m\in\mathbb{Z}} U_m \overline{\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \hat{U}(\xi) e^{\mathbf{i}mh\xi} \mathrm{d}\xi} \\
=\ & h \textstyle\sum_{m\in\mathbb{Z}} U_m \overline{U_m} = h\sum_{m\in\mathbb{Z}} |U_m|^2 = \|\mathbf{U}\|_h^2.
\end{aligned}$$ □

Remark 12.32. In this subsection, we typeset a grid function in bold font to emphasize that it is a vector with a countable number of components. In contrast, the Fourier transform of a grid function is a scalar continuous function and hence should be typeset in normal font.

Example 12.43. By (12.46), the Fourier transform of the grid function

$$U_m = \begin{cases} 1 & \text{if } |x_m| < 1, \\ \frac{1}{2} & \text{if } |x_m| = 1, \\ 0 & \text{if } |x_m| > 1, \end{cases}$$

with grid spacing $h = M^{-1}$ and $M\in\mathbb{N}^+$ is

$$\begin{aligned}
\hat{U}(\xi) =\ & \frac{h}{2\sqrt{2\pi}} \left(e^{-\mathbf{i}Mh\xi} + e^{\mathbf{i}Mh\xi} + 2\textstyle\sum_{m=-(M-1)}^{M-1} e^{-\mathbf{i}mh\xi} \right) \\
=\ & \frac{h}{\sqrt{2\pi}} \cos(\xi) + \frac{h}{\sqrt{2\pi}} \frac{\sin((M-\frac{1}{2})h\xi)}{\sin(\frac{1}{2}h\xi)} \\
=\ & \frac{h}{\sqrt{2\pi}} \cos(\xi) + \frac{h}{\sqrt{2\pi}} \frac{\sin(Mh\xi)\cos(\frac{1}{2}h\xi)}{\sin(\frac{1}{2}h\xi)} \\
& -\frac{h}{\sqrt{2\pi}} \frac{\cos(Mh\xi)\sin(\frac{1}{2}h\xi)}{\sin(\frac{1}{2}h\xi)} \\
=\ & \frac{h}{\sqrt{2\pi}} \sin(\xi) \cot\left(\tfrac{1}{2}h\xi\right).
\end{aligned}$$

By Parseval's relation in Theorem 12.42, we must have

$$\begin{aligned}
2 - \tfrac{1}{2}h =\ & 2h\left(\tfrac{1}{2}\right)^2 + h\textstyle\sum_{m=-(M-1)}^{M-1} 1 \\
=\ & \frac{h^2}{2\pi} \textstyle\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \sin^2(\xi) \cot^2\left(\frac{1}{2}h\xi\right) \mathrm{d}\xi,
\end{aligned}$$

which can be verified by directly evaluating the integral.

Example 12.44. For any constant $\alpha\in\mathbb{R}^+$, the Fourier transform of the grid function $U_m = e^{-\alpha|m|h}$ is given by

$$\begin{aligned}
\hat{U}(\xi) =\ & \frac{1}{\sqrt{2\pi}} \textstyle\sum_{m\in\mathbb{Z}} e^{-\mathbf{i}mh\xi} e^{-\alpha|m|h} h \\
=\ & \frac{h}{\sqrt{2\pi}} \left(1 + \textstyle\sum_{m\in\mathbb{Z}\setminus\{0\}} e^{-\mathbf{i}mh\xi} e^{-\alpha|m|h}\right) \\
=\ & \frac{h}{\sqrt{2\pi}} \left(1 + \frac{e^{-(\alpha-\mathbf{i}\xi)h}}{1-e^{-(\alpha-\mathbf{i}\xi)h}} + \frac{e^{-(\alpha+\mathbf{i}\xi)h}}{1-e^{-(\alpha+\mathbf{i}\xi)h}}\right) \\
=\ & \frac{h}{\sqrt{2\pi}} \frac{1-e^{-2\alpha h}}{1-2e^{-\alpha h}\cos(h\xi)+e^{-2\alpha h}}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\|\mathbf{U}\|_h^2 =\ & h\frac{1+e^{-2\alpha h}}{1-e^{-2\alpha h}} \\
=\ & \frac{h^2}{2\pi} \textstyle\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \left(\frac{1-e^{-2\alpha h}}{1-2e^{-\alpha h}\cos(h\xi)+e^{-2\alpha h}}\right)^2 \mathrm{d}\xi = \left\|\hat{U}\right\|_2^2,
\end{aligned}$$

which verifies Parseval's relation in Theorem 12.42.

Corollary 12.45. An MOL for a first-order equation (F.43) is von Neumann stable iff

$$\begin{aligned}
& \forall T>0,\ \exists h_0, k_0, C_T > 0,\ \exists S\in\mathbb{N} \text{ s.t.} \\
& \forall k\in(0,k_0], \forall h\in(0,h_0], \forall n\in\mathbb{N}^+, \\
& nk\le T \implies \left\|\hat{U}^n\right\|_2 \le C_T \textstyle\sum_{i=0}^{S} \left\|\hat{U}^i\right\|_2,
\end{aligned} \tag{12.49}$$

where C_T is a constant depending only on T and $\hat{U}^n$ is the Fourier transform of the grid function $\mathbf{U}^n$ advanced by the MOL.

Proof. This follows directly from Definition 12.39 and Theorem 12.42. □

Example 12.46. Consider the FTCS method by the von Neumann analysis. Plug (12.47) into (12.15) and we have

$$\begin{aligned}
U_j^{n+1} &= \frac{1}{\sqrt{2\pi}} \textstyle\int_{-\frac{\pi}{h}}^{+\frac{\pi}{h}} \chi_j(h\xi) \hat{U}^n(\xi) \mathrm{d}\xi \\
&= \frac{1}{\sqrt{2\pi}} \textstyle\int_{-\frac{\pi}{h}}^{+\frac{\pi}{h}} e^{\mathbf{i}jh\xi} \hat{U}^{n+1}(\xi) \mathrm{d}\xi,
\end{aligned}$$

where $\chi_j(h\xi) := re^{\mathbf{i}(j-1)h\xi} + re^{\mathbf{i}(j+1)h\xi} + (1-2r)e^{\mathbf{i}jh\xi}$ characterizes the FTCS method. Since the Fourier transform is unique, we have

$$\hat{U}^{n+1}(\xi) = g(h\xi)\hat{U}^n(\xi), \tag{12.50}$$

where $g(\xi)$ is the *amplification factor* for the wavenumber ξ

$$g(h\xi) = 1 - 4r\sin^2\left(\frac{\xi h}{2}\right).$$

To guarantee $|g(h\xi)| \le 1$, we take $1 - \frac{4\nu k}{h^2} \ge -1$, which implies (12.27), i.e. $k \le \frac{h^2}{2\nu}$. Therefore, the FTCS method is von Neumann stable under the constraint $k - \frac{h^2}{2\nu} \le 0$, but not so for $k = \Theta(h)$.

Theorem 12.47. An MOL with a one-step time integrator and constant coefficients is von Neumann stable if and only if

$$\begin{array}{l}\exists C, k_0, h_0 \in \mathbb{R}^+ \text{ s.t. } \forall h\xi \in [-\pi, \pi], k \in (0, k_0], h \in (0, h_0], \\ |g(h\xi, k, h)| \le 1 + Ck,\end{array} \tag{12.51}$$

where the constant C is independent of ξ, k, and h.

Proof. Parseval's equality in Theorem 12.42 and the definition of g in (12.50) yield

$$\|\mathbf{U}^n\|_h^2 = \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |g(h\xi, k, h)|^{2n} \left|\hat{U}^0(\xi)\right|^2 \mathrm{d}\xi.$$

If $|g(h\xi, k, h)| \le 1 + Ck$, we have

$$\begin{aligned}\|\mathbf{U}^n\|_h^2 &\le \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} (1+Ck)^{2n} \left|\hat{U}^0(\xi)\right|^2 \mathrm{d}\xi \\ &= (1+Ck)^{2n} \left\|\mathbf{U}^0\right\|_h^2.\end{aligned}$$

Then $n \le \frac{T}{k}$ yields

$$(1+Ck)^n \le (1+Ck)^{\frac{T}{k}} \le e^{CT}.$$

Hence $\|U^n\|_h \le e^{CT}\|U^0\|_h$ and the scheme is stable.

