Chapter 8

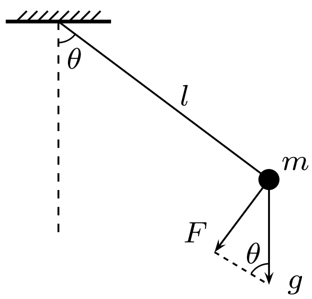
Initial Value Problems

Definition 8.1. A *system of ordinary differential equations* (ODEs) of dimension N is a set of differential equations of the form

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t), t), \tag{8.1}$$

where t is time, $\mathbf{u} \in \mathbb{R}^N$ is the evolutionary variable, and the RHS function has the signature $\mathbf{f} : \mathbb{R}^N \times (0, +\infty) \to \mathbb{R}^N$. In particular, (8.1) is an ODE for $N = 1$.

Definition 8.2. A system of ODEs is *linear* if its RHS function can be expressed as $\mathbf{f}(\mathbf{u}, t) = \alpha(t)\mathbf{u} + \boldsymbol{\beta}(t)$, and *nonlinear* otherwise; it is *homogeneous* if it is linear and $\boldsymbol{\beta}(t) = \mathbf{0}$; it is *autonomous* if $\mathbf{f}$ does not depend on t explicitly; and *nonautonomous* otherwise.



Example 8.3. For the simple pendulum shown above, the moment of inertial and the torque are

$$I = m\ell^2, \ \tau = -mg\ell \sin\theta,$$

and the equation of motion can be derived from Newton's second law $\tau = I\theta''(t)$ as

$$\theta''(t) = -\frac{g}{\ell}\sin\theta, \tag{8.2}$$

which admits a unique solution if we impose two initial conditions

$$\theta(0) = \theta_0, \ \theta'(0) = \omega_0.$$

Alternatively, (8.2) can be derived by the consideration that the total energy remains a constant with respect to time.

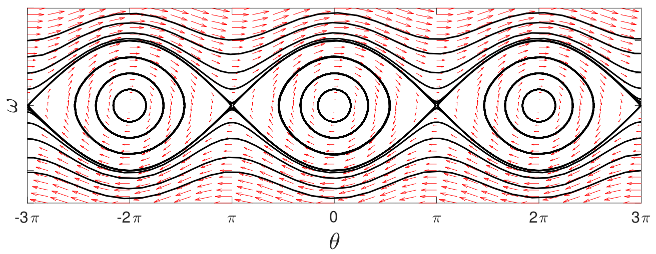
$$L = \frac{1}{2}m(\ell\theta')^2 + mg\ell(1 - \cos\theta);$$

$$\frac{\mathrm{d}L}{\mathrm{d}t} = 0 \Rightarrow \ m\ell^2\theta'\theta'' + mg\ell\theta'\sin\theta = 0.$$

The ODE (8.2) is second-order, nonlinear, and autonomous; it can be reduced to a first-order system as follows,

$$\omega = \theta', \ \mathbf{u} = \begin{pmatrix} \theta \\ \omega \end{pmatrix} \ \Rightarrow \ \mathbf{u}'(t) = \mathbf{f}(u) := \begin{pmatrix} \omega \\ -\frac{g}{\ell}\sin\theta \end{pmatrix}.$$

See the following plot for some solutions.



Definition 8.4. Given $T > 0$, $\mathbf{f} : \mathbb{R}^N \times [0, T] \to \mathbb{R}^N$, and $\mathbf{u}_0 \in \mathbb{R}^N$, the *initial value problem* (IVP) is to find $\mathbf{u}(t) \in \mathcal{C}^1$ such that

$$\begin{cases} \mathbf{u}(0) & = \mathbf{u}_0, \\ \mathbf{u}'(t) & = \mathbf{f}(\mathbf{u}(t), t), \ \forall t \in [0, T]. \end{cases} \tag{8.3}$$

Definition 8.5. The IVP in Definition 8.4 is *well-posed* if

(i) it admits a unique solution for any fixed $t > 0$,

(ii) $\exists c > 0$, $\hat{\epsilon} > 0$ s.t. $\forall \epsilon < \hat{\epsilon}$, the perturbed IVP

$$\mathbf{v}' = \mathbf{f}(\mathbf{v}, t) + \boldsymbol{\delta}(t), \qquad \mathbf{v}(0) = \mathbf{u}_0 + \boldsymbol{\epsilon}_0 \tag{8.4}$$

satisfies

$$\forall t \in [0, T], \begin{cases} \|\boldsymbol{\epsilon}_0\| < \epsilon \\ \|\boldsymbol{\delta}(t)\| < \epsilon \end{cases} \ \Rightarrow \ \|\mathbf{u}(t) - \mathbf{v}(t)\| \le c\epsilon. \tag{8.5}$$

8.1 Lipschitz continuity

Definition 8.6. A function $\mathbf{f} : \mathbb{R}^N \times [0, +\infty) \to \mathbb{R}^N$ is *Lipschitz continuous* in its first variable over some domain

$$\mathcal{D} = \{(\mathbf{u}, t) : \|\mathbf{u} - \mathbf{u}_0\| \le a, t \in [0, T]\} \tag{8.6}$$

if

$$\exists L \ge 0 \text{ s.t. } \forall (\mathbf{u}, t), (\mathbf{v}, t) \in \mathcal{D}, \ \|\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t)\| \le L\|\mathbf{u} - \mathbf{v}\|. \tag{8.7}$$

Example 8.7. If $\mathbf{f}(\mathbf{u},t)=\mathbf{f}(t)$, then $L=0$.

Example 8.8. If $\mathbf{f}\notin\mathcal{C}^0$, then $\mathbf{f}$ is not Lipschitz.

Definition 8.9. A subset $S\subset\mathbb{R}^n$ is *star-shaped* with respect to a point $p\in S$ if for each $x\in S$ the line segment from p to x lies in S.

Theorem 8.10. Let $S\subset\mathbb{R}^n$ be star-shaped with respect to $p=(p_1,p_2,\ldots,p_n)\in S$. For a continuously differentiable function $f:S\to\mathbb{R}^m$, there exist continuously differentiable functions $g_1(\mathbf{x}),g_2(\mathbf{x}),\ldots,g_n(\mathbf{x})$ such that

$$f(\mathbf{x})=f(p)+\sum_{i=1}^{n}(x_i-p_i)g_i(\mathbf{x}),\quad g_i(p)=\frac{\partial f}{\partial x_i}(p). \tag{8.8}$$

Proof. Since S is star-shaped, for any given $\mathbf{y}\in S$ and $t\in[0,1]$, $f(\mathbf{x})$ is defined for $\mathbf{x}=p+t(\mathbf{y}-p)$. Then the chain rule yields

$$\frac{\mathrm{d}}{\mathrm{d}t}f(\mathbf{x})=\sum_i\frac{\partial f}{\partial x_i}\frac{\mathrm{d}x_i}{\mathrm{d}t}=\sum_i(y_i-p_i)\frac{\partial f}{\partial x_i}(\mathbf{x}).$$

An integration with respect to t from 0 to 1 leads to

$$\begin{aligned}f(\mathbf{y})-f(p)&=\sum_i(y_i-p_i)g_i(\mathbf{y}),\\ g_i(\mathbf{y})&=\int_0^1\frac{\partial f}{\partial x_i}(p+t(\mathbf{y}-p))\mathrm{d}t,\end{aligned}$$

where the function $g_i(p)=\frac{\partial f}{\partial x_i}(p)$. □

Proposition 8.11. If $\mathbf{f}(\mathbf{u},t)$ is continuously differentiable on some compact convex set $\mathcal{D}\subseteq\mathbb{R}^{N+1}$, then $\mathbf{f}$ is Lipschitz on $\mathcal{D}$ with

$$L=\max_{i,j}\left|\frac{\partial f_i}{\partial u_j}\right|.$$

Proof. Indeed, for $N=1$, the mean value theorem states that if $f:[a,b]\to\mathbb{R}$ is continuous on $[a,b]$ and differentiable on (a,b), then $\exists c\in(a,b)$ s.t. $f'(c)=\frac{f(a)-f(b)}{a-b}$. The compactness of $\mathcal{D}$ and $f'(x)$ being continuous imply that $f'(x)$ is bounded.

If $N>1$, the convexity of $\mathcal{D}$ and the differentiability of $\mathbf{f}$ imply that the directional derivative of $\mathbf{f}$ exists along the line determined by any $(\mathbf{u},t)$ and $(\mathbf{v},t)$. The rest of the proof is similar to the 1D case. □

Lemma 8.12. Let (M,ρ) denote a complete metric space and $\phi:M\to M$ a contractive mapping in the sense that

$$\exists c\in[0,1)\text{ s.t. }\forall\eta,\zeta\in M,\ \rho(\phi(\eta),\phi(\zeta))\le c\rho(\eta,\zeta). \tag{8.9}$$

Then there exists a unique $\xi\in M$ such that $\phi(\xi)=\xi$.

Theorem 8.13 (Fundamental theorem of ODEs)**.** If $\mathbf{f}(\mathbf{u}(t),t)$ is Lipschitz continuous in $\mathbf{u}$ and continuous in t over some region $\mathcal{D}=\{(\mathbf{u},t):\|\mathbf{u}-\mathbf{u}_0\|\le a,t\in[0,T]\}$, then there is a unique solution to the IVP problem as in Definition 8.4 at least up to time $T^*=\min(T,\frac{a}{S})$ where $S=\max_{(\mathbf{u},t)\in\mathcal{D}}\|\mathbf{f}(\mathbf{u},t)\|$.

Proof. It suffices to prove the case of $a=+\infty$ since the minimum ensures that the solution $\mathbf{u}(t)$ remains in the domain $\mathcal{D}$ where the Lipschitz continuity holds.

Let (M,ρ) denote the complete metric space of continuous functions $\mathbf{u}:[0,T]\to\mathbb{R}^N$ such that $\mathbf{u}(0)=\mathbf{u}_0$. The metric is defined by

$$\rho(\mathbf{u},\mathbf{v})=\sup_{t\in[0,T]}\exp(-Kt)\|\mathbf{u}(t)-\mathbf{v}(t)\|,$$

where $K>L$.

For a given $\mathbf{u}\in M$, define $\boldsymbol{\phi}(\mathbf{u})$ as the solution $\mathbf{U}$ on $[0,T]$ to the IVP in Definition 8.4, which is solvable by integration as

$$\boldsymbol{\phi}(\mathbf{u})(t)=\mathbf{u}_0+\int_0^t\mathbf{f}(\mathbf{u}(s),s)\mathrm{d}s.$$

$\boldsymbol{\phi}$ is a contractive mapping because $\forall\mathbf{u},\mathbf{v}\in M$,

$$\begin{aligned}&\rho(\boldsymbol{\phi}(\mathbf{u}),\boldsymbol{\phi}(\mathbf{v}))\\
=&\sup_{t\in[0,T]}\exp(-Kt)\left\|\int_0^t\big(\mathbf{f}(\mathbf{u}(s),s)-\mathbf{f}(\mathbf{v}(s),s)\big)\mathrm{d}s\right\|\\
\le&\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\|\mathbf{f}(\mathbf{u}(s),s)-\mathbf{f}(\mathbf{v}(s),s)\|\,\mathrm{d}s\\
\le&L\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\|\mathbf{u}(s)-\mathbf{v}(s)\|\,\mathrm{d}s\\
\le&L\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\exp(-Ks)\,\|\mathbf{u}(s)-\mathbf{v}(s)\|\exp(Ks)\mathrm{d}s\\
\le&L\rho(\mathbf{u},\mathbf{v})\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\exp(Ks)\mathrm{d}s\\
\le&\frac{L}{K}\rho(\mathbf{u},\mathbf{v}).\end{aligned}$$

The rest follows from Lemma 8.12. □

Theorem 8.14. If $\mathbf{f}(\mathbf{u},t)$ is Lipschitz in $\mathbf{u}$ and continuous in t on $\mathcal{D}=\{(\mathbf{u},t):\mathbf{u}\in\mathbb{R}^N,t\in[0,T]\}$, then the IVP in Definition 8.4 is well-posed for all initial data.

Proof. Exercise. □

Example 8.15. Consider $N=1$, $u'(t)=\sqrt{u(t)}$, $u(0)=0$.

$$\lim_{u\to0}f'(u)=\lim_{u\to0}\frac{1}{2\sqrt{u}}=+\infty.$$

Hence $f(u)$ is not Lipschitz near $u=0$. However, $u(t)\equiv0$ and $u(t)=\frac14t^2$ are both solutions. Hence the Lipschitz condition is not necessary for existence.

Example 8.16. Consider the IVP $u'(t)=u^2$, $u_0=\eta>0$. The slope $f'(u)=2u\to+\infty$ as $u\to\infty$. So there is no unique solution on $[0,+\infty)$, but there might exist T^* such that unique solutions are guaranteed on $[0,T^*]$.

In fact, $u(t)=\frac{1}{\eta^{-1}-t}$ is a solution, but blows up at $t=1/\eta$. According to Theorem 8.13, $f(u)=u^2$ and we would like to maximize a/S. Since $S=\max_{\mathcal{D}}|f(u)|=(\eta+a)^2$, it is equivalent to find $\min_{a>0}(\eta+a)^2/a$:

$$(\eta+a)^2/a=2\eta+a+\eta^2\frac1a\ge2\eta+2\sqrt{\eta^2}=4\eta.$$

Hence $T^*=\frac{1}{4\eta}$. The estimation of T^* in Theorem 8.13 is thus quite conservative for this case.

Example 8.17. For the simple pendulum in Example 8.3, we have

$$|\sin\theta - \sin\theta^*| \le |\theta - \theta^*| \le \|\mathbf{u} - \mathbf{u}^*\|_\infty$$

since $\cos\theta^* \le 1$. In addition, we have $|\omega - \omega^*| \le \|\mathbf{u} - \mathbf{u}^*\|_\infty$.

$$\begin{aligned}\|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{u}^*)\|_\infty &= \max\left(|\omega - \omega^*|, \frac{g}{\ell}|\sin\theta - \sin\theta^*|\right)\\ &\le \max(\frac{g}{\ell}, 1)\|\mathbf{u} - \mathbf{u}^*\|_\infty.\end{aligned}$$

Therefore, $\mathbf{f}$ is Lipschitz continuous with $L = \max(g/l, 1)$.

8.2 Duhamel's principle

Definition 8.18. Two matrices A and B are *similar* if there exists a nonsingular matrix S such that

$$B = S^{-1}AS, \tag{8.10}$$

and $S^{-1}AS$ is called a *similarity transformation* of A.

Theorem 8.19. Two similar matrices A and B have the same set of eigenvalues.