We now prove that if the inequality in (12.51) does not hold for any $C \in \mathbb{R}^+$, then the scheme is not stable. To do this we show that we can achieve any amount of growth in the solution, that is, we show that the stability inequality in Definition 12.39 cannot hold.

Suppose for any $C \in \mathbb{R}^+$ there are constants $h_0, k_0 \in \mathbb{R}^+$ and an interval $[\theta_1, \theta_2]$ such that $\xi h \in [\theta_1, \theta_2]$, $h \in (0, h_0]$, and $k \in (0, k_0]$ imply $|g(\xi h, k, h)| \ge 1 + Ck$. Then we construct a function U_m^0 as

$$\hat{U}^0(\xi) = \begin{cases} 0 & \text{if } h\xi \notin [\theta_1, \theta_2], \\ \sqrt{h(\theta_2 - \theta_1)^{-1}} & \text{if } h\xi \in [\theta_1, \theta_2]. \end{cases}$$

Then $\left\|\mathbf{U}^0\right\|_h = \left\|\hat{U}^0\right\|_2 = 1$ and

$$\begin{aligned}\|\mathbf{U}^n\|_h^2 &= \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |g(h\xi, k, h)|^{2n} \left|\hat{U}^0(\xi)\right|^2 \mathrm{d}\xi \\ &= \int_{\frac{\theta_1}{h}}^{\frac{\theta_2}{h}} |g(h\xi, k, h)|^{2n} \frac{h}{\theta_2 - \theta_1} \mathrm{d}\xi \\ &\ge (1+Ck)^{2n} \ge \frac{1}{2}e^{2TC} \left\|\mathbf{U}^0\right\|_h^2,\end{aligned}$$

for n near $\frac{T}{k}$. Hence the scheme is unstable. $\square$

Remark 12.33. By Theorem F.32, the exact solution of the heat equation with homogeneous Dirichlet condition has an L_2 norm that is never increasing. In light of this physics, a numerical scheme should be energy stable. In the MOL approach, we can employ a B-stable ODE solver to show

$$\left\|\mathbf{U}^{n+1}\right\|_2 \le \|\mathbf{U}^n\|_2$$

because $V_j^n \equiv 0$ is a solution of the semi-discrete system (12.10) with homogeneous Dirichlet boundary conditions. From the viewpoint of Fourier modes, Theorem 12.5 decouples the energy of different Fourier modes and the exponential decay implies

$$\left\|\hat{U}^{n+1}\right\|_2 \le \left\|\hat{U}^n\right\|_2$$

and thus $\left\|\mathbf{U}^{n+1}\right\|_2 \le \|\mathbf{U}^n\|_2$.

Remark 12.34. From a mathematical viewpoint, the basic reason that von Neumann analysis works is the fact that the functions $e^{\mathbf{i}\xi x}$ with constant wavenumber ξ are eigenfunctions of the differentiation operator ∂_x:

$$\partial_x e^{\mathbf{i}\xi x} = \mathbf{i}\xi e^{\mathbf{i}\xi x}.$$

Exercise 12.48. Prove Lemma 12.25 via Von Neumann analysis. What can you say after comparing this proof with that for Exercise 12.26?

Remark 12.35. We have shown that the Crank-Nicolson method could be an excellent choice for a stiff problem such as the heat equation. When diffusion is coupled with other processes such as convection, the high-frequency oscillations caused by convection should be damped by diffusion. However, the Crank-Nicolson method is not a good choice for this type of problems because it cannot smooth out these oscillations effectively. This feature of the Crank-Nicolson method is often referred to as being "neutrally stable." To sum up, it is better to use an L-stable method for stiff problems with persistent high-frequency oscillations; this has already been illustrated in Example 12.38.

Remark 12.36. The von Neumann analysis of stability does not apply directly to problems with variable coefficients, i.e., when ν depends on x and t. In this case the reader is referred to Levy and Tadmor [1998].

Remark 12.37. The Lax equivalence theorem can also be proved by Fourier analysis; see Strikwerda [1989].

12.2 MOL for the advection equation

Definition 12.49. A second-order, constant-coefficient, linear partial differential equation (PDE) of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \tag{12.52}$$

is called a *hyperbolic PDE* if its coefficients satisfy

$$B^2 - 4AC > 0. \tag{12.53}$$

Definition 12.50. The *one-dimensional wave equation* is a hyperbolic PDE of the form

$$u_{tt} = a^2 u_{xx}, \tag{12.54}$$

where $a > 0$ is the *wave speed.*

Remark 12.38. The essential feature of (12.54) is captured in

$$u_t = -au_x$$

since $u_t = -au_x$ implies $u_{tt} = -au_{xt} = -au_{tx} = a^2u_{xx}$. This leads to the following definition.

Definition 12.51. The *one-dimensional advection equation* is

$$u_t = -au_x \text{ in } \Omega := (0,1)\times(0,T), \tag{12.55}$$

where $x \in (0,1)$ is the spatial location and $t \in (0,T)$ the time; the equation has to be supplemented with an *initial condition*

$$u(x,0) = \eta(x), \text{ on } (0,1)\times\{0\} \tag{12.56}$$

and appropriate boundary conditions at either $\{0\}\times(0,T)$ or $\{1\}\times(0,T)$, depending on the sign of a.

Theorem 12.52. The exact solution of the Cauchy problem (12.55) is

$$u(x,t) = \eta(x - at). \tag{12.57}$$

Proof. It is straightforward to verify that

$$u_t + au_x = -a\eta'(x-at) + a\eta'(x-at) = 0. \qquad \square$$

Remark 12.39. For the PDE (12.55), $u(x,t)$ depends on the frame (x,t). However, the PDE can be more simply described by one variable $\xi = \xi(x,t) = x - at$ in the sense that u only depends on ξ. In this case we refer to a level curve $\xi(x,t) = c \in \mathbb{R}$ as a *characteristic* of the PDE. For example, the characteristic curves for both PDEs (12.55) and (12.58) are the straight lines $x - at = c \in \mathbb{R}$.

Lemma 12.53. The hyperbolic equation

$$u_t + au_x = f(x,t) - bu, \quad u(x,0) = u_0(x) \tag{12.58}$$

with $a, b \in \mathbb{R}^+$ is solved by

$$u(x,t) = u_0(x-at)e^{-bt} + \int_0^t f(x - a(t-s), s)e^{-b(t-s)}\mathrm{d}s. \tag{12.59}$$

Proof. It is straightforward to verify that (12.59) is indeed the solution of (12.58). We give a derivation below. For the following coordinate transformation and its inverse,

$$\begin{aligned} \tau = t, &\quad \xi = x - at, \\ t = \tau, &\quad x = \xi + a\tau, \end{aligned}$$

we write $\tilde{u}(\xi,\tau) = u(x,t)$ to indicate that $\tilde{u}$ is expressed in the new coordinates (ξ,τ). Then the chain rule yields

$$\frac{\partial \tilde{u}}{\partial \tau} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\partial x}{\partial \tau} = \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x},$$

which gives an ODE on $\tilde{u}$:

$$\frac{\mathrm{d}\tilde{u}}{\mathrm{d}\tau} = -b\tilde{u} + f(\xi + a\tau, \tau). \tag{12.60}$$

By Duhamel's principle (Theorem 11.51), we have

$$\tilde{u}(\xi,\tau) = u_0(\xi)e^{-b\tau} + \int_0^\tau f(\xi + a\sigma, \sigma)e^{-b(\tau-\sigma)}\mathrm{d}\sigma.$$

Then the inverse transformation yields (12.59). $\square$

Remark 12.40. The idea behind the proof of Lemma 12.53 is to find a new space-time frame where the PDE (12.60) reduces to the simpler ODE (12.60) which we already know how to solve. For a variable speed $a(x,t)$, the characteristic curve is a general function of x and t.