Proof. Let $(\lambda, \mathbf{u})$ be an eigen-pair of B, i.e.,

$$B\mathbf{u} = \lambda\mathbf{u}.$$

Combine with (8.10), and we have

$$AS\mathbf{u} = SB\mathbf{u} = \lambda S\mathbf{u},$$

and thus λ is an eigenvalue of A with corresponding eigenvector $S\mathbf{u}$. □

Definition 8.20. $A \in \mathbb{C}^{m\times m}$ is *diagonalizable* if there exists a similarity transformation that maps A to a diagonal matrix Λ, i.e.,

$$\exists \text{ invertible } R \text{ s.t. } R^{-1}AR = \Lambda. \tag{8.11}$$

Definition 8.21. Let $A \in \mathbb{C}^{m\times m}$, then the *matrix exponential* e^{At} is defined by

$$e^{At} := I + At + \frac{1}{2!}A^2t^2 + \cdots = \sum_{j=0}^{\infty}\frac{1}{j!}A^jt^j. \tag{8.12}$$

Proposition 8.22. If A is diagonalizable, i.e., (8.11) holds, then

$$\begin{aligned}e^{At} &= RR^{-1} + R\Lambda R^{-1}t + \frac{1}{2!}R\Lambda R^{-1}R\Lambda R^{-1}t^2 + \cdots\\ &= R\sum_{j=0}^{\infty}\frac{t^j}{j!}\Lambda^jR^{-1} = Re^{\Lambda t}R^{-1}.\end{aligned}$$

Theorem 8.23. For a linear IVP $\mathbf{f}(\mathbf{u}, t) = A(t)\mathbf{u} + \mathbf{g}(t)$ with a constant matrix $A(t) = A$, the solution is

$$\mathbf{u}(t) = e^{At}\mathbf{u}_0 + \int_0^t e^{A(t-\tau)}\mathbf{g}(\tau)\mathrm{d}\tau. \tag{8.13}$$

Proof. For $N = 1$, (8.13) follows from Leibniz's formula

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\int_{a(x)}^{b(x)} f(x,y)\mathrm{d}y = \int_a^b \frac{\partial}{\partial x}f(x,y)\mathrm{d}y - f(x,a)\frac{\mathrm{d}a}{\mathrm{d}x}\\ + f(x,b)\frac{\mathrm{d}b}{\mathrm{d}x}.\end{aligned}$$

□

Example 8.24. Many linear problems are naturally formulated in the form of a single high-order ODE

$$v^{(m)}(t) - \sum_{j=1}^{m} c_j(t)v^{(m-j)} = \phi(t). \tag{8.14}$$

By setting $u_j(t) = v^{(j-1)}$ and $\mathbf{u} = [u_1, u_2, \ldots, u_m]^T$, we can rewrite (8.14) as

$$\mathbf{u}'(t) = A(t)\mathbf{u} + \mathbf{g}(t), \tag{8.15}$$

where $\mathbf{g}(t) = [0, \ldots, 0, \phi(t)]^T$ and

$$a_{ij}(t) = \begin{cases} 1 & \text{if } i = j-1,\\ c_{m+1-j}(t) & \text{if } i = m,\\ 0 & \text{otherwise.}\end{cases}$$

Theorem 8.25 (Superposition principle)**.** If $\hat{\mathbf{u}}$ is a solution to the IVP

$$\mathbf{u}'(t) = A(t)\mathbf{u} + \mathbf{g}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0 \tag{8.16}$$

and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are solutions to the homogeneous IVP $\mathbf{u}'(t) = A(t)\mathbf{u}$, $\mathbf{u}(0) = \mathbf{0}$, then for any constants $\alpha_1, \alpha_2, \ldots, \alpha_k$, the function

$$\mathbf{U}(t) = \hat{\mathbf{u}}(t) + \sum_{i=1}^{k}\alpha_i\mathbf{v}_i(t) \tag{8.17}$$

is a solution to (8.16).

Proof. It is trivial to verify the conclusion by differentiating (8.17). Due to the homogeneous initial conditions of $\mathbf{v}_i$'s, $\mathbf{U}(t)$ also satisfies the initial condition. □

8.3 Some basic numerical methods

Notation 8. In the following, we shall use k to denote the time step, and thus $t_n = nk$.

To numerically solve the IVP (8.3), we are given initial data $\mathbf{U}^0 = \mathbf{u}_0$, and want to compute approximations $\mathbf{U}^1, \mathbf{U}^2, \ldots$ such that

$$\mathbf{U}^n \approx \mathbf{u}(t_n).$$

Definition 8.26. The *(forward) Euler's method* solves the IVP (8.3) by

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^n, t_n), \tag{8.18}$$

which is based on replacing $\mathbf{u}'(t_n)$ with the forward difference $(\mathbf{U}^{n+1} - \mathbf{U}^n)/k$ and $\mathbf{u}(t_n)$ with $\mathbf{U}^n$ in $\mathbf{f}(\mathbf{u}, t)$.

Definition 8.27. The *backward Euler's method* solves the IVP (8.3) by

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^{n+1}, t_{n+1}), \tag{8.19}$$

which is based on replacing $\mathbf{u}'(t_{n+1})$ with the backward difference $(\mathbf{U}^{n+1} - \mathbf{U}^n)/k$ and $\mathbf{u}(t_{n+1})$ with $\mathbf{U}^{n+1}$ in $\mathbf{f}(\mathbf{u}, t)$.

Definition 8.28. The *trapezoidal method* is

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \frac{k}{2}\left(\mathbf{f}(\mathbf{U}^n, t_n) + \mathbf{f}(\mathbf{U}^{n+1}, t_{n+1})\right). \tag{8.20}$$

Definition 8.29. The *midpoint (or leapfrog) method* is

$$\mathbf{U}^{n+1} = \mathbf{U}^{n-1} + 2k\mathbf{f}(\mathbf{U}^n, t_n). \tag{8.21}$$

Example 8.30. Applying Euler's method (8.18) with step size $k = 0.2$ to solve the IVP

$$u'(t) = u, \quad u(0) = 1, \quad t \in [0, 1],$$

yields the following table:

n	U^n	$kf(U^n, t_n)$
0	1	0.2
1	1.2	$0.2 \times 1.2 = 0.24$
2	1.44	$0.2 \times 1.44 = 0.288$
3	1.728	$0.2 \times 1.728 = 0.3456$
4	2.0736	$0.2 \times 2.0736 = 0.41472$
5	2.48832	

8.4 Accuracy and convergence

Definition 8.31. The *local truncation error* (LTE) is the error caused by replacing continuous derivatives with finite difference formulas.

Example 8.32. For the leapfrog method, the local truncation error is

$$\begin{aligned}\boldsymbol{\tau}^n &= \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1})}{2k} - \mathbf{f}(\mathbf{u}(t_n), t_n) \\ &= \left[\mathbf{u}'(t_n) + \frac{1}{6}k^2\mathbf{u}'''(t_n) + O(k^4)\right] - \mathbf{u}'(t_n) \\ &= \frac{1}{6}k^2\mathbf{u}'''(t_n) + O(k^4).\end{aligned}$$

Definition 8.33. For a numerical method of the form

$$\mathbf{U}^{n+1} = \boldsymbol{\phi}(\mathbf{U}^{n+1}, \mathbf{U}^n, \ldots, \mathbf{U}^{n-m}),$$

the *one-step error* is defined by

$$\mathcal{L}^n := \mathbf{u}(t_{n+1}) - \boldsymbol{\phi}(\mathbf{u}(t_{n+1}), \mathbf{u}(t_n), \ldots, \mathbf{u}(t_{n-m})). \tag{8.22}$$

In other words, $\mathcal{L}^n$ is the error that would be introduced in one time step if the past values $\mathbf{U}^n, \mathbf{U}^{n-1}, \ldots$ were all taken to be the exact values from $\mathbf{u}(t)$.

Example 8.34. For the leapfrog method, the one-step error is

$$\begin{aligned}\mathcal{L}^n &= \mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1}) - 2k\mathbf{f}(\mathbf{u}(t_n), t_n) \\ &= \frac{1}{3}k^3\mathbf{u}'''(t_n) + O(k^5) \\ &= 2k\boldsymbol{\tau}^n.\end{aligned}$$

Definition 8.35. The *solution error* of a numerical method for solving the IVP in Definition 8.4 is

$$\mathbf{E}^N := \mathbf{U}^{T/k} - \mathbf{u}(T); \qquad \mathbf{E}^n = \mathbf{U}^n - \mathbf{u}(t_n). \tag{8.23}$$

Definition 8.36. A numerical method is *convergent* if the application of it to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\lim_{\substack{k\to 0\\ Nk=T}} \mathbf{U}^N = \mathbf{u}(T) \tag{8.24}$$

for every fixed $T > 0$.

8.5 Analysis of Euler's methods

8.5.1 Linear IVPs

In this section, we consider the convergence of Euler's method for solving linear IVPs of the form

$$\begin{cases}\mathbf{u}'(t) = \lambda\mathbf{u}(t) + \mathbf{g}(t), \\ \mathbf{u}(0) = \mathbf{u}_0,\end{cases} \tag{8.25}$$

where λ is either a scalar or a diagonal matrix.

Lemma 8.37. For the linear IVP (8.25), the solution errors and the local truncation error of Euler's method satisfy

$$\mathbf{E}^{n+1} = (1 + k\lambda)\mathbf{E}^n - k\boldsymbol{\tau}^n. \tag{8.26}$$

Proof. By Definition 8.31, we have

$$\begin{aligned}\boldsymbol{\tau}^n &= \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)}{k} - \mathbf{u}'(t_n) \\ &= \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)}{k} - (\lambda\mathbf{u}(t_n) + \mathbf{g}(t_n)),\end{aligned}$$

and therefore

$$\mathbf{u}(t_{n+1}) = (1 + k\lambda)\mathbf{u}(t_n) + k\mathbf{g}(t_n) + k\boldsymbol{\tau}^n.$$

Euler's method applied to the linear IVP (8.25) reads

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k(\lambda\mathbf{U}^n + \mathbf{g}(t_n)) = (1 + k\lambda)\mathbf{U}^n + k\mathbf{g}(t_n).$$

Subtracting the above two equations yields (8.26). □

Lemma 8.38. For the linear IVP (8.25), the solution error and the local truncation errors of Euler's method satisfy

$$\mathbf{E}^n = (1 + k\lambda)^n\mathbf{E}^0 - k\sum_{m=1}^{n}(1 + k\lambda)^{n-m}\boldsymbol{\tau}^{m-1}. \tag{8.27}$$

Proof. We proceed by induction on n.

The induction basis holds because of (8.26). Suppose (8.27) holds for all integers no greater than n. Then for $n + 1$, we have

$$\begin{aligned}\mathbf{E}^{n+1} &= (1 + k\lambda)\mathbf{E}^n - k\boldsymbol{\tau}^n \\ &= (1 + k\lambda)^{n+1}\mathbf{E}^0 - k\sum_{m=1}^{n+1}(1 + k\lambda)^{n+1-m}\boldsymbol{\tau}^{m-1},\end{aligned}$$

where the first equality follows from (8.26) and the second from the induction hypothesis. □

Theorem 8.39. Euler's method is convergent for solving the linear IVP (8.25).

Proof. We have

$$|1+k\lambda| \le 1+|k\lambda| \le e^{k|\lambda|},$$

and hence for $m < n \le T/k$

$$(1+k\lambda)^{n-m} \le e^{(n-m)k|\lambda|} \le e^{nk|\lambda|} \le e^{|\lambda|T},$$

then Lemma 8.38 yields

$$\begin{aligned}\|\mathbf{E}^n\| &\le e^{|\lambda|T}\left(\|\mathbf{E}^0\| + k\sum_{m=1}^{n}\|\boldsymbol{\tau}^{m-1}\|\right)\\ &\le e^{|\lambda|T}\left(\|\mathbf{E}^0\| + nk\max_{1\le m\le n}\|\boldsymbol{\tau}^{m-1}\|\right).\end{aligned}$$

For Euler's method, the local truncation error is

$$\boldsymbol{\tau}^n = \frac{\mathbf{u}(t_{n+1})-\mathbf{u}(t_n)}{k} - \mathbf{u}'(t_n) = \frac{1}{2}k\mathbf{u}''(t_n) + O(k^2),$$

hence

$$\|\mathbf{E}^N\| \le e^{|\lambda|T}(\|\mathbf{E}^0\| + TO(k)) = O(k),$$

where we have assumed that $\|\mathbf{E}^0\| = O(k)$. □

8.5.2 Nonlinear IVPs

Lemma 8.40. Consider a nonlinear IVP of the form

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t), t),$$

where $\mathbf{f}(\mathbf{u}, t)$ is continuous in t and is Lipschitz continuous in $\mathbf{u}$ with L as the Lipschitz constant. Euler's method satisfies

$$\|\mathbf{E}^{n+1}\| \le (1+kL)\|\mathbf{E}^n\| + k\|\boldsymbol{\tau}^n\|. \tag{8.28}$$

Proof. The definition of the local truncation error yields

$$\boldsymbol{\tau}^n = \frac{\mathbf{u}(t_{n+1})-\mathbf{u}(t_n)}{k} - \mathbf{f}(\mathbf{u}(t_n), t_n),$$

and hence

$$\mathbf{u}(t_{n+1}) = \mathbf{u}(t_n) + k\mathbf{f}(\mathbf{u}(t_n), t_n) + k\boldsymbol{\tau}^n,$$

the Euler's method is

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^n, t_n),$$

subtracting the above two equations gives

$$\mathbf{E}^{n+1} = \mathbf{E}^n + k\left(\mathbf{f}(\mathbf{U}^n, t_n) - \mathbf{f}(\mathbf{u}(t_n), t_n)\right) - k\boldsymbol{\tau}^n,$$

the triangle inequality and Lipschitz continuity of $\mathbf{f}$ yield

$$\begin{aligned}\|\mathbf{E}^{n+1}\| &\le \|\mathbf{E}^n\| + k\|\mathbf{f}(\mathbf{U}^n, t_n) - \mathbf{f}(\mathbf{u}(t_n), t_n)\| + k\|\boldsymbol{\tau}^n\|\\ &\le (1+kL)\|\mathbf{E}^n\| + k\|\boldsymbol{\tau}^n\|.\end{aligned}$$ □

Theorem 8.41. For the nonlinear IVP in Lemma 8.40, Euler's method is convergent.

Proof. Follow the same procedure as in section 8.5.1 to show that

$$\|\mathbf{E}^N\| \le e^{LT}T\|\boldsymbol{\tau}\| = O(k) \text{ as } k \to 0.$$ □

8.5.3 Zero stability and absolute stability

Example 8.42. Consider the scalar IVP

$$u'(t) = \lambda(u - \cos t) - \sin t,$$

with $\lambda = -2100$ and $u(0) = 1$. The exact solution is clearly

$$u(t) = \cos t.$$

The following table shows the error at time $T = 2$ when Euler's method is used with various values of k.

k	$E(T)$
2.00e-4	1.98e-8
4.00e-4	3.96e-8
8.00e-4	7.92e-8
9.50e-4	3.21e-7
9.76e-4	5.88e+35
1.00e-3	1.45e+76

The first three lines confirm the first-order accuracy of Euler's method, but something dramatic happens between $k = 9.76$e-4 and $k = 9.50$e-4. What's going on?

Definition 8.43. The Euler's method

$$U^{n+1} = (1+k\lambda)U^n$$

for solving the scalar IVP

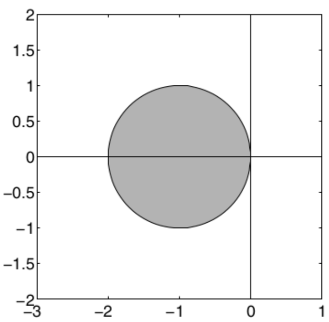
$$u'(t) = \lambda u(t) \tag{8.29}$$

is *absolutely stable* if

$$|1+k\lambda| \le 1. \tag{8.30}$$

Definition 8.44. The *region of absolute stability* for Euler's method applied to (8.29) is the set of all points

$$\{z \in \mathbb{C} : |1+z| \le 1\}. \tag{8.31}$$

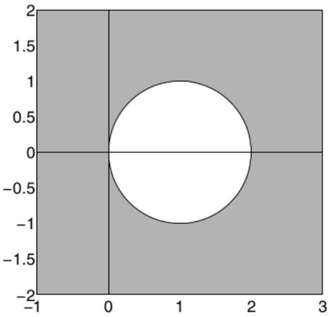


Example 8.45. The backward Euler's method applied to (8.29) reads

$$U^{n+1} = U^n + k\lambda U^{n+1} \Rightarrow U^{n+1} = \frac{1}{1-k\lambda}U^n.$$

Hence the region of absolute stability for backward Euler's method is

$$\{z \in \mathbb{C} : |1-z| \ge 1\}. \tag{8.32}$$



Lemma 8.46. Consider an autonomous, homogeneous, and linear system of IVPs

$$\mathbf{u}'(t) = A\mathbf{u} \tag{8.33}$$

where $\mathbf{u} \in \mathbb{R}^N$, $N > 1$, and A is diagonalizable with eigenvalues as λ_i's. Euler's method is absolutely stable if each $z_i := k\lambda_i$ is within the stability region (8.31).