Lemma 12.54. The hyperbolic equation

$$u_t + a(x,t)u_x = f(x,t,u), \quad u(x,0) = u_0(x) \tag{12.61}$$

is equivalent to the ODE system

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{bmatrix} x \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} a(x(\tau),\tau) \\ f(x(\tau),\tau,\tilde{u}) \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ \tilde{u}(\xi,0) \end{bmatrix} = \begin{bmatrix} \xi \\ u_0(\xi) \end{bmatrix}, \tag{12.62}$$

where $\tau = t$ and $\tilde{u} = \tilde{u}(\xi,\tau)$.

Proof. For the new frame (ξ,τ), we select ξ as the initial value of the ODE

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = a(x(\tau),\tau),$$

i.e., $x(0) = \xi$. Then the chain rule and the coordinate transformation $\tilde{u}(\xi,\tau) = u(x,t)$ yield

$$\frac{\partial \tilde{u}}{\partial \tau} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\partial x}{\partial \tau} = \frac{\partial u}{\partial t} + a(x(\tau),\tau)\frac{\partial u}{\partial x}.$$

Then (12.61) gives another ODE

$$\frac{\mathrm{d}\tilde{u}}{\mathrm{d}\tau} = f(x(\tau),\tau,\tilde{u}),$$

which completes the proof. $\square$

Remark 12.41. Once again, the PDE is converted to an ODE system; however, we have one more ODE to accomodate the variable speed. After all, an explicit solution of the PDE is not possible since the form of $a(x,t)$ is unknown.

The case that the characteristic speed $a(x,t,u)$ depends on u requires special treatments and is out of the scope of this book.

Example 12.55. By Lemma 12.54, the PDE

$$u_t + xu_x = \alpha u, \quad u(x,0) = \begin{cases} 1 & \text{if } x \in [0,1]; \\ 0 & \text{otherwise} \end{cases} \tag{12.63}$$

with $-\alpha \in \mathbb{R}^+$ is solved by

$$u(x,t) = \begin{cases} e^{\alpha t} & \text{if } x \in [0, e^t]; \\ 0 & \text{otherwise.} \end{cases}$$

Definition 12.56. A system of PDEs

$$\mathbf{u}_t + A(x,t)\mathbf{u}_x + B(x,t)\mathbf{u} = F(x,t) \tag{12.64}$$

with $\mathbf{u}(x,0) = \mathbf{u}_0(x)$ is *hyperbolic* if there is a matrix function $P(x,t)$ such that $P(x,t)AP(x,t)^{-1} = \Lambda(x,t)$ is diagonal with real eigenvalues and the matrix norms of $P(x,t)$ and $P(x,t)^{-1}$ are bounded for all $(x,t) \in \mathbb{R} \times [0,+\infty)$.

Remark 12.42. The characteristic curves for (12.64) are solutions to the ODEs

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \lambda_i(x,t), \quad y_i(0) = \xi_i$$

where λ_i is the ith diagonal of $\Lambda(x,t)$ in Definition 12.56. Define $\mathbf{v} = P\mathbf{u}$ and we have the ODEs

$$\mathbf{v}_t + \Lambda\mathbf{v}_x = G(x,t)\mathbf{v} + P(x,t)F(x,t),$$

where $G(x,t) = (P_t + \Lambda P_x - PB)P^{-1}$.

Example 12.57. The Euler equations are

$$\frac{\partial}{\partial t}\begin{bmatrix} p \\ u \end{bmatrix} + \begin{bmatrix} 0 & \kappa_0 \\ \frac{1}{\rho_0} & 0 \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} p \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{12.65}$$

The equation for the pressure p can be further written as

$$p_{tt} = a^2 p_{xx} \text{ with } a = \pm\sqrt{\kappa_0/\rho_0}.$$

Remark 12.43. For hyperbolic PDEs on a finite interval instead of the whole real line, we have to impose boundary conditions on some of the boundary points. The resulting problem is thus an initial-boundary value problem.

For the model problem in Definition 12.51 with $a > 0$, the characteristic line goes from left to right and thus a boundary condition has to be specified at $x = 0$ while no boundary condition should be specified at $x = 1$. For $u(0,t) = g(t)$, the solution of (12.55) is given by

$$u(x,t) = \begin{cases} u_0(x-at) & \text{if } x - at \ge 0, \\ g(t - a^{-1}x) & \text{if } x - at \le 0. \end{cases}$$

In particular, we should have $g(0) = u_0(0)$ along the characteristic line $x - at = 0$.

For a hyperbolic system such as (12.64), we need to classify the characteristics into *incoming characteristic*s (characteristics that enter the domain at a boundary point) and *outgoing characteristic*s (characteristics that leaves the domain). In order for a hyperbolic intial-boundary value problem to be well-posed, the number of boundary conditions must equal that of incoming characteristics; see [Strikwerda, 1989, p 9] for a simple example.

In this chapter, we confine ourselves to periodic domains to avoid the issue of boundary conditions.

12.2.1 Classical MOLs for the advection equation

Example 12.58. Discretize the advection equation (12.55) in space at grid point x_j by

$$U_j'(t) = -\frac{a}{2h}\left(U_{j+1}(t) - U_{j-1}(t)\right), \quad 2 \le j \le m, \tag{12.66}$$

where $U_j(t) \approx u(x_j,t)$ for $j = 1, 2, \cdots, m+1$. For periodic boundary conditions we have

$$u(0,t) = u(1,t), \tag{12.67}$$

thus the discretizations of (12.55) at $j = 1$ and $j = m+1$ are

$$U_1'(t) = -\frac{a}{2h}\left(U_2(t) - U_{m+1}(t)\right), \tag{12.68}$$

$$U_{m+1}'(t) = -\frac{a}{2h}\left(U_1(t) - U_m(t)\right). \tag{12.69}$$

Then the semi-discrete system can be written as

$$\mathbf{U}'(t) = A\mathbf{U}(t), \tag{12.70}$$

where

$$A = -\frac{a}{2h}\begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{bmatrix}, \tag{12.71}$$

and $\mathbf{U}(t) = [U_1(t), U_2(t), \cdots, U_{m+1}(t)]^T$.

Remark 12.44. Although the matrix A in (12.71) is not symmetric, it is anti-symmetric and is a normal operator as in Definition B.199.

Lemma 12.59. The eigenvalues of A in (12.70) are

$$\lambda_p = -\frac{\mathbf{i}a}{h}\sin(2\pi ph) \text{ for } p = 1, 2, \ldots, m+1. \tag{12.72}$$

The corresponding eigenvector $\mathbf{w}^p$ has components

$$w_j^p = e^{2\pi\mathbf{i}pjh} \text{ for } j = 1, 2, \ldots, m+1. \tag{12.73}$$

Proof. For $j = 2, 3, \ldots, m$, we have

$$\begin{aligned}(A\mathbf{w}^p)_j &= -\tfrac{a}{2h}\left(w_{j+1}^p - w_{j-1}^p\right) \\ &= -\tfrac{a}{2h}e^{2\pi\mathbf{i}pjh}\left(e^{2\pi\mathbf{i}ph} - e^{-2\pi\mathbf{i}ph}\right) \\ &= -\tfrac{\mathbf{i}a}{h}\sin(2\pi ph)e^{2\pi\mathbf{i}pjh} \\ &= \lambda_p w_j^p.\end{aligned}$$

Similarly for $j = 1$ and $j = m+1$. ☐

Notation 14. The *Courant number* for the advection equation (12.55) is

$$\mu := \frac{ak}{h}, \tag{12.74}$$

where k and h are respectively the uniform time-step size and the uniform grid size in an MOL.

The FTCS method

Definition 12.60. The FTCS method for the advection equation (12.55) is

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right), \tag{12.75}$$

or in matrix form

$$\mathbf{U}^{n+1} = (I + kA)\mathbf{U}^n. \tag{12.76}$$

Corollary 12.61. The FTCS method for the advection equation (12.55) is unconditionally unstable in the sense of absolute stability, c.f. Definition 12.23.

Proof. The RAS of the forward Euler's method is

$$\mathcal{R} = \{z \in \mathbb{C} : |1 + z| \le 1\}.$$

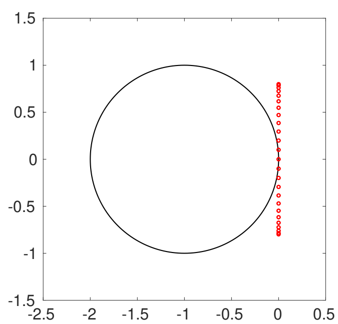
For (12.76), we have

$$z_p = k\lambda_p = -\mathbf{i}\mu \sin(2\pi ph),$$

which lies on the imaginary axis between $-\mathbf{i}\mu$ and $\mathbf{i}\mu$. Hence

$$\forall k > 0\ \forall p = 1, 2, \ldots, m+1, \quad z_p \notin \mathcal{R},$$

which implies the instability, as shown below. □



Remark 12.45. Hereafter we shall present three remedies of the FTCS method.

- Instead of using $k = O(h)$, we let $k \to 0$ faster than h (Lemma 12.62);
- Switch to an ODE solver whose RAS covers the unit interval on the imaginary axis (Definition 12.63);
- Perturb FTCS so that $\operatorname{Re}\lambda \le 0$ (Definition 12.64).

Lemma 12.62. The FTCS method for the advection equation is Lax-Richtmyer stable for $k = O(h^2)$, c.f. Definition 12.28.

Proof. Since λ_p is purely imaginary, we have

$$|1 + k\lambda_p|^2 = 1 + k\frac{k}{h^2}a^2\sin^2(2\pi ph) \le 1 + k\alpha,$$

for $\alpha = \frac{k}{h^2}a^2 = O(1)$. Since A is skew-symmetric, A is a normal operator as in Definition B.199. By the spectral theorem B.204, A has a diagonal matrix with respect to some orthonormal basis, so does $I + kA$. Thus we have

$$\|(I + kA)^n\|_2 \le \left(\|I + kA\|_2^2\right)^{\frac{n}{2}} \le (1 + k\alpha)^{\frac{n}{2}} \le e^{\frac{\alpha}{2}T},$$

where the second inequality follows from the fact that A and $I + kA$ have the same eigenvectors. This shows the uniform boundedness of the iteration matrix needed for Lax-Richtmyer stability. □

Remark 12.46. The z_p's of the FTCS method are still outside of the RAS of forward Euler, however, they all approach 0 as $h \to 0$ since now we have $k = O(h^2)$. Therefore, although the FTCS is unconditionally unstable in the sense of absolute stability, it is conditionally stable in the sense of Lax-Richtmyer stability. As the main reason, absolute stability is a sufficient but not necessary condition for the boundedness of computed results while Lax-Richtmyer stability is both sufficient and necessary.

The leapfrog method

Definition 12.63. The *leapfrog method* for the advection equation (12.55) is

$$\frac{U_j^{n+1} - U_j^{n-1}}{2k} = -\frac{a}{2h}\left(U_{j+1}^n - U_{j-1}^n\right),$$

or, equivalently

$$U_j^{n+1} = U_j^{n-1} - \mu\left(U_{j+1}^n - U_{j-1}^n\right). \tag{12.77}$$

Remark 12.47. Recall from Example 11.136 that the RAS of the midpoint method in Definition 11.57 is

$$\mathcal{R} = \{z \in \mathbb{C} : \operatorname{Re}(z) = 0,\ |z| \le 1\},$$

hence the leapfrog method (12.77) is stable in solving the advection equation so long as $|\mu| \le 1$ holds.

Lax-Friedrichs

Definition 12.64. The *Lax-Friedrichs method* for the advection equation (12.55) is

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right). \tag{12.78}$$

Remark 12.48. As the only difference of the Lax-Friedrichs method from FTCS, U_j^n is replaced with the average of its two neighbors so that the discretization corresponds more to the advection-diffusion equation than to the advection equation.

Lemma 12.65. Consider the IVP system

$$\mathbf{U}'(t) = A_\epsilon \mathbf{U}(t), \tag{12.79}$$

where

$$A_\epsilon = A + \frac{\epsilon}{h^2}\begin{bmatrix} -2 & 1 & & & & 1 \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ 1 & & & & 1 & -2 \end{bmatrix} \tag{12.80}$$

with A defined in (12.71). The eigenvalues of A_ϵ are

$$\lambda_p = -\frac{\mathbf{i}a}{h}\sin(2\pi ph) - \frac{2\epsilon}{h^2}\left[1 - \cos(2\pi ph)\right] \tag{12.81}$$

for $p = 1, 2, \ldots, m+1$. The corresponding eigenvector $\mathbf{w}^p$ has components

$$w_j^p = e^{2\pi \mathbf{i} pjh} \text{ for } j = 1, 2, \ldots, m+1. \tag{12.82}$$

Proof. This follows from Lemma 12.59 and Lemma 7.25 about eigenpairs of the discrete Laplacian. □

Remark 12.49. The IVP (12.79) can be considered as the semi-discrete system of the MOL obtained from the advection-diffusion equation

$$u_t + au_x = \epsilon u_{xx}.$$

Lemma 12.66. The Lax-Friedrichs method can be regarded as the MOL obtained by applying the forward Euler to the semi-discrete system (12.79) with $\epsilon = \frac{h^2}{2k}$.

Proof. The Lax-Friedrichs method can be rewritten as

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{1}{2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n\right),$$

which is equivalent to

$$\frac{U_j^{n+1} - U_j^n}{k} + a\left(\frac{U_{j+1}^n - U_{j-1}^n}{2h}\right) = \epsilon \frac{U_{j-1}^n - 2U_j^n + U_{j+1}^n}{h^2};$$

and this shows the standard discretization from the advection-diffusion equation. □

Theorem 12.67. The Lax-Friedrichs method (12.78) is convergent provided that $|\mu| \le 1$.

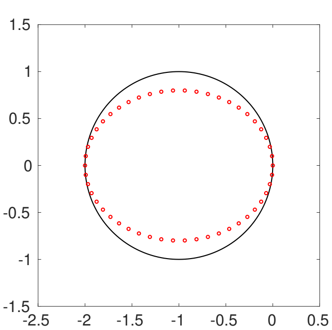
Proof. By Lemma 12.66, we have

$$z_p = k\lambda_p = -\mathbf{i}\mu \sin(2\pi ph) - \frac{2k\epsilon}{h^2}\left[1 - \cos(2\pi ph)\right],$$

thus z_p's lie on an ellipse centered at $\frac{-2k\epsilon}{h^2} = -1$ with semi-axes $\left(\frac{2k\epsilon}{h^2}, \mu\right) = (1, \mu)$. If $|\mu| \le 1$, then this ellipse lies entirely inside the absolute region of stability of the forward Euler's method. Hence the Lax-Friedrichs method is convergent provided that $|\mu| \le 1$. □

Remark 12.50. As a natural consequence to the change discussed in Remark 12.48, the z_p's in the Lax-Friedrics now have nonzero real parts.

Example 12.68. For $a = 1$, $h = \frac{1}{50}$, $k = 0.8h$, some z_p's of Lax-Friedrichs are plotted below.



Remark 12.51. The shortcoming of the Lax-Friedrichs method (12.78) is that it is only first-order accurate in time. To fix this limit, we proceed to the Lax-Wendroff method.

Lax-Wendroff

Remark 12.52. In practice, second-order accuracy both in time and in space is desirable. One way to achieve this is to use the trapezoidal method as the ODE solver, but the additional expense in solving linear systems is unnecessary since the advection equation is not stiff at all and explicit methods work perfectly fine. The Lax-Wendroff method is an explicit method free of the accuracy mismatch in the Lax-Friedrichs method.