Proof. Applying Euler's method to (8.33) gives

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

Since A is diagonalizable, we have $AR = R\Lambda$ where R contains the eigenvectors of A that span $\mathbb{R}^N$. then

$$R^{-1}\mathbf{U}^{n+1} = R^{-1}(I + kA)RR^{-1}\mathbf{U}^n.$$

Set $\mathbf{V} := R^{-1}\mathbf{U}$ and we have

$$\mathbf{V}^{n+1} = (I + k\Lambda)\mathbf{V}^n.$$

After advancing $\mathbf{V}^0$ to $\mathbf{V}^n$, we use $\mathbf{U}^n = R\mathbf{V}^n$ to recover the solution of (8.33). □

Definition 8.47. The *law of mass action* states that the rate of a chemical reaction is proportional to the product of the concentration of the reacting substances, with each concentration raised to a power equal to the coefficient that occurs in the reaction.

Example 8.48. For the reaction

$$\alpha\mathrm{A} + \beta\mathrm{B} \underset{\mathrm{k}_-}{\overset{\mathrm{k}_+}{\rightleftharpoons}} \sigma\mathrm{S} + \tau\mathrm{T},$$

the forward reaction rate is $k_+[A]^\alpha[B]^\beta$ and the backward reaction rate is $k_-[S]^\sigma[T]^\tau$.

Example 8.49. Consider

$$\mathrm{A} + \mathrm{B} \underset{\mathrm{c}_2}{\overset{\mathrm{c}_1}{\rightleftharpoons}} \mathrm{AB}.$$

Let

$$\mathbf{u} := \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} [A] \\ [B] \\ [AB] \end{bmatrix}.$$

Then we have

$$\begin{aligned} u_1' &= -c_1u_1u_2 + c_2u_3; \\ u_2' &= -c_1u_1u_2 + c_2u_3; \\ u_3' &= c_1u_1u_2 - c_2u_3, \end{aligned}$$

which can be written more compactly as

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}).$$

Let $\mathbf{v}(t) := \mathbf{u}(t) - \bar{\mathbf{u}}$ with $\bar{\mathbf{u}}$ independent on time. Then

$$\begin{aligned} \mathbf{v}'(t) = \mathbf{u}'(t) &= \mathbf{f}(\mathbf{u}(t)) = \mathbf{f}(\mathbf{v} + \bar{\mathbf{u}}) \\ &= \mathbf{f}(\bar{\mathbf{u}}) + \mathbf{f}'(\bar{\mathbf{u}})\mathbf{v}(t) + O(\|\mathbf{v}\|^2), \end{aligned}$$

and hence

$$\mathbf{v}'(t) = A\mathbf{v}(t) + \mathbf{b},$$

where $A = \mathbf{f}'(\bar{\mathbf{u}})$ is the Jacobian, i.e.,

$$A = \begin{bmatrix} -c_1u_2 & -c_1u_1 & c_2 \\ -c_1u_2 & -c_1u_1 & c_2 \\ c_1u_2 & c_1u_1 & -c_2 \end{bmatrix},$$

with eigenvalues $\lambda_1 = -c_1(u_1 + u_2) - c_2$ and $\lambda_2 = \lambda_3 = 0$. Since $u_1 + u_2$ is simply the total concentration of species A and B present, they can be bounded by $u_1(0)+u_2(0)+u_3(0)$.

Example 8.50. For the reaction

$$\mathrm{A} \underset{\mathrm{c}_2}{\overset{\mathrm{c}_1}{\rightleftharpoons}} \mathrm{B},$$

we obtain the linear IVPs

$$\begin{cases} u_1' = -c_1u_1 + c_2u_2; \\ u_2' = c_1u_1 - c_2u_2. \end{cases}$$

8.5.4 Review of Jordan canonical form

Theorem 8.51 (Factorization of a polynomial over $\mathbb{C}$)**.** If $p \in \mathcal{P}(\mathbb{C})$ is a nonconstant polynomial, then p has a unique factorization (except for the order of the factors) of the form

$$p(z) = c(z - \lambda_1)\cdots(z - \lambda_m), \tag{8.34}$$

where $c, \lambda_1, \ldots, \lambda_m \in \mathbb{C}$.

Definition 8.52. Let $A \in \mathbb{C}^{m\times m}$, then the *characteristic polynomial* of A is

$$p_A(z) = \det(zI - A). \tag{8.35}$$

Proposition 8.53. Let $A \in \mathbb{C}^{m\times m}$, then λ is an eigenvalue of A iff λ is a root of the characteristic polynomial of A.

Exercise 8.54. Show that

$$p_M(z) = z^s + \sum_{j=0}^{s-1} \alpha_j z^j.$$

is the characteristic polynomial of

$$M = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{s-2} & -\alpha_{s-1} \end{bmatrix} \in \mathbb{C}^{s\times s}. \tag{8.36}$$

Definition 8.55. If the characteristic polynomial $p_A(z)$ has a factor $(z-\lambda)^n$, then λ is said to have *algebraic multiplicity* $m_a(\lambda) = n$.

Definition 8.56. Let λ be an eigenvalue of $A \in \mathbb{C}^{m\times m}$, the *eigenspace* of A corresponding to λ is

$$\begin{aligned}\mathcal{N}(A-\lambda I) &= \{\mathbf{u}\in\mathbb{C}^m : (A-\lambda I)\mathbf{u} = \mathbf{0}\} \\ &= \{\mathbf{u}\in\mathbb{C}^m : A\mathbf{u} = \lambda\mathbf{u}\}.\end{aligned} \tag{8.37}$$

The dimension of $\mathcal{N}(A-\lambda I)$ is the *geometric multiplicity* $m_g(\lambda)$ of λ.

Proposition 8.57. Geometric multiplicity and algebraic multiplicity satisfy

$$1 \le m_g(\lambda) \le m_a(\lambda). \tag{8.38}$$

Definition 8.58. An eigenvalue λ of A is *defective* if

$$m_g(\lambda) < m_a(\lambda). \tag{8.39}$$

A is *defective* if A has one or more defective eigenvalues.

Proposition 8.59. A nondefective matrix A is diagnolizable, i.e.,

$$\exists \text{ nonsingular } R \text{ s.t. } R^{-1}AR = \Lambda \text{ is diagonal.} \tag{8.40}$$

Theorem 8.60 (Schur decomposition)**.** For each square matrix A, there exists a unitary matrix Q such that

$$A = QUQ^{-1}, \tag{8.41}$$

where U is upper triangular.

Definition 8.61. A *Jordan block* of order k has the form

$$J(\lambda, k) = \lambda I_k + S_k, \tag{8.42}$$

where

$$(S_k)_{i,j} = \begin{cases} 1, & i = j-1, \\ 0, & \text{otherwise.}\end{cases}$$

Example 8.62. The Jordan blocks of orders 1, 2, and 3 are

$$J(\lambda,1) = \lambda, \quad J(\lambda,2) = \begin{bmatrix}\lambda & 1\\ 0 & \lambda\end{bmatrix}, \quad J(\lambda,3) = \begin{bmatrix}\lambda & 1 & 0\\ 0 & \lambda & 1\\ 0 & 0 & \lambda\end{bmatrix}.$$

Theorem 8.63 (Jordan canonical form)**.** Every square matrix A can be expressed as

$$A = RJR^{-1}, \tag{8.43}$$

where R is invertible and J is a block diagonal matrix of the form

$$J = \begin{bmatrix} J(\lambda_1,k_1) & & & \\ & J(\lambda_2,k_2) & & \\ & & \ddots & \\ & & & J(\lambda_s,k_s)\end{bmatrix}. \tag{8.44}$$

Each $J(\lambda_i,k_i)$ is a Jordan block of some order k_i and $\sum_{i=1}^s k_i = m$. If λ is an eigenvalue of A with algebraic multiplicity m_a and geometric multiplicity m_g, then λ appears in m_g blocks and the sum of the orders of these blocks is m_a.

8.6 Linear multistep methods

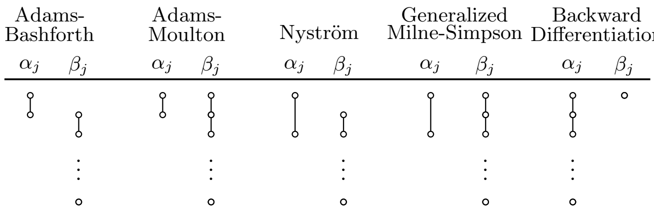
Definition 8.64. For solving the IVP (8.3), an *s-step linear multistep method* (LMM) has the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = k\sum_{j=0}^{s}\beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{8.45}$$

where $\alpha_s = 1$ is assumed WLOG.

Definition 8.65. An LMM (8.45) is *explicit* if $\beta_s = 0$; otherwise it is *implicit*.

8.6.1 Classical formulas



Definition 8.66. An *Adams formula* is an LMM of the form

$$\mathbf{U}^{n+s} = \mathbf{U}^{n+s-1} + k\sum_{j=0}^{s}\beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{8.46}$$

where β_j's are chosen to maximize the order of accuracy.

Definition 8.67. An *Adams-Bashforth formula* is an Adams formula with $\beta_s = 0$. An *Adams-Moulton formula* is an Adams formula with $\beta_s \neq 0$.

Example 8.68. Euler's method is the 1-step Adams-Bashforth formula with

$$s = 1,\ \alpha_1 = 1,\ \alpha_0 = -1,\ \beta_1 = 0,\ \beta_0 = 1.$$

Example 8.69. The trapezoidal method is a 1-step Adams-Moulton formula with

$$s = 1,\ \alpha_1 = 1,\ \alpha_0 = -1,\ \beta_1 = \beta_0 = \frac{1}{2}.$$

Another 1-step Adams-Moulton formula is the backward Euler's method.

Definition 8.70. A *Nyström formula* is an LMM of the form

$$\mathbf{U}^{n+s} = \mathbf{U}^{n+s-2} + k\sum_{j=0}^{s-1}\beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{8.47}$$

where β_j's are chosen to give order s.

Example 8.71. The midpoint method is the 2-step Nyström formula with

$$s = 2,\ \alpha_2 = 1,\ \alpha_1 = 0,\ \alpha_0 = -1,\ \beta_1 = 2,\ \beta_0 = 0.$$

Definition 8.72. A *backward differentiation formula* (BDF) is an LMM of the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = k\beta_s \mathbf{f}(\mathbf{U}^{n+s}, t_{n+s}), \tag{8.48}$$

where α_j's are chosen to give order s.

Example 8.73. The backward Euler's method is the 1-step BDF with

$$s = 1,\ \alpha_1 = \beta_1 = 1,\ \alpha_0 = -1.$$

8.6.2 Consistency and accuracy

Definition 8.74. The *characteristic polynomials* or *generating polynomials* for the LMM (8.45) are

$$\rho(\zeta) = \sum_{j=0}^{s} \alpha_j \zeta^j; \qquad \sigma(\zeta) = \sum_{j=0}^{s} \beta_j \zeta^j. \tag{8.49}$$

Example 8.75. The forward Euler's method (8.18) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = 1, \tag{8.50}$$

while the backward Euler's method (8.19) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = \zeta. \tag{8.51}$$

Example 8.76. The trapezoidal method (8.20) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = \frac{1}{2}(\zeta + 1), \tag{8.52}$$

and the midpoint method (8.21) has

$$\rho(\zeta) = \zeta^2 - 1, \qquad \sigma(\zeta) = 2\zeta. \tag{8.53}$$

Notation 9. Denote by Z a *time shift operator* that acts on both discrete functions according to

$$Z\mathbf{U}^n = \mathbf{U}^{n+1} \tag{8.54}$$

and on continuous functions according to

$$Z\mathbf{u}(t) = \mathbf{u}(t + k). \tag{8.55}$$

Definition 8.77. The *difference operator of an LMM* is an operator $\mathcal{L}$ on the linear space of continuously differentiable functions given by

$$\mathcal{L} = \rho(Z) - k\mathcal{D}\sigma(Z), \tag{8.56}$$

where $\mathcal{D}\mathbf{u}(t_n) = \mathbf{u}_t(t_n) := \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_n)$, Z the time shift operator, and ρ, σ are the characteristic polynomials for the LMM.

Lemma 8.78. The one-step error of the LMM (8.45) is

$$\mathcal{L}\mathbf{u}(t_n) = C_0\mathbf{u}(t_n) + C_1 k\mathbf{u}_t(t_n) + C_2 k^2 \mathbf{u}_{tt}(t_n) + \cdots, \tag{8.57}$$

where

$$\begin{aligned}
C_0 &= \textstyle\sum_{j=0}^{s} \alpha_j \\
C_1 &= \textstyle\sum_{j=0}^{s} (j\alpha_j - \beta_j) \\
C_2 &= \textstyle\sum_{j=0}^{s} \left(\frac{1}{2} j^2 \alpha_j - j\beta_j\right) \\
&\vdots \\
C_q &= \textstyle\sum_{j=0}^{s} \left(\frac{1}{q!} j^q \alpha_j - \frac{1}{(q-1)!} j^{q-1} \beta_j\right).
\end{aligned} \tag{8.58}$$

Proof. By definition of the one-step error (8.22), we have

$$\begin{aligned}
\mathcal{L}\mathbf{u}(t_n) &= \sum_{j=0}^{s} \alpha_j \mathbf{u}(t_{n+j}) - k \sum_{j=0}^{s} \beta_j \mathbf{f}(\mathbf{u}(t_{n+j}), t_{n+j}) \\
&= \sum_{j=0}^{s} \alpha_j \mathbf{u}(t_{n+j}) - k \sum_{j=0}^{s} \beta_j \mathbf{u}'(t_{n+j}).
\end{aligned}$$

Taylor's theorem yields

$$\begin{aligned}
\mathbf{u}(t_{n+j}) &= \mathbf{u}(t_n) + jk\mathbf{u}'(t_n) + \frac{1}{2}(jk)^2 \mathbf{u}''(t_n) + \cdots \\
\mathbf{u}'(t_{n+j}) &= \mathbf{u}'(t_n) + jk\mathbf{u}''(t_n) + \frac{1}{2}(jk)^2 \mathbf{u}'''(t_n) + \cdots
\end{aligned}$$

Substitution of the above into $\mathcal{L}\mathbf{u}(t_n)$ yields (8.57). □

Notation 10. We write $f(x) = \Theta(g(x))$ as $x \to 0$ if there exist constants $C, C' > 0$ and $x_0 > 0$ such that $Cg(x) \le f(x) \le C'g(x)$ for all $x \le x_0$.

Definition 8.79. An LMM has *order of accuracy p* if

$$\mathcal{L}\mathbf{u}(t_n) = \Theta(k^{p+1}) \text{ as } k \to 0, \tag{8.59}$$

i.e., if in (8.58) we have $C_0 = C_1 = \cdots = C_p = 0$ and $C_{p+1} \ne 0$. Then C_{p+1} is called the *error constant*.

Definition 8.80. An LMM is *preconsistent* if $\rho(1) = 0$, i.e. $\sum_{i=0}^{s} \alpha_i = 0$ or $\sum_{i=0}^{s-1} \alpha_i = -1$.