Definition 12.69. The *Lax-Wendroff method* for the advection equation (12.55) is

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right). \end{aligned} \tag{12.83}$$

Remark 12.53. The Lax-Wendroff method (12.83) follows from the idea of Taylor series methods, i.e.,

$$\begin{aligned} u(x, t+k) &= u(x,t) + ku_t(x,t) + \frac{1}{2}k^2 u_{tt}(x,t) + O(k^3) \\ &= u - kau_x + \frac{1}{2}k^2 a^2 u_{xx} + O(k^3), \end{aligned}$$

Lemma 12.70. The Lax-Wendroff method (12.83) is second-order accurate both in space and in time.

Proof. We calculate the LTE as

$$\begin{aligned} \tau(x,t) =& \frac{u(x,t+k) - u(x,t)}{k} + a\frac{u(x+h,t) - u(x-h,t)}{2h} \\ &- \frac{ka^2}{2}\frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} \\ =& u_t(x,t) + \frac{k}{2}u_{tt}(x,t) + au_x(x,t) - \frac{ka^2}{2}u_{xx}(x,t) \\ &+ O(k^2 + h^2) \\ =& O(k^2 + h^2), \end{aligned}$$

where the first step follows from the definition of LTE, the second from Taylor expansions and the last from $u_t = -au_x$ and $u_{tt} = -au_{tx} = a^2 u_{xx}$. □

Lemma 12.71. The Lax-Wendroff method (12.83) can be considered as the MOL obtained by applying the forward Euler to the semi-discrete system (12.79) with $\epsilon = \frac{1}{2}ka^2$.

Proof. The proof is similar to that for Lemma 12.66. □

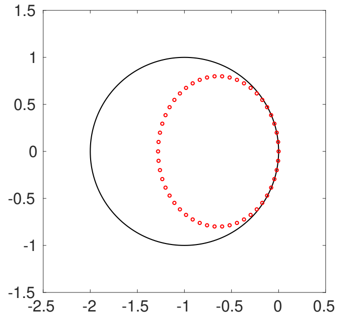
Theorem 12.72. The Lax-Wendroff method (12.83) is convergent provided $|\mu| \le 1$.

Proof. By Lemma 12.71, we have

$$z_p = k\lambda_p = -\mathbf{i}\mu \sin(2\pi ph) + \mu^2\left[\cos(2\pi ph) - 1\right].$$

These values all lie on an ellipse centered at $-\mu^2$ with semi-axes of length μ^2 and $|\mu|$. If $|\mu| \le 1$, all of these values lie inside the stability region of the forward Euler's method, thus ensuring the stability of the Lax-Wendroff method. □

Example 12.73. For $a = 1$, $h = \frac{1}{50}$, $k = 0.8h$, we have $\epsilon = 0.008$ for Lax-Wendroff; some z_p's are shown below.



The upwind method

Remark 12.54. So far we have only used centered (symmetric) spatial discretization. For one-sided finite difference formulas, the simplest is

$$u_x(x_j, t) = \frac{1}{h}(U_j - U_{j-1}) + O(h),$$

or

$$u_x(x_j, t) = \frac{1}{h}(U_{j+1} - U_j) + O(h).$$

Coupling these two equations with forward difference in time leads to the upwind method.

Definition 12.74. The *upwind method* for the advection equation (12.55) is

$$U_j^{n+1} = \begin{cases} U_j^n - \mu\left(U_j^n - U_{j-1}^n\right) & \text{if } a \geq 0; \\ U_j^n - \mu\left(U_{j+1}^n - U_j^n\right) & \text{if } a < 0. \end{cases} \tag{12.84}$$

Lemma 12.75. The upwind method (12.84) can be considered as the MOL obtained by applying the forward Euler to the semi-discrete system (12.79) with $\epsilon = \frac{h}{2}|a|$.

Proof. We only prove the case of $a > 0$. Then the upwind method can be rewritten as

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{\mu}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right),$$

which is the forward Euler's method applied to (12.79) with $\epsilon = \frac{ah}{2}$. □

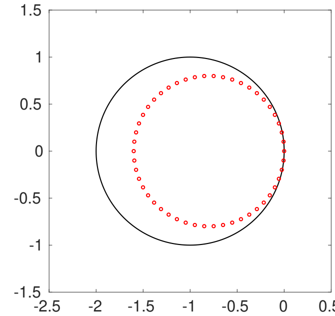
Theorem 12.76. For $a > 0$, the upwind method is convergent if and only if $\mu \leq 1$; for $a < 0$, the upwind method is convergent if and only if $\mu \geq -1$.

Proof. We only prove the case of $a > 0$. By Lemmas 12.75 and 12.65, we have

$$z_p = k\lambda_p = -\mathbf{i}\mu\sin(2\pi ph) + \mu[\cos(2\pi ph) - 1].$$

These values all lie on a circle centered at $(-\mu, 0)$ with radius μ. If $\mu \leq 1$, then all of these values lie inside the RAS of the forward Euler's method, thus ensuring the stability of the upwind method. For any choice of k, h satisfying $\mu > 1$, z_p would lie outside of the RAS and hence be unstable. □

Example 12.77. For $a = 1$, $h = \frac{1}{50}$, $k = 0.8h$, we have $\epsilon = 0.01$ for the upwind method; some z_p's are shown below.



Remark 12.55. We have emphasized the fact that all of the three methods, Lax-Wendroff, upwind, and Lax-Friedrichs, can be written in the same form (12.79) with different values of ϵ,

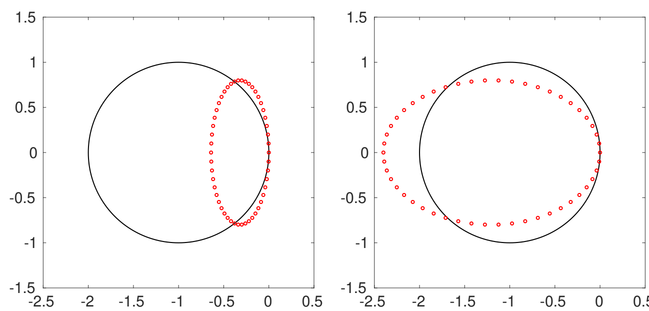
$$\epsilon_{LW} = \frac{a^2k}{2} = \frac{ah\mu}{2}, \quad \epsilon_{UP} = \frac{|a|h}{2}, \quad \epsilon_{LF} = \frac{h^2}{2k} = \frac{ah}{2\mu}.$$

Since $\epsilon_{LW} = |\mu|\epsilon_{UP}$ and $\epsilon_{UP} = |\mu|\epsilon_{LF}$, we have

$$\forall \mu \in (-1, 0) \cup (0, 1), \qquad 0 < \epsilon_{LW} < \epsilon_{UP} < \epsilon_{LF}.$$

This emphasis on Lemma 12.65 shall be clear in Section 12.2.3. Note that not all values of ϵ would work.

Example 12.78. For (12.79) with $a = 1$, the following plots show that some of the z_p's lie outside of the RAS for $\epsilon = 0.004$ and $\epsilon = 0.015$.



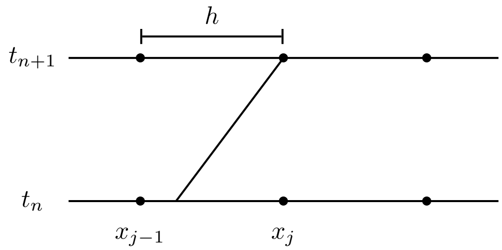
Remark 12.56. The condition $a > 0$ must be present in the proof of Theorem 12.76, otherwise the circle will be in the wrong half-plane. Physically, this also makes sense because

- $a < 0$ would mean that the first equation in the proof of Theorem 12.76 is a discretization of an advection-diffusion equation with negative diffusion coefficient. But what if a is really negative? The answer is to switch to the other one-sided finite difference formula.
- The exact solution for $[t, t + k]$ is

$$u(x_j, t + k) = u(x_j - ak, t), \tag{12.85}$$

hence the information comes from the left if $a > 0$, and from the right if $a < 0$.

Corollary 12.79. The upwind method is equivalent to characteristic tracing followed by a linear interpolation.



Proof. If we take $\mu = 1$, then the upwind method (12.84) reduces to

$$U_j^{n+1} = U_j^n - U_j^n + U_{j-1}^n = U_{j-1}^n.$$

Therefore for exact initial conditions, this method yields the exact solution by simply shifting the solution.