Definition 8.81. An LMM is *consistent* if it has order of accuracy $p \ge 1$.

Example 8.82. For Euler's method, the coefficients C_j's in (8.58) can be computed directly from Example 8.68 as $C_0 = C_1 = 0, C_2 = \frac{1}{2}, C_3 = \frac{1}{6}$.

Exercise 8.83. Compute the first five coefficients C_j's of the trapezoidal rule and the midpoint rule from Examples 8.69 and 8.71.

Example 8.84. A necessary condition for $\|\mathbf{E}^n\| = O(k)$ is $\|\mathcal{L}\mathbf{u}(t_n)\| = O(k^2)$ since there are $\frac{T}{k}$ time steps, and hence the first two terms in (8.57) should be zero, i.e.,

$$\sum_{j=0}^{s} \alpha_j = 0, \qquad \sum_{j=0}^{s} j\alpha_j = \sum_{j=0}^{s} \beta_j, \tag{8.60}$$

which is equivalent to

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1). \tag{8.61}$$

Second-order accuracy further requires

$$\frac{1}{2} \sum_{j=0}^{s} j^2 \alpha_j = \sum_{j=0}^{s} j\beta_j.$$

In general, pth-order accuracy requires (8.60) and

$$\forall q = 2, \ldots, p, \quad \sum_{j=0}^{s} \frac{1}{q!} j^q \alpha_j = \sum_{j=0}^{s} \frac{1}{(q-1)!} j^{q-1} \beta_j. \tag{8.62}$$

Exercise 8.85. Express conditions of $\mathcal{L} = O(k^3)$ using characteristic polynomials.

Exercise 8.86. Derive coefficients of LMMs shown below by the method of undetermined coefficients and a programming language with symbolic computation such as `Matlab`.

Adams-Bashforth formulas in Definition 8.67

s	p	β_s	β_{s-1}	β_{s-2}	β_{s-3}	β_{s-4}
1	1	0	1			
2	2	0	$\frac{3}{2}$	$-\frac{1}{2}$		
3	3	0	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	
4	4	0	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$

Adams-Moulton formulas in Definition 8.67

s	p	β_s	β_{s-1}	β_{s-2}	β_{s-3}	β_{s-4}
1	1	1				
1	2	$\frac{1}{2}$	$\frac{1}{2}$			
2	3	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		
3	4	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	
4	5	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

BDF formulas in Definition 8.72

s	p	α_s	α_{s-1}	α_{s-2}	α_{s-3}	α_{s-4}	β_s
1	1	1	-1				1
2	2	1	$-\frac{4}{3}$	$\frac{1}{3}$			$\frac{2}{3}$
3	3	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$		$\frac{6}{11}$
4	4	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	$\frac{12}{25}$

Example 8.87. To derive coefficients of the 2nd-order Adams-Bashforth formula, we interpolate $\mathbf{f}(t)$ by a linear polynomial

$$q(t) = \mathbf{f}^{n+1} - \frac{t_{n+1} - t}{k}(\mathbf{f}^{n+1} - \mathbf{f}^n)$$

and then calculate

$$\mathbf{U}^{n+2} - \mathbf{U}^{n+1} = \int_{t_{n+1}}^{t_{n+2}} q(t)\mathrm{d}t = \frac{3}{2}k\mathbf{f}^{n+1} - \frac{1}{2}k\mathbf{f}^n.$$

Lemma 8.88. An LMM with $\sigma(1) \neq 0$ has order of accuracy p if and only if

$$\frac{\rho(e^\kappa)}{\sigma(e^\kappa)} = \kappa + \Theta(\kappa^{p+1}) \qquad \text{as } \kappa \to 0. \tag{8.63}$$

where $\kappa := k\mathcal{D}$.

Proof. By Taylor's theorem,

$$\mathbf{u}(t_{n+1}) = \mathbf{u}(t_n) + k\mathbf{u}_t(t_n) + \frac{1}{2}k^2\mathbf{u}_{tt}(t_n) + \cdots$$

By Notation 9, we also have $\mathbf{u}(t_{n+1}) = Z\mathbf{u}(t_n)$. A comparison of the two equalities yields

$$Z = 1 + (k\mathcal{D}) + \frac{1}{2!}(k\mathcal{D})^2 + \cdots + \frac{1}{n!}(k\mathcal{D})^n + \cdots = e^{k\mathcal{D}},$$

where the last step follows from Definition 8.21. Set $\kappa = k\mathcal{D}$ and we have

$$\mathcal{L} = \rho(e^\kappa) - \kappa\sigma(e^\kappa) = C_0 + C_1\kappa + C_2\kappa^2 + \cdots,$$

where C_j's are the coefficients in Lemma 8.78. By Definition 8.79, an LMM has order of accuracy p if and only if the term between the equal signs in the above equation is $\Theta(\kappa^{p+1})$ as $\kappa \to 0$. Since $\sigma(e^\kappa)$ is an analytic function of κ and it is nonzero at $\kappa = 0$, we can divide through to get (8.63). □

Theorem 8.89. An LMM with $\sigma(1) \neq 0$ has order of accuracy p if and only if

$$\begin{aligned}\frac{\rho(z)}{\sigma(z)} =\ & \log z + \Theta\left((z-1)^{p+1}\right) \\ =\ & \left[(z-1) - \tfrac{1}{2}(z-1)^2 + \tfrac{1}{3}(z-1)^3 - \cdots\right] \\ & +\Theta((z-1)^{p+1}).\end{aligned} \tag{8.64}$$

as $z \to 1$.

Proof. To get from (8.63) to the first equality, we make the change of variables $z = e^\kappa, \kappa = \log z$, noting that $\Theta(\kappa^{p+1})$ as $\kappa \to 0$ has the same meaning as $\Theta\left((z-1)^{p+1}\right)$ as $z \to 1$ since $e^\kappa = 1$ and $\mathrm{d}(e^\kappa)/\mathrm{d}\kappa \neq 0$ at $\kappa = 0$. The second equality is just the usual Taylor series for $\log z$ at 1. □

Example 8.90. The trapezoidal rule has $\rho(z) = z - 1$ and $\sigma(z) = \frac{1}{2}(z+1)$. A comparison of (8.64) with the expansion

$$\frac{\rho(z)}{\sigma(z)} = \frac{z-1}{\frac{1}{2}(z+1)} = (z-1)\left[1 - \frac{z-1}{2} + \frac{(z-1)^2}{4} - \cdots\right]$$

confirms that the trapezoidal rule has order 2 with error constant $-\frac{1}{12}$.

Exercise 8.91. For the third-order BDF formula in Definition 8.72 and Exercise 8.86, derive its characteristic polynomials and apply Theorem 8.89 to verify that the order of accuracy is indeed 3.

Exercise 8.92. Prove that an s-step LMM has order of accuracy p if and only if, when applied to an ODE $u_t = q(t)$, it gives exact results whenever q is a polynomial of degree $< p$, but not whenever q is a polynomial of degree p. Assume arbitrary continuous initial data u_0 and exact numerical initial data $v^0, \cdots, v^{s-1}$.

8.6.3 Zero stability

Example 8.93 (A consistent but unstable LMM)**.** The LMM

$$\mathbf{U}^{n+2} - 3\mathbf{U}^{n+1} + 2\mathbf{U}^n = -k\mathbf{f}(\mathbf{U}^n, t_n) \tag{8.65}$$

has a one-step error given by

$$\begin{aligned}\mathcal{L}\mathbf{u}(t_n) &= \mathbf{u}(t_{n+2}) - 3\mathbf{u}(t_{n+1}) + 2\mathbf{u}(t_n) + k\mathbf{u}'(t_n) \\ &= \frac{1}{2}k^2\mathbf{u}''(t_n) + O(k^3),\end{aligned}$$

so the method is consistent with first-order accuracy. But the solution error may not exhibit first order accuracy, or even convergence. Consider the trivial IVP

$$u'(t) = 0, \qquad u(0) = 0,$$

with solution $u(t) \equiv 0$. The LMM (8.65) reads in this case

$$U^{n+2} = 3U^{n+1} - 2U^n \Rightarrow U^{n+2} - U^{n+1} = 2(U^{n+1} - U^n),$$

and therefore

$$U^n = 2U^0 - U^1 + 2^n(U^1 - U^0).$$

If we take $U^0 = 0$ and $U^1 = k$, then

$$U^n = k(2^n - 1) = k(2^{T/k} - 1) \to +\infty \text{ as } k \to 0.$$

Definition 8.94. An s-step LMM is *zero-stable* if all solutions $\{\mathbf{U}^n\}$ of the recurrence

$$\rho(Z)\mathbf{U}^n = \sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = \mathbf{0} \tag{8.66}$$

are bounded as $n \to +\infty$.

Theorem 8.95. An LMM is zero-stable if and only if all the roots of $\rho(z)$ satisfy $|z| \le 1$, and any root with $|z| = 1$ is simple.

Proof. (8.66) is equivalent to $\mathbf{U}^{n+s} + \sum_{j=0}^{s-1} \alpha_j \mathbf{U}^{n+j} = 0$, and this s-step recurrence formula can be expressed as a one-step matrix operation

$$\mathbf{V}^{n+1} = M\mathbf{V}^n,$$

where M is the *companion matrix* (8.36) and

$$\mathbf{V}^n = \begin{bmatrix} u^n & u^{n+1} & \cdots & u^{n+s-1} \end{bmatrix}^T.$$

Hence

$$\mathbf{V}^n = M^n \mathbf{V}^0.$$

By Exercise 8.54, the characteristic polynomial of M is $\rho(z)$, i.e., $p_M(z) = \rho(z)$. Therefore the set of eigenvalues of M is the same as the set of roots of ρ, and these eigenvalues determine how the powers M^n behave asymptotically as $n \to +\infty$. The scalar sequence $\{\mathbf{U}^n\}_{n=0}^{+\infty}$ is bounded as $n \to +\infty$ if and only if the vector sequence $\{\mathbf{V}^n\}$ is bounded, and $\{\mathbf{V}^n\}$ is bounded if and only if all elements of M^n is bounded. Since $\|\mathbf{V}^n\| \le \|M^n\|\|\mathbf{V}^0\|$, the zero-stability is now equivalent to the power-boundedness of M.

By Theorem 8.63, we have

$$M = RJR^{-1} \Rightarrow M^n = RJ^nR^{-1}.$$

Therefore M^n's growth or boundedness is determined by the boundedness of

$$J_i^n = \begin{bmatrix} \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \binom{n}{2}\lambda_i^{n-2} & \cdots & \binom{n}{m_i-1}\lambda_i^{n-m_i+1} \\ & \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \cdots & \binom{n}{m_i-2}\lambda_i^{n-m_i+2} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} \\ & & & & \lambda_i^n \end{bmatrix},$$

which follows from $J_i^n = (\lambda_i I + \eta)^n$ where η is the nilpotent matrix with $\eta^{m_i} = \mathbf{0}$,

$$\eta_{ij} = \begin{cases} 1 & \text{if } j - i = 1; \\ 0 & \text{otherwise.} \end{cases}$$

By Definition 8.56, the dimension of the eigenspace of the companion matrix M is 1 for each eigenvalue of M because the upper-right $(s-1)\times(s-1)$ block of $zI - M$ is nonsingular for any $z \in \mathbb{C}$. Hence the geometric multiplicity $m_g(\lambda)$ is 1 for any eigenvalue λ of M. By Theorem 8.63, there is exactly one Jordan block for each eigenvalue of M.

As $n \to \infty$, the nonzero elements of J_i^n approach 0 if $|\lambda_i| < 1$ and ∞ if $|\lambda_i| > 1$. For $|\lambda_i| = 1$, they are bounded in the case of a 1×1 block, but unbounded if $m_i \ge 2$. □

8.6.4 Linear difference equations

Definition 8.96. A *system of linear difference equations* is a set of equations of the form

$$X_n = A_n X_{n-1} + \phi_n, \tag{8.67}$$

where $n, s \in \mathbb{N}^+$, $X_n \in \mathbb{C}^s$, $\phi_n \in \mathbb{C}^s$, and $A_n \in \mathbb{C}^{s\times s}$. With the initial vector X_0 specified, the system of linear difference equations becomes an initial value problem. The system is *homogeneous* if $\phi_n = \mathbf{0}$.

Example 8.97. A linear difference equation of the form

$$y_n = \alpha_{n1}y_{n-1} + \alpha_{n2}y_{n-2} + \cdots + \alpha_{ns}y_{n-s} + \psi_n$$

can be easily recast in the form (8.67) by writing

$$X_n = \begin{bmatrix} y_n \\ y_{n-1} \\ \vdots \\ y_{n-s+1} \end{bmatrix}, A_n = \begin{bmatrix} \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{ns} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \phi_n = \begin{bmatrix} \psi_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Theorem 8.98. The problem (8.67) with initial value X_0 has the unique solution

$$\begin{aligned} X_n = {} & \left(\prod_{i=1}^{n} A_i\right) X_0 \\ & + \left(\prod_{i=2}^{n} A_i\right)\phi_1 + \left(\prod_{i=3}^{n} A_i\right)\phi_2 + \cdots + A_n\phi_{n-1} + \phi_n, \end{aligned} \tag{8.68}$$

where

$$\prod_{i=m}^{n} A_i = \begin{cases} A_n A_{n-1} \cdots A_{m+1} A_m & \text{if } m \le n; \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Proof. For $n = 1$, (8.68) reduces to (8.67). The rest of the proof is a straightforward induction. □

Theorem 8.99. Let θ_n be the solution to the homogeneous linear difference equation

$$\theta_{n+s} + \sum_{i=0}^{s-1} \alpha_i \theta_{n+i} = 0 \tag{8.69}$$

with constant coefficients α_i's and the initial values

$$\begin{bmatrix} \theta_0 \\ \theta_{-1} \\ \vdots \\ \theta_{-s+2} \\ \theta_{-s+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{8.70}$$

Then the inhomogeneous equation

$$y_{n+s} + \sum_{i=0}^{s-1} \alpha_i y_{n+i} = \psi_{n+s} \tag{8.71}$$

with the initial values $y_0, y_1, \cdots, y_{s-1}$ is uniquely solved by

$$y_n = \sum_{i=0}^{s-1} \theta_{n-i}\tilde{y}_i + \sum_{i=s}^{n} \theta_{n-i}\psi_i \tag{8.72}$$

where

$$\begin{bmatrix} \tilde{y}_{s-1} \\ \tilde{y}_{s-2} \\ \tilde{y}_{s-3} \\ \vdots \\ \tilde{y}_1 \\ \tilde{y}_0 \end{bmatrix} = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_{s-2} & \theta_{s-1} \\ 0 & 1 & \theta_1 & \cdots & \theta_{s-3} & \theta_{s-2} \\ 0 & 0 & 1 & \cdots & \theta_{s-4} & \theta_{s-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \theta_1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} y_{s-1} \\ y_{s-2} \\ y_{s-3} \\ \vdots \\ y_1 \\ y_0 \end{bmatrix}. \tag{8.73}$$

Exercise 8.100. Prove Theorem 8.99.

8.6.5 Convergence

Definition 8.101. Given initial values

$$\forall i = 0, 1, \ldots, s-1, \quad \mathbf{U}^i = \phi^i(\mathbf{u}(0), k)$$

satisfying

$$\forall i = 0, 1, \ldots, s-1, \quad \lim_{k\to 0} \|\phi^i(\mathbf{u}(0), k) - \mathbf{u}(0)\| = 0, \tag{8.74}$$

an LMM is said to be *convergent* if it yields

$$\lim_{\substack{k\to 0 \\ Nk=T}} \mathbf{U}^N = \mathbf{u}(T) \tag{8.75}$$

for *any* fixed $T > 0$ and *any* IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t.