For $\mu < 1$, using characteristic tracing, we know

$$u(x_j, t + k) = u(x_j - ak, t).$$

Linear interpolating $u(x_j - ak, t)$ yields

$$u(x_j - ak, t) = \mu U_{j-1}^n + (1-\mu)\, U_j^n + O(h^2),$$

which leads to the upwind method

$$U_j^{n+1} = \mu U_{j-1}^n + (1-\mu)\, U_j^n = U_j^n - \mu \left(U_j^n - U_{j-1}^n\right). \qquad \square$$

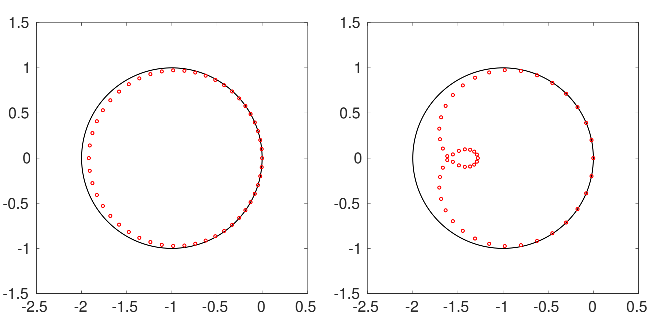
The Beam-Warming method

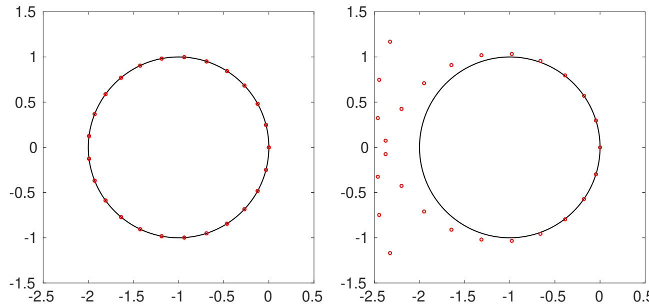
Definition 12.80. The *Beam-Warming method* solves the advection equation (12.55) by

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(3U_j^n - 4U_{j-1}^n + U_{j-2}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_j^n - 2U_{j-1}^n + U_{j-2}^n\right) \quad \text{if } a \geq 0; \qquad (12.86) \\ U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(-3U_j^n + 4U_{j+1}^n - U_{j+2}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_j^n - 2U_{j+1}^n + U_{j+2}^n\right) \quad \text{if } a < 0. \qquad (12.87) \end{aligned}$$

Exercise 12.81. Show that the Beam-Warming method is second-order accurate both in time and in space.

Exercise 12.82. Show that the Beam-Warming methods (12.86) and (12.87) are stable for $\mu \in [0, 2]$ and $\mu \in [-2, 0]$, respectively. Reproduce the following plots for $\mu = 0.8$, 1.6, 2, and 2.4.





12.2.2 The CFL condition

Definition 12.83. For the advection equation (12.55), the *domain of dependence* of a point $(X, T) \in \Omega$ is

$$\mathcal{D}_{\text{ADV}}(X, T) = \{X - aT\}. \qquad (12.88)$$

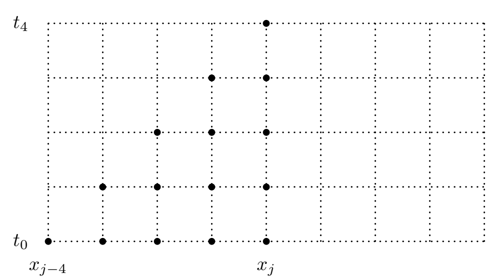
Remark 12.57. The solution $u(X, T)$ at some fixed point (X, T) depends on the initial data η at the single point $X - aT$ since $u(X, T) = \eta(X - aT)$.

Definition 12.84. The *numerical domain of dependence* of a grid point (x_j, t_n) is the set of all grid points x_i such that U_i^0 at (x_i, t_0) affects U_j^n.

$$\mathcal{D}_N(x_j, t_n) = \{x_i : U_i^0 \text{ affects } U_j^n\}. \qquad (12.89)$$

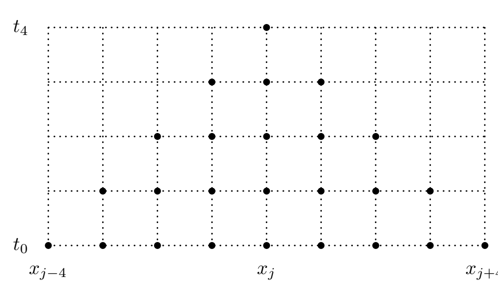
Remark 12.58. The word "affects" in Definition 12.84 represents a partial order on the grid points that is specified by the formula of an MOL.

Example 12.85. The numerical domain of dependence of a grid point (x_j, t_4) for the upwind method with $a > 0$ and $\mu \notin \{0, 1\}$ is shown as the solid dots below.



Exercise 12.86. Plot the numerical domains of dependence of the grid point (x_j, t_3) for the upwind method with $a < 0$ and $\mu = 0, -1, -2$.

Example 12.87. The numerical domain of dependence of a grid point (x_j, t_4) for the Lax-Wendroff method with $\mu \notin \{0, 1, -1\}$ is shown as the solid dots below.



Exercise 12.88. Plot the numerical domains of dependence of the grid point (x_j, t_3) for the Lax-Wendroff method with $\mu = +1, -1$.

Theorem 12.89 (Courant-Friedrichs-Lewy)**.** A numerical method can be convergent only if its numerical domain of dependence contains the domain of dependence of the PDE, at least in the limit of $k, h \to 0$.

Proof. It suffices to say that if some point p in the domain of dependence is not contained in the numerical domain of dependence, then we have no control over the value of p in the numerical method. Consequently, the numerical method cannot converge. □

Example 12.90. In solving the advection equation with $a = 1$, any choice of $k > h$ will result in instability. For $k = 3h$, the numerical domain of dependence for U_j^3, as shown in Example 12.85, is $\{x_j, x_{j-1}, x_{j-2}, x_{j-3}\}$ while the domain of dependence is the singleton set $\{x_j - 9h\}$. The former does not contain the latter and thus this choice of k will result in divergence.

Example 12.91. The heat equation

$$\begin{cases} u_t = \nu u_{xx} \\ u(x,0) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(x-\bar{x})^2}, \end{cases} \tag{12.90}$$

has its exact solution as

$$u(x,t) = \frac{1}{\sqrt{4\pi\nu t + \frac{\pi}{\beta}}} e^{-(x-\bar{x})^2/(4\nu t + \frac{1}{\beta})}. \tag{12.91}$$

The domain of dependence is the whole line, i.e.,

$$\mathcal{D}_{\mathrm{DIFF}}(X,T) = (-\infty, +\infty), \tag{12.92}$$

because an initial point source

$$\lim_{\beta\to\infty} u(x,0) = \delta(x - \bar{x})$$

instantaneously affects each point on the real line:

$$\lim_{\beta\to\infty} u(x,t) = \frac{1}{\sqrt{4\pi\nu t}} e^{-\frac{(x-\bar{x})^2}{4\nu t}}.$$

This is very much an artifact of the mathematical model rather than the true physics.

Remark 12.59. In solving the heat equation, $k = O(h^2)$ is needed to maintain stability for explicit one-step methods. As then when we make the grid finer by a factor of 2 in space it will become finer by a factor of 4 in time, and hence the numerical domain of dependence will cover a wider interval at time $t = 0$. As $k \to 0$ the numerical domain of dependence will spread to cover the entire real line, and hence the CFL condition is satisfied in this case.

An implicit method such as the Crank-Nicolson method satisfies the CFL condition for any time step k. In this case the numerical domain of dependence is the entire real line because the tridiagonal linear system couples together all points in such a manner that the solution at each point depends on the data at all points (i.e., the inverse of a tridiagonal matrix is dense).