Lemma 8.102. A convergent LMM is zero-stable.

Proof. Suppose the LMM is not zero-stable. Then there is an unbounded sequence η that satisfies the linear difference equation

$$\sum_{i=0}^{s} \alpha_i \eta_{n+i} = 0.$$

Define another sequence ζ by

$$\zeta_n := \max_{i=0}^{n} |\eta_i|$$

so that ζ converges monotonically to ∞. Consider the IVP

$$u'(t) = 0, \quad u(0) = 0$$

with $T = 1$. For n steps, $k = \frac{1}{n}$. The initial values

$$\forall i = 0, 1, \ldots, s-1, \quad U^i = \eta_i / \zeta_n$$

clearly satisfy (8.74). By definition of the sequence ζ, the computed solution $U^n = \eta_n/\zeta_n$. Because the sequence ζ_n is unbounded, there exist an infinite number of values of n for which $|\zeta_n|$ is greater than the greatest magnitude among previous members of this sequence. These values of n satisfy $|\eta_n/\zeta_n| = 1$ and thus the sequence $n \mapsto \eta_n/\zeta_n$ cannot converge to 0. □

Lemma 8.103. A convergent LMM is preconsistent.

Proof. By (8.75) and the continuity of $\mathbf{u}$ in time, we have

$$\lim_{k\to 0} U^N = \lim_{k\to 0} U^{N-1} = \cdots = \lim_{k\to 0} U^{N-s} = \mathbf{u}(T),$$

where $N = T/k$. Substituting this equation into the limit of the LMM equation (8.45) yields preconsistency as in Definition 8.80. □

Lemma 8.104. A convergent LMM is consistent.

Exercise 8.105. Prove Lemma 8.104.

Lemma 8.106. For an autonomous IVP, the one-step error of a consistent LMM satisfies

$$\|\mathcal{L}\mathbf{u}(t_n)\| \le \sum_{j=0}^{s-1} \Big(\frac{1}{2}(s-j)^2|\alpha_j| + (s-j)|\beta_j|\Big) LMk^2, \tag{8.76}$$

where L is the Lipschitz constant, and M is an upper bound of $\|\mathbf{f}(\mathbf{u}(t))\|$ on $t \in [0, T]$.

Proof. By definition of the one-step error (8.22), we have

$$\begin{aligned}
\mathcal{L}\mathbf{u}(t_n) =& \sum_{j=0}^{s} \alpha_j \mathbf{u}(t_{n+j}) - k\sum_{j=0}^{s} \beta_j \mathbf{u}'(t_{n+j}) \\
=& \sum_{j=0}^{s-1} \alpha_j \mathbf{u}(t_{n+j}) - \sum_{j=0}^{s-1} \alpha_j \mathbf{u}(t_{n+s}) \\
& - k\sum_{j=0}^{s-1} \Big((j-s)\alpha_j - \beta_j\Big)\mathbf{u}'(t_{n+s}) - k\sum_{j=0}^{s-1} \beta_j \mathbf{u}'(t_{n+j}) \\
=& \sum_{j=0}^{s-1} \alpha_j \Big(\mathbf{u}(t_{n+j}) - \mathbf{u}(t_{n+s}) - (j-s)k\mathbf{u}'(t_{n+s})\Big) \\
& + k\sum_{j=0}^{s-1} \beta_j \Big(\mathbf{u}'(t_{n+s}) - \mathbf{u}'(t_{n+j})\Big),
\end{aligned}$$

where the second step follows from the consistency condition (8.60), i.e.,

$$\alpha_s = -\textstyle\sum_{j=0}^{s-1} \alpha_j,$$
$$\beta_s = \textstyle\sum_{j=0}^{s} j\alpha_j - \sum_{j=0}^{s-1} \beta_j = \sum_{j=0}^{s-1} \Big((j-s)\alpha_j - \beta_j\Big).$$

Taylor expansions yield the identity

$$\begin{aligned}
&\mathbf{u}(t_{n+j}) - \mathbf{u}(t_{n+s}) - (j-s)k\mathbf{u}'(t_{n+s}) \\
=& k\int_{s-j}^{0} [\mathbf{f}(\mathbf{u}(t_{n+s} - \xi k)) - \mathbf{f}(\mathbf{u}(t_{n+s}))]\, \mathrm{d}\xi,
\end{aligned}$$

which, together with the Lipschitz condition, implies

$$\begin{aligned}
&\|\mathbf{u}(t_{n+j}) - \mathbf{u}(t_{n+s}) - (j-s)k\mathbf{u}'(t_{n+s})\| \\
\le& kL\int_{0}^{s-j} \|\mathbf{u}(t_{n+s} - \xi k) - \mathbf{u}(t_{n+s})\|\, \mathrm{d}\xi \\
\le& \frac{1}{2}(s-j)^2 k^2 LM,
\end{aligned}$$

where the second step follows from the mean value theorem and the condition of M being an upper bound of $\|\mathbf{f}(\mathbf{u}(t))\|$. Similarly, we have

$$\|\mathbf{f}(\mathbf{u}(t_{n+s})) - \mathbf{f}(\mathbf{u}(t_{n+j}))\| \le LM(s-j)k.$$

Take a norm of $\mathcal{L}\mathbf{u}(t_n)$, apply the above two inequalities, and we have (8.76). □

Lemma 8.107. For an autonomous IVP, the solution errors of a consistent LMM with $k < k_0$ and $k_0|\beta_s|L < 1$ satisfy

$$\left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| \le Ck \max_{i=0}^{s-1} \|\mathbf{E}^{n+i}\| + Dk^2, \quad (8.77)$$

where both C and D are positive constants.

Proof. By definitions of the LMM, its one-step errors, and its solution errors, we have

$$\begin{aligned}
&\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i} \\
=&\mathbf{U}^{n+s} - \mathbf{u}(t_{n+s}) + \sum_{i=0}^{s-1} \alpha_i (\mathbf{U}^{n+i} - \mathbf{u}(t_{n+i})) \\
=&k \sum_{i=0}^{s} \beta_i (\mathbf{f}(\mathbf{U}^{n+i}) - \mathbf{f}(\mathbf{u}(t_{n+i}))) - \mathcal{L}\mathbf{u}(t_n),
\end{aligned}$$

which yields

$$\begin{aligned}
&\left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| \\
\le& \|\mathcal{L}\mathbf{u}(t_n)\| + k|\beta_s| \left\|\mathbf{f}(\mathbf{U}^{n+s}) - \mathbf{f}(\mathbf{u}(t_{n+s}))\right\| \\
&+ k \sum_{i=0}^{s-1} |\beta_i| \left\|\mathbf{f}(\mathbf{U}^{n+i}) - \mathbf{f}(\mathbf{u}(t_{n+i}))\right\| \\
\le& \|\mathcal{L}\mathbf{u}(t_n)\| + kL|\beta_s| \left\|\mathbf{E}^{n+s}\right\| + kL \sum_{i=0}^{s-1} |\beta_i| \left\|\mathbf{E}^{n+i}\right\| \\
\le& \|\mathcal{L}\mathbf{u}(t_n)\| + kL|\beta_s| \left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| \\
&+ kL \sum_{i=0}^{s-1} |\alpha_i\beta_s| \left\|\mathbf{E}^{n+i}\right\| + kL \sum_{i=0}^{s-1} |\beta_i| \left\|\mathbf{E}^{n+i}\right\|.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&(1 - kL|\beta_s|) \left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| \\
\le& kL \sum_{i=0}^{s-1} (|\alpha_i\beta_s| + |\beta_i|) \left\|\mathbf{E}^{n+i}\right\| + \|\mathcal{L}\mathbf{u}(t_n)\|.
\end{aligned}$$

For any $k < k_0 < \frac{1}{|\beta_s L|}$, dividing both sides by $(1 - kL|\beta_s|)$ and applying Lemma 8.106 yield (8.77). ☐

Theorem 8.108. An LMM is convergent if and only if it is consistent and zero-stable.

Proof. We only prove the sufficiency since the necessity has been stated in Lemmas 8.102 and 8.104. By Lemma 8.106, we have

$$\mathbf{E}^{n+s} = -\sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i} + \psi_{n+s},$$

where $\|\psi_n\| \le Ck \max_{i=1}^{s} \|\mathbf{E}^{n-i}\| + Dk^2$ for any k sufficiently small. Then the zero-stability of the LMM and Theorem 8.99 imply the existence of bounded constants θ_i's such that

$$\mathbf{E}^n = \sum_{i=0}^{k-1} \theta_{n-i} \widetilde{\mathbf{E}^i} + \sum_{i=k}^{n} \theta_{n-i} \psi_i,$$

where $\widetilde{\mathbf{E}^i}$s are linear combinations of $\mathbf{E}^j$'s for $i, j = 0, 1, \ldots, s-1$; see (8.73). Note that, in order to apply Theorem 8.99, we have shifted $\mathbf{E}^{n+i}$ to $\mathbf{E}^{n+i-s}$. It follows that

$$\|\mathbf{E}^n\| \le \theta_m \sum_{i=0}^{s-1} \left\|\widetilde{\mathbf{E}^i}\right\| + \theta_m Cks \sum_{i=s}^{n-1} \|\mathbf{E}^i\| + \theta_m D(n-s+1)k^2,$$

where $\theta_m = \sup_{i=1}^{n} |\theta_i|$ and the factor s of the second summation is introduced to account for the fact that a local maximum value of $\|\mathbf{E}^{n-i}\|$ may apprear in at most s adjacent terms. Define a sequence (v_i) as

$$\begin{cases}
v_0 = \theta_m \sum_{i=0}^{s-1} \left\|\widetilde{\mathbf{E}^i}\right\|; \\
v_1 = \theta_m Dk^2 + v_0; \\
\quad \cdots \\
v_n = \theta_m Cks \sum_{i=1}^{n-1} v_i + n\theta_m Dk^2 + v_0,
\end{cases}$$

where $\lim_{k\to 0} v_0 = 0$ because Definition 8.101 implies $\lim_{k\to 0} \left\|\widetilde{\mathbf{E}^i}\right\| = 0$ for each $i = 0, 1, \ldots, s-1$. It is straightforward to show that, for $n > 1$,

$$v_n + \frac{Dk}{Cs} = (1 + \theta_m Cks)\left(v_{n-1} + \frac{Dk}{Cs}\right),$$

which implies

$$\begin{aligned}
v_n &= -\tfrac{Dk}{Cs} + (1 + \theta_m Cks)^{n-1}\left(v_1 + \tfrac{Dk}{Cs}\right) \\
&= (1 + \theta_m Cks)^{n-1} v_0 + \left[(1 + \theta_m Cks)^n - 1\right] \tfrac{Dk}{Cs} \\
&< \exp(\theta_m Csnk) v_0 + \left[\exp(\theta_m Csnk) - 1\right] \tfrac{Dk}{Cs}.
\end{aligned}$$

For $n = T/k$, we have $\lim_{k\to 0} v_n = 0$. The proof is completed by the fact of $\|\mathbf{E}^n\| < v_n$ for each n. ☐

Theorem 8.109. Consider an IVP of which $\mathbf{f}(\mathbf{u}, t)$ is p times continuously differentiable with respect to both t and $\mathbf{u}$. For a convergent LMM with consistency of order p and with its initial conditions satisfying

$$\forall i = 0, 1, \ldots, s-1, \qquad \|\mathbf{U}^i - \mathbf{u}(t_i)\| = O(k^p),$$

its numerical solution of the IVP satisfies

$$\|\mathbf{U}^{t/k} - \mathbf{u}(t)\| = O(k^p) \quad (8.78)$$

for all $t \in [0, T]$ and sufficiently small $k > 0$.

Exercise 8.110. Prove Theorem 8.109.

8.6.6 Absolute stability

Definition 8.111. The *stability polynomial* of an LMM is

$$\pi_\kappa(\zeta) := \rho(\zeta) - \kappa\sigma(\zeta) = \sum_{j=0}^{s} (\alpha_j - \kappa\beta_j)\zeta^j. \quad (8.79)$$

Definition 8.112. An LMM is *absolutely stable* for some κ if all solutions $\{\mathbf{U}^n\}$ of

$$\pi_\kappa(\zeta)\mathbf{U}^n = \left[\rho(\zeta) - \kappa\sigma(\zeta)\right]\mathbf{U}^n = \mathbf{0}$$

are bounded as $n \to +\infty$.

Theorem 8.113 (*Root condition* for absolute stability)**.** An LMM is absolutely stable for $\kappa := k\lambda$ if and only if all the zeros of $\pi_\kappa(\zeta)$ satisfy $|\zeta| \le 1$, and any zero with $|\zeta| = 1$ is simple.

Proof. This proof is the same as that of Theorem 8.95. $\square$

Definition 8.114. The *region of absolute stability (RAS)* for an LMM is the set of all $\kappa \in \mathbb{C}$ for which the method is absolutely stable.

Example 8.115. For Euler's method (8.18),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa = \zeta - (1 + \kappa), \tag{8.80}$$

with the single root $\zeta_1 = 1 + \kappa$. Thus the RAS for Euler's method is the disk:

$$\mathcal{R} = \{\kappa : |1 + \kappa| \le 1\}. \tag{8.81}$$

Example 8.116. For backward Euler's method (8.19),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa\zeta = (1 - \kappa)\zeta - 1, \tag{8.82}$$

with root $\zeta_1 = (1 - \kappa)^{-1}$. Thus the RAS for backward Euler's method is:

$$\mathcal{R} = \{\kappa : |(1 - \kappa)^{-1}| \le 1\} = \{\kappa : |1 - \kappa| \ge 1\}. \tag{8.83}$$

Example 8.117. For the trapezoidal method (8.20),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \frac{1}{2}\kappa(\zeta + 1) = \left(1 - \frac{1}{2}\kappa\right)\zeta - \left(1 + \frac{1}{2}\kappa\right). \tag{8.84}$$

Thus the RAS for the trapezoidal method is the left half-plane:

$$\begin{aligned}\mathcal{R} &= \left\{\kappa \in \mathbb{C} : \left|\frac{2 + \kappa}{2 - \kappa}\right| \le 1\right\} \\ &= \{\kappa \in \mathbb{C} : \mathrm{Re}\,\kappa \le 0\}.\end{aligned} \tag{8.85}$$

Example 8.118. For the midpoint method (8.21),

$$\pi_\kappa(\zeta) = \zeta^2 - 2\kappa\zeta - 1. \tag{8.86}$$

$\pi_z(\zeta) = 0$ implies

$$2\kappa = \zeta - \frac{1}{\zeta}.$$

Since $\zeta = ae^{i\theta}$ and $\frac{1}{\zeta} = a^{-1}e^{-i\theta}$, there are always one zero with $|\zeta_1| \le 1$ and another zero with $|\zeta_2| \ge 1$, depending on the sign of κ. The only possibility for both roots to have a modulus no greater than one is $|\zeta_1| = |\zeta_2| = 1 = a$. So the stability region consists only of the open interval from $-i$ to i on the imaginary axis:

$$\mathcal{R} = \{\kappa \in \mathbb{C} : \kappa = i\alpha \text{ with } |\alpha| < 1\}. \tag{8.87}$$

Definition 8.119. The *boundary locus* method finds the RAS of an LMM (ρ, σ) with $\sigma(e^{i\theta}) \ne 0$ by steps as follows.

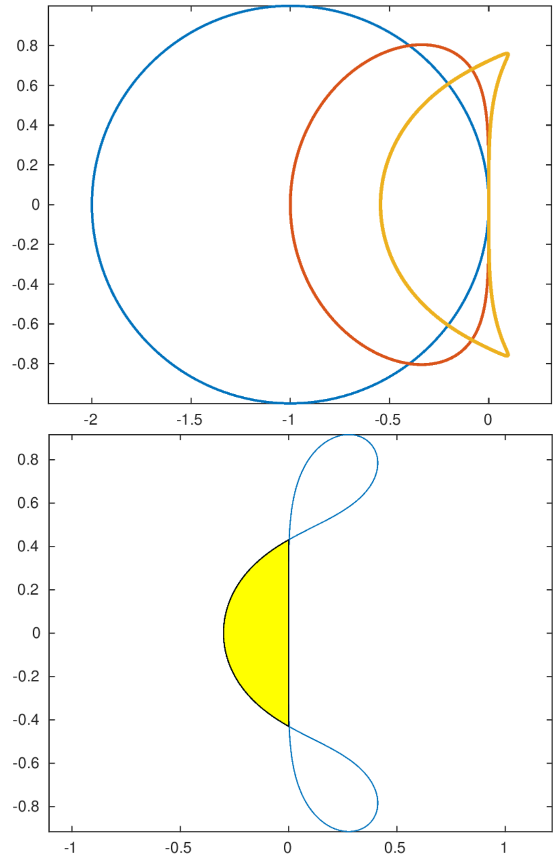
(a) compute the *root locus curve*

$$\gamma(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \qquad \theta = [0, 2\pi]; \tag{8.88}$$

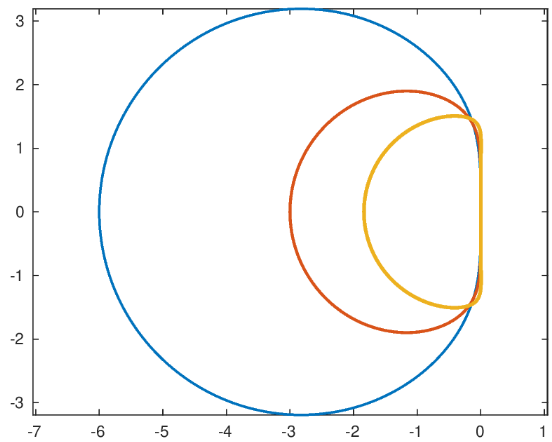
(b) the closed curve γ divides the complex plane $\mathbb{C}$ into a number of connected regions;

(c) for each connected region $S \subset \mathbb{C}$, choose a convenient interior point $\kappa_p \in S$ and solve the equation $\rho(\zeta) - \kappa_p\sigma(\zeta) = 0$: S is part of the RAS if all roots are in the unit disk; otherwise S is not.

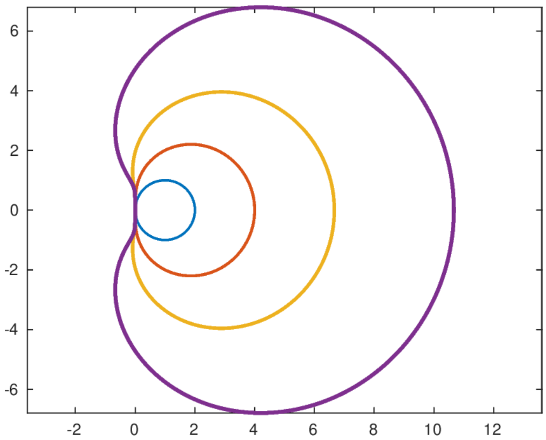
Example 8.120. The RASs of Adams-Bashforth formulas are shown below, with the first plot as those of $p = 1, 2, 3$ and the second as that of $p = 4$. Each RAS is bounded.



Example 8.121. The RASs of Adams-Moulton formulas with $p = 3, 4, 5$ are shown below. Each RAS is bounded.



Example 8.122. The RASs of backward differentiation formulas with $p = 1, 2, 3, 4$ are shown below. Each RAS is unbounded.



Exercise 8.123. Write a program to reproduce the RAS plots in Examples 8.120, 8.121, and 8.122.

8.6.7 The first Dahlquist barrier

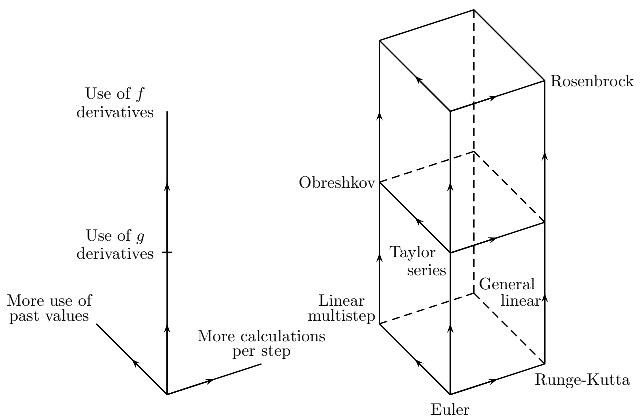
The proofs of conclusions in this subsection can be found in *Hairer et. al. 1993 Solving Ordinary Differential Equations I, Springer 2nd ed.*

Theorem 8.124. The s-step Adams and Nystrom formulas are stable for all $s \geq 1$. The s-step backward differentiation formulas are stable for $s = 1, 2, \ldots, 6$, but unstable for $s \geq 7$.

Theorem 8.125. The order of accuracy p of a stable s-step LMM satisfies

$$p \leq \begin{cases} s & \text{if the LMM is explicit,} \\ s+1 & \text{else if } s \text{ is odd,} \\ s+2 & \text{else if } s \text{ is even.} \end{cases} \tag{8.89}$$

8.7 Runge-Kutta methods



Definition 8.126. A *one-step method* or *multistage method* constructs numerical solutions of a scalar IVP (8.3) at each time step $n = 0, 1, \ldots$ by a formula of the form

$$U^{n+1} = U^n + k\Phi(U^n, t_n; k), \tag{8.90}$$

where the *increment function* $\Phi : \mathbb{R} \times [0, T] \times (0, +\infty) \to \mathbb{R}$ is given in terms of the function $f : \mathbb{R} \times [0, T] \to \mathbb{R}$ in (8.3).

8.7.1 Classical formulas

Definition 8.127. The *modified Euler method* or the *improved polygon method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + \frac{k}{2}y_1, t_n + \frac{k}{2}), \\ U^{n+1} = U^n + ky_2. \end{cases} \tag{8.91}$$

Definition 8.128. The *improved Euler method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + ky_1, t_n + k), \\ U^{n+1} = U^n + \frac{k}{2}(y_1 + y_2). \end{cases} \tag{8.92}$$

Definition 8.129. *Heun's third-order formula* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + \frac{k}{3}y_1, t_n + \frac{k}{3}), \\ y_3 = f(U^n + \frac{2k}{3}y_2, t_n + \frac{2k}{3}), \\ U^{n+1} = U^n + \frac{k}{4}(y_1 + 3y_3). \end{cases} \tag{8.93}$$

Definition 8.130. The *classical fourth-order Runge-Kutta method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + \frac{k}{2}y_1, t_n + \frac{k}{2}), \\ y_3 = f(U^n + \frac{k}{2}y_2, t_n + \frac{k}{2}), \\ y_4 = f(U^n + ky_3, t_n + k), \\ U^{n+1} = U^n + \frac{k}{6}(y_1 + 2y_2 + 2y_3 + y_4). \end{cases} \tag{8.94}$$

Definition 8.131. An *s-stage explicit Runge-Kutta (ERK) method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + ka_{2,1}y_1, t_n + c_2k), \\ y_3 = f(U^n + k(a_{3,1}y_1 + a_{3,2}y_2), t_n + c_3k), \\ \quad \cdots \\ y_s = f(U^n + k(a_{s,1}y_1 + \ldots + a_{s,s-1}y_{s-1}), t_n + c_sk), \\ U^{n+1} = U^n + k(b_1y_1 + b_2y_2 + \ldots + b_sy_s), \end{cases} \tag{8.95}$$

where $a_{i,j}$, b_i, and c_i are real coefficients for $i, j = 1, 2, \ldots, s$, $a_{i,j} = 0$ for $i \leq j$, and

$$\forall i = 1, 2, \ldots, s, \quad c_i = \sum_{j=1}^{s} a_{i,j}. \tag{8.96}$$

Definition 8.132. An *s-stage Runge-Kutta method* is a one-step method of the form

$$\begin{cases} y_i = f(U^n + k\sum_{j=1}^{s} a_{i,j}y_j, t_n + c_ik), \\ U^{n+1} = U^n + k\sum_{j=1}^{s} b_jy_j, \end{cases} \tag{8.97}$$

where $i = 1, 2, \ldots, s$ and the cooefficients $a_{i,j}$, b_j, c_i are real.

Definition 8.133. The *Butcher tableau* is one way to organize the coefficients of a Runge-Kutta method as follows.

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \cdots & a_{1,s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & \cdots & a_{s,s} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{8.98}$$

Definition 8.134. An *implicit Runge-Kutta (IRK) method* is a Runge-Kutta method with at least one $a_{i,j} \neq 0$ for $i \leq j$. A *diagonal implicit Runge-Kutta (DIRK) method* is an IRK method with $a_{i,j} = 0$ whenever $i < j$. A *singly diagonal implicit Runge-Kutta (SDIRK) method* is a DIRK method with $a_{1,1} = a_{2,2} = \cdots = a_{s,s} = \gamma \neq 0$.

Example 8.135. The Butcher tableau of an s-stage ERK method is

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a_{2,1} & 0 & & & \\ c_3 & a_{3,1} & a_{3,2} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \tag{8.99}$$

Example 8.136. The Butcher tableau of the classical fourth-order RK method (8.94), is

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{8.100}$$

Exercise 8.137. Write down the Butcher tableaux of the modified Euler method, the improved Euler method, and Heun's third-order method.

Definition 8.138. The *TR-BDF2 method* is a second-order DIRK method of the form

$$\begin{cases} U^* = U^n + \frac{k}{4}\left(f(U^n, t_n) + f(U^*, t_n + \frac{k}{2})\right), \\ U^{n+1} = \frac{1}{3}\left(4U^* - U^n + kf(U^{n+1}, t_{n+1})\right). \end{cases} \tag{8.101}$$

Exercise 8.139. Rewrite the TR-BDF2 method in the standard form of a Runge-Kutta method and derive its Butcher tableau.

8.7.2 Consistency and convergence

Definition 8.140. The *one-step error of a multistage method* (8.90) is

$$\mathcal{L}u(t_n) := u(t_{n+1}) - u(t_n) - k\Phi(u(t_n), t_n; k). \tag{8.102}$$

Definition 8.141. A multistage method is said to have *order of accuracy* p if

$$\mathcal{L}u(t_n) = \Theta(k^{p+1}) \text{ as } k \to 0. \tag{8.103}$$

Definition 8.142. A multistage method is *consistent* if

$$\lim_{k\to 0} \frac{1}{k}\mathcal{L}u(t_n) = 0. \tag{8.104}$$

Example 8.143. For the modified Euler method, we have

$$\frac{U^{n+1} - U^n}{k} = f\left(U^n + \frac{k}{2}f(U^n, t_n), t_n + \frac{k}{2}\right) \tag{8.105}$$

and thus the one-step error is

$$\begin{aligned} \mathcal{L}u(t_n) =& u(t_{n+1}) - u(t_n) \\ & - kf\left(u(t_n) + \frac{k}{2}f(u(t_n), t_n), t_n + \frac{k}{2}\right) \\ =& u(t_{n+1}) - u(t_n) - kf\left(u(t_n) + \frac{1}{2}ku'(t_n), t_n + \frac{k}{2}\right) \\ =& ku'\left(t_n + \frac{k}{2}\right) + O(k^3) \\ & - kf\left(u\left(t_n + \frac{k}{2}\right) + O(k^2), t_n + \frac{k}{2}\right) \\ =& ku'\left(t_n + \frac{k}{2}\right) + O(k^3) - kf\left(u\left(t_n + \frac{k}{2}\right), t_n + \frac{k}{2}\right) \\ =& O(k^3), \end{aligned}$$

where the second and last equality hold since u satisfies the IVP and the third and fourth follow from Taylor expansions. Hence the method is at least second-order accurate.

Exercise 8.144. Derive the $O(k^3)$ term in Example 8.143 to verify that it does not valish.

Theorem 8.145. A multistage method is consistent if and only if

$$\lim_{k\to 0} \Phi(u, t; k) = f(u, t) \tag{8.106}$$

for any (u, t) in the domain of f.

Proof. Definition 8.126 and a Taylor expansion of $u(t_{n+1})$ at t_n yield

$$\frac{\mathcal{L}u(t_n)}{k} = f(u(t_n), t_n) - \Phi(u(t_n), t_n; k) + \Theta(k).$$

The proof is completed by taking limit of the above equation in the asymptotic range of $k \to 0$, c.f. Definition 8.142. □

Corollary 8.146. The Euler method is consistent.

Proof. This follows from Theorem 8.145 and the fact that $\Phi(u, t; 0) = f(u, t)$ for Euler's method. □

Definition 8.147. A multistage method is *convergent* if its solution error tends to zero as $k \to 0$ for any $T > 0$ and for any initial condition $u_0 = u(0) + o(1)$, i.e.,

$$\lim_{k\to 0; Nk=T} U^N = u(T). \tag{8.107}$$

Lemma 8.148. Let (ξ_n) be a sequence in $\mathbb{R}$ such that

$$|\xi_{n+1}| \leq (1 + C)|\xi_n| + D, \quad n \in \mathbb{N} \tag{8.108}$$

for some positive constants C and D. Then we have

$$|\xi_n| \leq e^{nC}|\xi_0| + \frac{D}{C}(e^{nC} - 1), \quad n \in \mathbb{N}. \tag{8.109}$$

Proof. The induction basis $n = 0$ clearly holds. Now suppose (8.109) holds for n, then for the inductive step, we have

$$\begin{aligned}|\xi_{n+1}| &\le (1+C)e^{nC}|\xi_0| + (1+C)\frac{D}{C}(e^{nC}-1) + D\\ &\le e^{(n+1)C}|\xi_0| + \frac{D}{C}(e^{(n+1)C}-1),\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second from $1 + C \le e^C$. Thus the estimate (8.109) holds for $n + 1$ as well. □

Theorem 8.149. Suppose the increment function Φ that describes a multistage method is continuous (in u, t, and k) and satisfies a Lipschitz condition

$$|\Phi(u,t;k) - \Phi(v,t;k)| \le M|u-v| \tag{8.110}$$

for all (u, t) and (v, t) in the domain of f and for all sufficiently small k. Also suppose that the initial condition satsifies $|E^0| = O(k)$. Then the multistage method is convergent if and only if it is consistent. Furthermore, if the method has order of accuracy p, i.e., $\mathcal{L}u(t_n) \le Kk^{p+1}$, and the initial condition satsifies $|E^0| = O(k^{p+1})$, then its solution error can be bounded as

$$|E^n| \le \frac{K}{M}\left(e^{MT}-1\right)k^p. \tag{8.111}$$

Proof. For sufficiency, we assume that the multistage method is consistent and compute

$$\begin{aligned}|E^{n+1} - E^n| &= |(U^{n+1} - U^n) - (u(t_{n+1}) - u(t_n))\,|\\ &= |k\Phi(U^n, t_n; k) - (u(t_{n+1}) - u(t_n))\,|\\ &= |k\Phi(U^n, t_n; k) - k\Phi(u(t_n), t_n; k) - \mathcal{L}u(t_n)|\\ &\le kM|U^n - u(t_n)| + kc(k),\end{aligned}$$

where the last step follows from the Lipschitz condition (8.110) and $\lim_{k\to 0} c(k) = \lim_{k\to 0} \frac{1}{k}\max|\mathcal{L}u(t)| = 0$. Hence we have

$$|E^{n+1}| \le (1 + kM)|E^n| + kc(k).$$

Applying Lemma 8.148 with $C = kM$ and $D = kc(k)$ yields

$$\begin{aligned}|E^n| &\le |E^0|e^{nkM} + \frac{c(k)}{M}\left(e^{nkM}-1\right)\\ &= |E^0|e^{MT} + \frac{c(k)}{M}\left(e^{MT}-1\right),\end{aligned}$$

which establishes the convergence since $|E^0|$ and $c(k)$ both tend to 0 as $k \to 0$. In particular, (8.111) follows from this inequality and the condition of $c(k) \le Kk^p$.