12.2.3 Modified equations

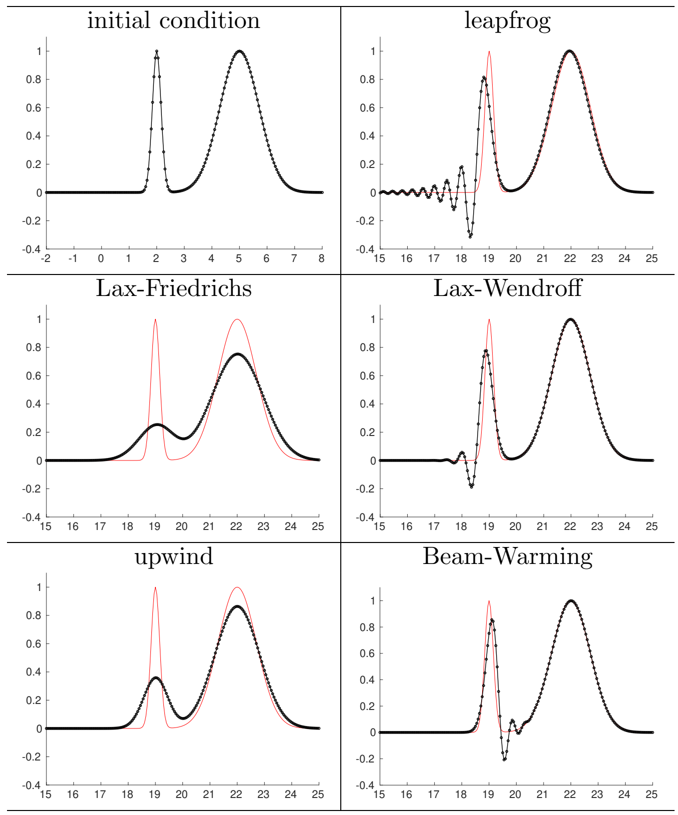
Example 12.92. For the advection equation

$$u_t + u_x = 0$$

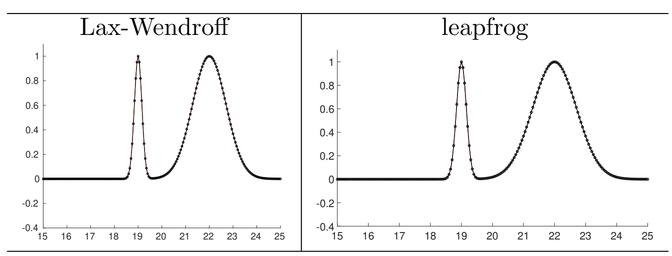
with initial condition

$$u(x,0) = \eta(x) = \exp(-20(x-2)^2) + \exp(-(x-5)^2), \tag{12.93}$$

the exact solution at $t = T$ is simply the initial data shifted by T. We solve this problem with $h = 0.05$ to $T = 17$ using the leapfrog method, the Lax-Friedrichs method, the Lax-Wendroff method, the upwind method, and the Beam-Warming method. The final results with $k = 0.8h$ are shown below.



If we keep all parameters the same except the change $k = h$, we have the following results.



These results invite a number of questions as follows.

(a) Why are the solutions of Lax-Friedrichs and upwind much smoother than those of the other three methods?

(b) What caused the ripples in the solutions of the three methods in the right column?

(c) Why does the numerical solution of the leapfrog method contain more oscillations than that of the Lax-Wendroff method?

(d) For the Lax-Wendroff method, why do the ripples of numerical solutions lag behind the true crest?

(e) For the Beam-Warming method, why do the ripples of numerical solutions move ahead of the true crest?

(f) Why are numerical results with $k = h$ much better than those with $k = 0.8h$?

These questions concern the physics behind the different features of the results of different methods; they can be answered by the modified equations.

Definition 12.93. The *modified equation of an MOL for solving a PDE* (the original equation) is another PDE obtained from the formula of the MOL by

(1) replacing U_j^n in the MOL formula with a smooth function $v(x_j, t_n)$,

(2) expanding all terms in Taylor series at (x_j, t_n),

(3) neglecting potentially high-order terms.

Remark 12.60. By Definition 12.93, the grid function $v(x_j, t_n)$ satisfies the modified equation better than the original equation. Thus the modified equation helps us to understand the behavior of the numerical solution.

Example 12.94. Consider the upwind method for solving the advection equation

$$U_j^{n+1} = U_j^n - \mu\left(U_j^n - U_{j-1}^n\right).$$

The modified equation can be derived as follows.

(1) Replace U_j^n with $v(x_j, t_n)$ and we have

$$v(x, t+k) = v(x,t) - \mu\left(v(x,t) - v(x-h,t)\right).$$

(2) Expand all terms in Taylor series at (x, t) in a way similar to the calculation of the LTE.

$$\begin{aligned} 0 =& \frac{v(x,t+k) - v(x,t)}{k} + \frac{a}{h}\left(v(x,t) - v(x-h,t)\right) \\ =& \left(v_t + \frac{1}{2}kv_{tt} + \frac{1}{6}k^2 v_{ttt} + \cdots\right) \\ &+ a\left(v_x - \frac{1}{2}hv_{xx} + \frac{1}{6}h^2 v_{xxx} + \cdots\right), \end{aligned}$$

and thus

$$v_t + av_x = \frac{1}{2}\left(ahv_{xx} - kv_{tt}\right) - \frac{1}{6}\left(ah^2 v_{xxx} + k^2 v_{ttt}\right) + \cdots,$$

differentiating with respect to t and x gives

$$\begin{aligned} v_{tt} &= -av_{xt} + \frac{1}{2}\left(ahv_{xxt} - kv_{ttt}\right) + \cdots, \\ v_{tx} &= -av_{xx} + \frac{1}{2}\left(ahv_{xxx} - kv_{xtt}\right) + \cdots. \end{aligned}$$

Combining these gives

$$v_{tt} = a^2 v_{xx} + O(h+k).$$

Therefore we have

$$v_t + av_x = \frac{1}{2}ah\left(1-\mu\right)v_{xx} + O(h^2 + k^2),$$

(3) Neglect the high-order terms and we have the modified equation as

$$v_t + av_x = \frac{1}{2}ah\left(1-\mu\right)v_{xx} := \beta v_{xx}, \tag{12.94}$$

which is satisfied better by the grid function than the advection equation $v_t + av_x = 0$.

Remark 12.61. We emphasize that $v_t + av_x \neq 0$ in step (2) of Example 12.94; otherwise the numerical solution would be the exact solution. However, in the special case of $a = 1$ and $k = h$, (12.94) reduces to the advection equation, and in this case we already know that the exact solution to the advection equation is recovered by the upwind method.

Remark 12.62. According to Lemma F.31, the solution to the modified equation (12.94) is

$$v(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{-\beta\xi^2 t} e^{i\xi(x-at)} \mathrm{d}\xi,$$

here the term "$e^{-\beta\xi^2 t}$" indicates dissipation while "$e^{i\xi(x-at)}$" represents a translation with exactly the same speed as that of the original equation. The dissipative nature of the term "$e^{-\beta\xi^2 t}$" answers Question (a) in Example 12.92. Notice that the position of the crests in the results of the upwind and Lax-Friedrichs methods are the same as that of the exact solution.

Example 12.95. The modified equation of the Lax-Wendroff method for the advection equation is

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = 0, \tag{12.95}$$

which can be derived as follows.