For necessity, we assume that the multistage method is convergent, i.e., the multistage method (8.90) converges to the solution of

$$u'(t) = f(u,t), \quad u(0) = u_0,$$

for all final time $T > 0$. Consider

$$g(u,t) := \Phi(u,t;0)$$

and observe that by Theorem 8.145 the multistage method is consistent with the new IVP

$$u'(t) = g(u,t), \quad u(0) = u_0.$$

Since we have already shown that consistency implies convergence, the multistage method also converge to this new IVP. Hence the solutions of the two IVPs coincide we have $f(u(\tau), \tau) = g(u(\tau), \tau)$ for all $(u(\tau), \tau)$ in the domain of f. Then the continuity of Φ in k at $k = 0$ implies

$$\begin{aligned}&\forall\epsilon > 0,\ \exists\delta \text{ s.t. } \forall k < \delta,\ \forall t \in [0,T],\\ &\quad |\Phi(u,t;k) - f(u,t)|\\ &\quad \le |\Phi(u,t;0) - f(u,t)| + |\Phi(u,t;k) - \Phi(u,t;0)|\\ &\quad <\epsilon,\end{aligned}$$

which implies uniform convergence of $\Phi(u, t; k)$ to f. Then the proof is completed by Theorem 8.145. □

Corollary 8.150. Both the modified Euler method and the improved Euler method are convergent. If f in the IVP is twice continuously differentiable, then each of them has order of accuracy two.

Proof. For the modified Euler method (8.91), we have

$$\Phi(u,t;k) = f\left(u + \frac{k}{2}f(u,t), t + \frac{k}{2}\right),$$

which clearly satisfies the consistency condition (8.106), and hence by Theorem 8.149, it only remains to verify the Lipschitz condition of Φ. From the Lipschitz condition for f we obtain

$$\begin{aligned}&|\Phi(u,t;k) - \Phi(v,t;k)|\\ &= \left|f\left(u + \frac{k}{2}f(u,t), t + \frac{k}{2}\right) - f\left(v + \frac{k}{2}f(v,t), t + \frac{k}{2}\right)\right|\\ &\le L\left(|u-v| + \frac{k}{2}\,|f(u,t) - f(v,t)|\right)\\ &\le L\left(1 + \frac{kL}{2}\right)|u-v|,\end{aligned}$$

hence Φ also satisfies a Lipschitz condition.

If f is twice continuously differentiable, then by Example 8.143, the one-step error of the modified Euler method satisfies

$$\mathcal{L}u(t_n) \le Kk^3,$$

Therefore the modified Euler method (8.91) has order of accuracy two by Theorem 8.39.

The same result concerning the improved Euler method (8.92) can be proved in a similar manner. □

Lemma 8.151. The one-step error of the classical Runge-Kutta method (8.94) is

$$\mathcal{L}u(t_n) = O(k^5). \tag{8.112}$$

Exercise 8.152. Prove Lemma 8.151.

Corollary 8.153. The classical Runge-Kutta method (8.94) is convergent. If f in the IVP is four-times continuously differentiable, then it is convergent with order of accuracy four.

Proof. The function Φ describing the classical Runge-Kutta method (8.94) is given by

$$\Phi = \frac{1}{6}(\Phi_1 + 2\Phi_2 + 2\Phi_3 + \Phi_4),$$

where

$$\begin{aligned}
\Phi_1(u,t;k) &= f(u,t),\\
\Phi_2(u,t;k) &= f\left(u + \frac{k}{2}\Phi_1(u,t;k), t + \frac{k}{2}\right),\\
\Phi_3(u,t;k) &= f\left(u + \frac{k}{2}\Phi_2(u,t;k), t + \frac{k}{2}\right),\\
\Phi_4(u,t;k) &= f(u + k\Phi_3(u,t;k), t + k).
\end{aligned}$$

From this, consistency follows immediately by Theorem 8.145. Since Φ clearly satisfies a Lipschitz condition, it follows from Theorem 8.149 that the classical Runge-Kutta method (8.94) is convergent.

If f is four-times continuously differentiable, Lemma 8.151 shows that the classical Runge-Kutta method (8.94) has a one-step error of $O(k^5)$, hence it has order of accuracy four by Theorem 8.149. □

8.7.3 Absolute stability

Definition 8.154. The *stability function of a one-step method* is a ratio of two polynomials

$$R(z) = \frac{P(z)}{Q(z)} \tag{8.113}$$

that satisfies

$$U^{n+1} = R(z)U^n \tag{8.114}$$

for the test problem $u'(t) = \lambda u$ where $z := k\lambda$.

Example 8.155. The fourth-order Runge-Kutta method has its stability function as

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4. \tag{8.115}$$

Example 8.156. The trapezoidal rule, when viewed as a one-step method has its stability function as

$$R(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, \tag{8.116}$$

which is also the root of the LMM stability polynomial in Example 8.117.

Exercise 8.157. Show that the TR-BDF2 method (8.101) has

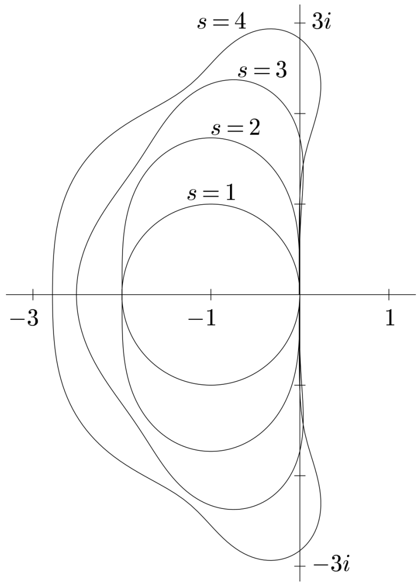
$$R(z) = \frac{1+\frac{5}{12}z}{1-\frac{7}{12}z+\frac{1}{12}z^2}, \tag{8.117}$$

and $R(z) - e^z = O(z^3)$ as $z \to 0$.

Definition 8.158. The *region of absolute stability (RAS) of a one-step method* is a subset of the complex plane

$$\mathcal{R} := \{z \in \mathbb{C} : |R(z)| \le 1\}. \tag{8.118}$$

Example 8.159. The boundaries of RASs for ERKs with $s = 1, 2, 3, 4$ are shown below.



8.8 Stiff IVPs

Example 8.160. Consider the IVP

$$u'(t) = \lambda(u - \cos t) - \sin t, \quad u(0) = \eta. \tag{8.119}$$

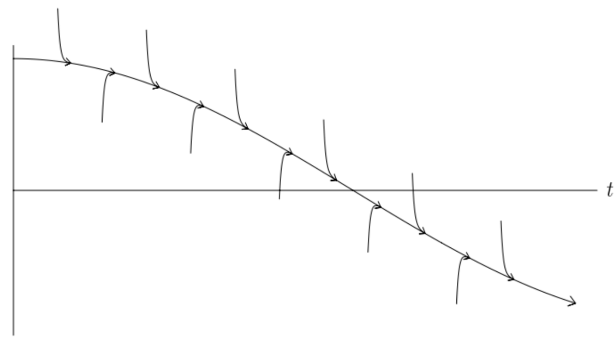
By Duhamel's principle (8.13), the exact solution is

$$\begin{aligned}
u_\eta(t) &= e^{\lambda t}\eta - \int_0^t e^{\lambda(t-\tau)}(\lambda\cos\tau + \sin\tau)\mathrm{d}\tau\\
&= e^{\lambda t}\eta - \int_0^t \lambda e^{\lambda(t-\tau)}\cos\tau\mathrm{d}\tau - \int_0^t e^{\lambda(t-\tau)}\sin\tau\mathrm{d}\tau\\
&= e^{\lambda t}(\eta - 1) + \cos t,
\end{aligned}$$

where the third equality follows from the integration-by-parts formula.

If $\eta = \cos(0) = 1$, then $u_1(t) = \cos t$ is the unique solution. If $\eta \ne 1$ and $\lambda < 0$, then the solution curve $u_\eta(t)$ decays exponentially to $u_1(t)$.

A negative λ with large magnitude has a dominant effect on nearby solutions of the ODE corresponding to different initial data; the following picture shows some solution curves with $\lambda = -100$.



For six values of k, the following table compares the results at $T = 1$ computed by the second-order Adams-Bashforth and the second-order BDF method.

k	AB2	BDF2
0.2	14.40	0.5404
0.1	-5.70×10^4	0.54033
0.05	-1.91×10^9	0.540309
0.02	-5.77×10^{10}	0.5403034
0.01	0.5403019	0.54030258
0.005	0.54030222	0.54030238
$\vdots$	$\vdots$	$\vdots$
0	0.540302306	0.540302306

The results indicate the curious effect that this property of the ODE has on numerical computations. To achieve a solution error $E(T) \le \epsilon = 4 \times 10^{-5}$, the BDF2 method may use $k = 0.1$, the AB2 method has to use $k \le 0.01$ while the time scale of the IVP is 1.

8.8.1 The notion of stiffness

Definition 8.161. An IVP is said to be *stiff in an interval* if for some initial condition any numerical method with a finite RAS is forced to use a time-step size that is excessively smaller than the time scale of the true solution of the IVP.

Formula 8.162. A general way of reducing an IVP

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}, t)$$

to a collection of scalar, linear model problems of the form

$$w_i'(t) = \lambda_i w_i(t), \quad i = 1, 2, \cdots, n$$

consists of steps as follows.

(a) Linearization: at the neighborhood of a particular solution $\mathbf{u}^*(t)$, we write

$$\mathbf{u}(t) = \mathbf{u}^*(t) + (\delta\mathbf{u})(t)$$

and apply Taylor expansion

$$\mathbf{f}(\mathbf{u}, t) = \mathbf{f}(\mathbf{u}^*, t) + J(t)\|\delta\mathbf{u}\| + o(\|\delta\mathbf{u}\|)$$

to obtain

$$(\delta\mathbf{u})'(t) = J(t)(\delta\mathbf{u}).$$

(b) Freezing coefficients: set

$$A = J(t^*),$$

where t^* is the particular time of interest.

(c) Diagonalization: assume A is diagonalizable by V and we write

$$(\delta\mathbf{u})'(t) = V(V^{-1}AV)V^{-1}(\delta\mathbf{u}).$$

Define $\mathbf{w} := V^{-1}(\delta\mathbf{u})$ and we have a collection of decoupled scalar IVPs,

$$\mathbf{w}'(t) = \Lambda\mathbf{w}(t),$$

where $\Lambda = V^{-1}AV$ is the diagonal matrix.

Definition 8.163. For an IVP

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{b}(t) \tag{8.120}$$

where $\mathbf{u}, \mathbf{f} \in \mathbb{R}^n$ and A is a constant, diagonalizable, $n \times n$ matrix with eigenvalues $\lambda_i \in \mathbb{C}, i = 1, 2, \cdots, n$, its *stiffness ratio* is

$$\frac{\max_{\lambda \in \Lambda(A)} |\mathrm{Re}\,\lambda|}{\min_{\lambda \in \Lambda(A)} |\mathrm{Re}\,\lambda|}. \tag{8.121}$$

Example 8.164. Consider the linear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -1000 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad t \in [0, 1] \tag{8.122}$$

with initial value $\mathbf{u}(0) = (1, 1)^T$. Suppose we want

$$\|\mathbf{E}\|_\infty \le \epsilon,$$

that is

$$|U_1^N - e^{-1000}| \le \epsilon, \quad |U_2^N - e^{-1}| \le \epsilon.$$

If (8.122) is solved by a p-th order LMM with time step k. To obtain U_2^N sufficiently accurately, we need $k = O(\epsilon^{1/p})$. But to obtain U_1^N sufficiently accurately, if the formula has a stability region of finite size like the Euler formula, we need k to be on the order 10^{-3}. Most likely this is a much tighter restriction.

Example 8.165. Consider the nonlinear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -u_1 u_2 \\ \cos(u_1) - \exp(u_2) \end{pmatrix}. \tag{8.123}$$

The Jacobian matrix is

$$J = -\begin{pmatrix} u_2 & u_1 \\ \sin(u_1) & \exp(u_2) \end{pmatrix}.$$

Near a point t with $u_1(t) = 0$ and $u_2(t) \gg 1$, the matrix is diagonal with widely differing eigenvalues and the behavior will probably be stiff.

Example 8.166. Read Example 8.2 (pp 167) in the book by Leveque.

8.8.2 A-stability and L-stability

Definition 8.167. An ODE method is *A-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \mathrm{Re}\, z \le 0\} \subseteq \mathcal{R}. \tag{8.124}$$

Example 8.168. The backward Euler's method and trapezoidal method are A-stable.

Theorem 8.169 (Dahlquist's Second Barrier)**.** The order of accuracy of an implicit A-stable LMM satisfies $p \le 2$. An explicit LMM cannot be A-stable.

Definition 8.170. An ODE method is *$A(\alpha)$-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \pi - \alpha \le \arg(z) \le \pi + \alpha\} \subseteq \mathcal{R}. \tag{8.125}$$

It is *A(0)-stable* if it is A(α)-stable for some $\alpha > 0$.

Example 8.171. As shown in Example 8.122, the BDFs are A(α)-stable with $\alpha = 90°$ for $p = 1, 2$ and $\alpha \approx 86°$, $73°$, $51°$, and $17°$ for $p = 3, 4, 5, 6$ respectively. Note the large drop of α from $p = 5$ to $p = 6$.

Definition 8.172. A one-step method is *L-stable* if it is A-stable and

$$\lim_{z\to\infty} |R(z)| = 0, \tag{8.126}$$

where $U^{n+1} = R(z)U^n$.

Example 8.173. We use the trapezoidal and backward Euler's methods to solve the IVP (8.119) with $\lambda = -10^6$. The following table shows the errors at $T = 3$ with various values of k and the initial data $u(0) = \eta$.

	k	Backward Euler	Trapezoidal
	0.4	4.7770e-02	4.7770e-02
$\eta = 1$	0.2	9.7731e-08	4.7229e-10
	0.1	4.9223e-08	1.1772e-10
	0.4	4.7770e-02	4.5219e-01
$\eta = 1.5$	0.2	9.7731e-08	4.9985e-01
	0.1	4.9223e-08	4.9940e-01

The results are caused by the fact that the backward Euler's method is L-stable while the trapezoidal method is not.

Exercise 8.174. Reproduce the results in Example 8.173.