First, replace U_j^n with $v(x,t)$ and we have

$$\begin{aligned} &\frac{v(x,t+k) - v(x,t)}{k} + \frac{a}{2h}\left(v(x+h,t) - v(x-h,t)\right) \\ =&\frac{a^2 k}{2h^2}\left(v(x+h,t) - 2v(x,t) + v(x-h,t)\right). \end{aligned}$$

Second, we expand all terms in Taylor series about (x,t) to derive the modified equation as (for brevity, omit the dependence on (x,t))

$$v_t + \frac{1}{2}kv_{tt} + \frac{1}{6}k^2 v_{ttt} + a\left(v_x + \frac{h^2}{6}v_{xxx}\right) = \frac{k}{2}a^2 v_{xx} + O(k^3),$$

where we have assumed $h = O(k)$. It follows that

$$v_t + av_x = -\frac{1}{2}k\left(v_{tt} - a^2 v_{xx}\right) - \frac{1}{6}k^2 v_{ttt} - \frac{ah^2}{6}v_{xxx} + O(k^3).$$

Differentiate the above equation and we have

$$\begin{aligned}
v_{tt} &= -av_{xt} - \frac{1}{2}kv_{ttt} + \frac{k}{2}a^2 v_{xxt} + O(k^2),\\
v_{xt} &= -av_{xx} - \frac{k}{2}v_{xtt} + \frac{k}{2}a^2 v_{xxx} + O(k^2),\\
v_{xxt} &= -av_{xxx} + O(k),\\
v_{xtt} &= -av_{xxt} + O(k) = a^2 v_{xxx} + O(k),\\
v_{ttt} &= -av_{xtt} + O(k) = -a^3 v_{xxx} + O(k).
\end{aligned}$$

Combining the above equations gives

$$\begin{aligned}
v_{tt} =& a^2 v_{xx} + \frac{ka}{2}(v_{xtt} - a^2 v_{xxx}) + \frac{ka^3}{2}v_{xxx} - \frac{ka^3}{2}v_{xxx} + O(k^2)\\
=& a^2 v_{xx} + O(k^2).
\end{aligned}$$

Hence we have

$$v_t + av_x = \frac{k^2}{6}a^3 v_{xxx} - \frac{ah^2}{6}v_{xxx} + O(k^3),$$

which implies (12.95).

Example 12.96. By Lemma F.31, the solution to the modified equation (12.95) is

$$v(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i\xi(x - C_p t)} \mathrm{d}\xi.$$

For Lax-Wendroff, (12.95) and Example F.46 yield

$$\begin{aligned}
C_p(\xi) &= a - \frac{ah^2}{6}\left(1-\mu^2\right)\xi^2,\\
C_g(\xi) &= a - \frac{ah^2}{2}\left(1-\mu^2\right)\xi^2.
\end{aligned}$$

First, the phase velocity $C_p \neq a$ for $\mu \neq 1$, and its value depends on ξ; this answers Question (b) of Example 12.92. For $\mu \neq 1$, both C_p and C_g have a magnitude smaller than $|a|$, hence numerical oscillations lag behind the true wave crest; this answers Question (d) of Example 12.92.

Exercise 12.97. Show that the modified equation of the leapfrog method is also (12.95). However, if one more term of higher-order derivative had been retained, the modified equation of the leapfrog method would have been

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = \epsilon_f v_{xxxxx} \tag{12.96}$$

while that of the Lax-Wendroff method would have been

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = \epsilon_w v_{xxxx}. \tag{12.97}$$

Remark 12.63. Due to the symmetry of the modified equation of the leapfrog method,

$$\frac{v(x,t+k) - v(x,t-k)}{2k} + a\frac{v(x+h,t) - v(x-h,t)}{2h} = 0,$$

all even derivatives drop out. Because of the absence of even derivatives, there is no dissipation in the numerical solution of the leapfrog method. In contrast, in the modified equation of the Lax-Wendroff method, the fourth-order derivative term dissipates high-frequency modes and is responsible for a more smooth solution than that of the leapfrog method. This answers Question (c) of Example 12.92.

Exercise 12.98. Show that the modified equation of the Beam-Warming method applied to the advection equation (12.55) with $a \geq 0$ is

$$v_t + av_x + \frac{ah^2}{6}\left(-2 + 3\mu - \mu^2\right)v_{xxx} = 0. \tag{12.98}$$

Thus we have

$$\begin{aligned}
C_p(\xi) &= a + \frac{ah^2}{6}(\mu-1)(\mu-2)\xi^2,\\
C_g(\xi) &= a + \frac{ah^2}{2}(\mu-1)(\mu-2)\xi^2.
\end{aligned}$$

How do these facts answer Question (e) of Example 12.92?

Exercise 12.99. What if $\mu = 1$? Discuss this case for both Lax-Wendroff and leapfrog methods to answer Question (f) of Example 12.92.

12.2.4 Von Neumann analysis

Remark 12.64. The computation of eigenvalues of the iteration matrix and the absolute stability analysis of time-stepping method can be combined into von Neumann analysis when the PDE solution is of the form

$$u(x,t) = e^{\mathbf{i}\xi x_j - \omega t_n} = [g(\xi)]^n e^{\mathbf{i}\xi j h}.$$

The ultimate reason that von Neumann analysis works is the fact that linear PDEs can be analyzed effectively by Fourier transforms; see Appendix F.

Example 12.100. For the advection equation (12.55) with $a \geq 0$, the von Neumann analysis of the upwind method yields its amplification factor as

$$g(\xi) = (1-\mu) + \mu e^{-\mathbf{i}\xi h}. \tag{12.99}$$

Indeed, substituting $U_j^n = [g(\xi)]^n e^{\mathbf{i}\xi j h}$ into the upwind method

$$U_j^{n+1} = U_j^n - \mu\left(U_j^n - U_{j-1}^n\right)$$

gives

$$g(\xi) = 1 - \mu\left(1 - e^{-\mathbf{i}\xi h}\right) = (1-\mu) + \mu e^{-\mathbf{i}\xi h}.$$

Therefore

$$\begin{aligned}
|g(\xi)|^2 &= (1 - \mu + \mu\cos(\xi h))^2 + \mu^2 \sin^2(\xi h)\\
&= 1 + 2\mu(\mu-1)(1-\cos(\xi h)),
\end{aligned}$$

and hence the method is stable ($|g(\xi)| \leq 1$) provided $\mu \leq 1$.

Exercise 12.101. Apply the von Neumann analysis to the Lax-Friedrichs method to derive its amplification factor as

$$g(\xi) = \cos(\xi h) - \mu\mathbf{i}\sin(\xi h). \tag{12.100}$$

For which values of μ would the method be stable?

Exercise 12.102. Apply the von Neumann analysis to the Lax-Wendroff method to derive its amplification factor as

$$g(\xi) = 1 - 2\mu^2 \sin^2 \frac{\xi h}{2} - \mathbf{i}\mu \sin(\xi h). \tag{12.101}$$

For which values of μ would the method be stable?

Example 12.103. When performing the analysis of modified equations, we typically neglect some higher-order terms of ξh in deriving the group velocity and the phase velocity. For ξh sufficiently small, the modified equation would be a reasonable model. However, for large ξh the terms we have neglected may play an equally important role. In this case it might be better to use an approach similar to von Neumann analysis by setting

$$v(x_j, t_n) := e^{\mathbf{i}(\xi x_j - \omega t_n)}. \tag{12.102}$$

For the leapfrog method, this form yields

$$\sin(\omega k) = \mu \sin(\xi h), \tag{12.103}$$

which yields the group velocity as

$$\frac{\mathrm{d}\omega}{\mathrm{d}\xi} = \pm \frac{a\cos(\xi h)}{\sqrt{1-\mu^2\sin^2(\xi h)}}, \tag{12.104}$$

where the $\pm$ follows from the multivalued dispersion relation (12.103). For high-frequency modes satisfying $\xi h \approx \pi$, the group velocity may have a sign different from that of a.

12.3 Programming assignments

12.3.1 MOL for the heat equation

Write subroutines to solve the test problem in Example 12.38. More specifically, you need to

(a) reproduce all plots in Example 12.38,

(b) rerun your subroutine for BTCS and the collocation method in Example 11.258 (coupled with centered difference in space) with $r = \frac{1}{2h}$, show their results after 1, 2, and 10 steps,

(c) write a subroutine for FTCS and show its results after 1, 2, and 10 steps with $r = \frac{1}{2}, 1$.

(d) write a subroutine for the 1-stage Gauss-Legendre RK method in Example 11.227 (coupled with centered difference in space) and show its results after 1, 2, and 10 steps with $r = 1, \frac{1}{2h}$.

(e) collect the results, compare them, and explain your understanding in your own words in a report.

12.3.2 MOL for the advection equation

Reproduce all results in Example 12.92.