Chapter 9

Boundary Value Problems

Chapter 10

Parabolic Problems

10.1 Parabolic equations

Definition 10.1. A second-order, constant-coefficient, linear partial differential equation (PDE) of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \tag{10.1}$$

is called a *parabolic PDE* if its coefficients satisfy

$$B^2 - 4AC = 0. \tag{10.2}$$

Definition 10.2. The *one-dimensional heat equation* is a parabolic PDE of the form

$$u_t = \nu u_{xx} \text{ in } \Omega := (0,1) \times (0,T), \tag{10.3}$$

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*

$$u(x,0) = \eta(x), \text{ on } (0,1) \times \{0\} \tag{10.4}$$

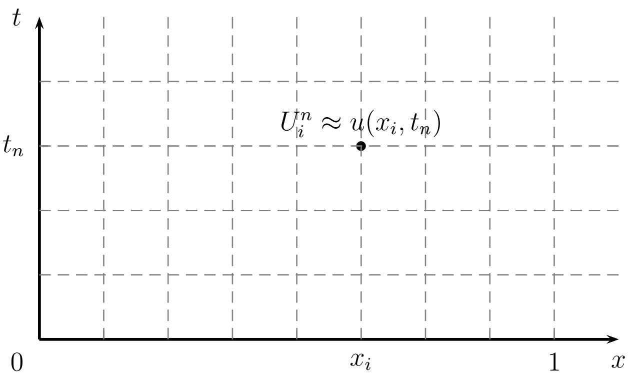
and appropriate boundary conditions at $\{0,1\} \times (0,T)$.

10.2 Method of lines (MOL)

Notation 11. The space-time domain of the PDE (10.3) can be discretized by the rectangular grids

$$x_i = ih, \quad t_n = nk, \tag{10.5}$$

$h = \frac{1}{m+1}$ 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).



Definition 10.3. The *method of lines* (MOL) is a technique for solving PDEs via

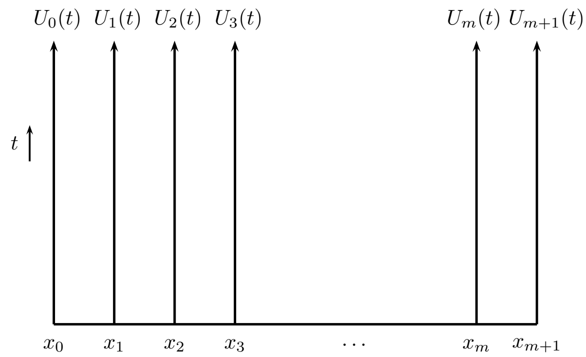
(a) discretizing the spatial derivatives while leaving the time variable continuous;

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

Example 10.4. Discretize the heat equation (10.3) in space at grid point x_i by

$$U_i'(t) = \frac{\nu}{h^2}\Big(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)\Big), \tag{10.6}$$

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



For Dirichlet conditions

$$\begin{cases} u(0,t) = g_0(t), & \text{on } \{0\} \times (0,T); \\ u(1,t) = g_1(t), & \text{on } \{1\} \times (0,T), \end{cases} \tag{10.7}$$

this semi-discrete system (10.6) can be written as

$$\mathbf{U}'(t) = A\mathbf{U}(t) + g(t), \tag{10.8}$$

where

$$A = \frac{\nu}{h^2}\begin{bmatrix} -2 & +1 & & & & \\ +1 & -2 & +1 & & & \\ & +1 & -2 & +1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & +1 & -2 & +1 \\ & & & & +1 & -2 \end{bmatrix}, \tag{10.9}$$

$$\mathbf{U}(t) := \begin{bmatrix} U_1(t) \\ U_2(t) \\ U_3(t) \\ \vdots \\ U_{m-1}(t) \\ U_m(t) \end{bmatrix}, \quad g(t) = \frac{\nu}{h^2} \begin{bmatrix} g_0(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}. \tag{10.10}$$

Definition 10.5. The *FTCS (forward in time, centered in space) method* solves the heat equation (10.3) by

$$\frac{U_i^{n+1} - U_i^n}{k} = \frac{\nu}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n), \tag{10.11}$$

or, equivalently

$$U_i^{n+1} = U_i^n + 2r(U_{i-1}^n - 2U_i^n + U_{i+1}^n), \tag{10.12}$$

where $r := \frac{k\nu}{2h^2}$.

Example 10.6. For homogeneous Dirichlet boundary conditions, the FTCS method can be written as

$$\mathbf{U}^{n+1} = (I + kA)\mathbf{U}^n, \tag{10.13}$$

where A is the matrix in (10.9) and

$$\mathbf{U}^n := \begin{bmatrix} U_1^n \\ U_2^n \\ \vdots \\ U_m^n \end{bmatrix}. \tag{10.14}$$

Definition 10.7. The *Crank-Nicolson method* solves the heat equation (10.3) by

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{k} &= \frac{1}{2}\Big(f(U^n, t_n) + f(U^{n+1}, t_{n+1})\Big) \\ &= \frac{\nu}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}), \end{aligned} \tag{10.15}$$

or, equivalently

$$\begin{aligned} &-rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} \\ =&rU_{i-1}^n + (1-2r)U_i^n + rU_{i+1}^n. \end{aligned} \tag{10.16}$$

Exercise 10.8. Show that the matrix form of the Crank-Nicolson method for solving the heat equation (10.3) with Dirichlet conditions is

$$\left(I - \frac{k}{2}A\right)\mathbf{U}^{n+1} = \left(I + \frac{k}{2}A\right)\mathbf{U}^n + \mathbf{b}^n, \tag{10.17}$$

where $r = \frac{k\nu}{2h^2}$ and

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

10.3 Accuracy and Consistency

Definition 10.9. The *local truncation error (LTE) of an MOL* for solving a PDE is the error caused by replacing continuous derivatives with finite difference formulas.

Example 10.10. The LTE of the FTCS method in Definition 10.5 is

$$\begin{aligned} \tau(x,t) =& \frac{u(x,t+k) - u(x,t)}{k} \\ &- \frac{\nu}{h^2}\Big(u(x-h,t) - 2u(x,t) + u(x+h,t)\Big) \\ =&\Big(u_t + \frac{1}{2}ku_{tt} + \frac{1}{6}k^2u_{ttt} + \cdots\Big) \\ &- \nu\Big(u_{xx} + \frac{1}{12}h^2u_{xxxx} + \cdots\Big) \\ =&\Big(\frac{1}{2}k - \frac{\nu}{12}h^2\Big)u_{xxxx} + O(k^2 + h^4), \end{aligned}$$

where the first step follows from the Definition 8.31, the second from Taylor expansions and the last from $u_t = \nu u_{xx}$ and $u_{tt} = \nu u_{xxt} = \nu u_{txx} = \nu u_{xxxx}$. Due to $\tau(x,t) = O(k + h^2)$, this method is said to be second order accurate in space and first order accurate in time.

Exercise 10.11. Show that the Crank-Nicolson method in Definition 10.7 is second order accurate in both space and time by calculating the LTE as

$$\tau(x,t) = O(k^2 + h^2).$$

Definition 10.12. An MOL is said to be *consistent* if

$$\lim_{k,h\to 0} \tau(x,t) = 0. \tag{10.18}$$

Definition 10.13. The *solution error* of an MOL is

$$E_i^n = U_i^n - u(x_i, t_n), \tag{10.19}$$

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

10.4 Stability

Lemma 10.14. The eigenvalues λ_p and eigenvectors $\mathbf{w}^p$ of A in (10.9) are

$$\lambda_p = -\frac{4\nu}{h^2}\sin^2\Big(\frac{p\pi h}{2}\Big), \tag{10.20}$$

$$w_j^p = \sin(p\pi jh), \tag{10.21}$$

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

Example 10.15. For the FTCS method (10.11) to be absolutely stable, we must have $|1 + k\lambda| \le 1$ for each eigenvalue in (10.20), which implies $-2 \le -4\nu k/h^2 \le 0$ and thus limits the time-step size to

$$k \le \frac{h^2}{2\nu}. \tag{10.22}$$

Definition 10.16. An MOL is said to be *unconditionally stable* for a PDE if in solving the semi-discrete system of the PDE its ODE solver is absolutely stable for any $k > 0$.

Lemma 10.17. 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 $k > 0$. □

Corollary 10.18. The Crank-Nicolson method (10.16) is unconditionally stable for the heat equation.

Proof. The ODE solver of the Crank-Nicolson method (10.16) is the trapezoidal rule, which is A-stable and hence $A(\alpha)$-stable. The proof is completed by Lemma 10.17. □

Definition 10.19. A linear MOL of the form

$$\mathbf{U}^{n+1} = B(k)\mathbf{U}^n + b^n(k). \tag{10.23}$$

is *Lax-Richtmyer stable* if

$$\forall T > 0,\ \exists C_T > 0, \quad \forall k > 0, \forall n \in \mathbb{N}^+ \text{ satisfying } nk \le T, \\ \|B(k)^n\| \le C_T. \tag{10.24}$$

Definition 10.20. A linear MOL (10.23) is said to have *strong stability* if

$$\|B\|_2 \le 1. \tag{10.25}$$

Corollary 10.21. The Crank-Nicolson method has strong stability with

$$B = \Big(I - \frac{k}{2}A\Big)^{-1}\Big(I + \frac{k}{2}A\Big). \tag{10.26}$$

Proof. (10.26) follows directly from Exercise 10.8. The symmetry of A implies the symmetry of B and thus the spectral radius of B satisfies

$$\rho(B) = \frac{1 + k\lambda_p/2}{1 - k\lambda_p/2} \le 1.$$

Then the proof is completed by Definition 10.20. □

10.5 Convergence

Theorem 10.22 (Lax Equivalence Theorem)**.** A consistent linear MOL (10.23) is convergent if and only if it is Lax-Richtmyer stable.

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

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

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

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

Subtracting (*) from (10.23) gives the difference equation for the global error $E^n = U^n - \hat{U}^n$:

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

and hence, by induction,

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

from which we obtain

$$\|E^N\| \le \|B^N\|\|E^0\| + k\sum_{n=1}^{N} \|B^{N-n}\|\|\tau^{n-1}\|.$$

If the method is Lax-Richtmyer stable, then for $Nk < T$, we have

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

the RHS goes to 0 as $k \to 0$. □

Corollary 10.23. The Crank-Nicolson method is convergent for any $k > 0$.

Proof. This follows from Theorem 10.22 and Corollary 10.21. □

Example 10.24. For the FTCS method, (10.13) implies

$$B = I + kA \tag{10.27}$$

and thus the convergence depends on

$$\rho(B) \le 1 + O(k),$$

which is a form of Lax-Richtmyer stability.

Exercise 10.25. Prove the necessity part of Theorem 10.22.

10.6 Von Neumann analysis

Theorem 10.26. The exact solution to the heat equation (10.3) with Dirichlet conditions $g_0(t) = g_1(t) = 0$ is

$$u(x,t) = \sum_{j=0}^{\infty} \hat{u}_j(t)\sin(\pi j x), \tag{10.28}$$

where

$$\hat{u}_j(t) = \exp(-j^2\pi^2\nu t)\hat{u}_j(0), \tag{10.29}$$

and $\hat{u}_j(0)$ is determined as the Fourier coefficients of the initial data $\eta(x)$.

Proof. It is straightforward to verify that (10.28) is indeed the solution of (10.3). □

Example 10.27. Consider the FTCS method. To apply von Neumann analysis we consider how this method works on a single wave number ξ, *i.e.*, we set

$$U_j^n = [g(\xi)]^n e^{ix_j\xi}. \tag{10.30}$$

Then we expect that

$$U_j^{n+1} = g(\xi)U_j^n, \tag{10.31}$$

where $g(\xi)$ is the amplification factor for this wave number. Inserting these expressions into (10.12) gives

$$g(\xi)U_j^n = \left[1 + \frac{\nu k}{h^2}\left(e^{-i\xi h} - 2 + e^{i\xi h}\right)\right]U_j^n,$$

i.e.,

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

To guarantee $|g(\xi)| \le 1$, we take

$$1 - \frac{4\nu k}{h^2} \ge -1,$$

which implies (10.22), i.e. $k \le \frac{h^2}{2\nu}$.

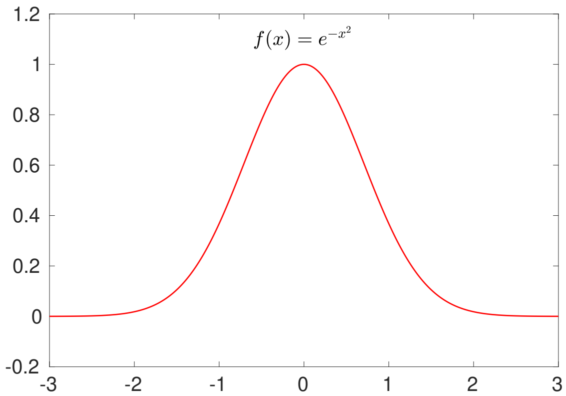
Exercise 10.28. For the Crank-Nicolson method, show that the modulus of its amplification factor is never greater than 1 for any choice of $k, h > 0$.

10.7 Green's function of the heat equation in $(-\infty, +\infty)$

Definition 10.29. A Gaussian function, often simply referred to as a *Gaussian*, is a function of the form

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}}, \tag{10.32}$$

for arbitrary real constants a, b and non-zero c.



Lemma 10.30.

$$\int_{-\infty}^{+\infty} ae^{-\frac{(x-b)^2}{2c^2}}\,dx = ac\sqrt{2\pi}. \tag{10.33}$$

Proof. By the trick of combining two one-dimensional Gaussians and the Polar coordinate transformation, we have

$$\begin{aligned}
\int_{-\infty}^{+\infty} e^{-x^2}\,dx &= \sqrt{\left(\int_{-\infty}^{+\infty} e^{-x^2}dx\right)\left(\int_{-\infty}^{+\infty} e^{-y^2}dy\right)} \\
&= \sqrt{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)}\,dxdy} \\
&= \sqrt{\int_0^{2\pi}\int_0^{+\infty} e^{-r^2}r\,drd\theta} \\
&= \sqrt{2\pi\cdot -\frac{1}{2}e^{-r^2}\Big|_0^{+\infty}} \\
&= \sqrt{\pi}.
\end{aligned}$$

and hence

$$\int_{-\infty}^{+\infty} ae^{-\frac{(x-b)^2}{2c^2}}\,dx = \sqrt{2}ac\int_{-\infty}^{+\infty} e^{-y^2}dy = ac\sqrt{2\pi},$$

where it follows from the transformation of $x = b+\sqrt{2}cy$. $\square$

Lemma 10.31. The Fourier transform of a Gaussian centered at the origin is another such Gaussian.

Proof. First we consider the case $f(x) = e^{-x^2}$, then

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}e^{-i\xi x}\,dx,$$

Differentiating with respect to ξ yields

$$\begin{aligned}
\frac{d}{d\xi}\hat{f}(\xi) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}(-ix)e^{-i\xi x}\,dx \\
&= \frac{i}{2\sqrt{2\pi}}\int_{-\infty}^{+\infty} \frac{de^{-x^2}}{dx}e^{-i\xi x}\,dx \\
&= -\frac{\xi}{2\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}e^{-i\xi x}\,dx \\
&= -\frac{\xi}{2}\hat{f}(\xi),
\end{aligned}$$

where the third line follows from the integration by parts formula. The unique solution to this ordinary differential equation is given by

$$\hat{f}(\xi) = c\cdot e^{-\frac{\xi^2}{4}},$$

where $c = \hat{f}(0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}\,dx = \frac{\sqrt{2}}{2}$. The proof is completed by the dilation property (E.10) of Fourier transform. In particular, the Fourier transform of a Gaussian with $a = 1$, $b = 0$, and c is another Gaussian with $a' = c$, $b' = 0$, and $c' = \frac{1}{c}$. $\square$

Lemma 10.32. For any $u \in L^2$ satisfying

$$\forall n \in \mathbb{N}, \qquad \lim_{x\to\pm\infty} u^{(n)}(x) = 0, \tag{10.34}$$

we have

$$\widehat{\frac{\partial^2 u}{\partial x^2}} = -\xi^2\hat{u}. \tag{10.35}$$

Proof. Repeated application of (10.34) yields

$$\begin{aligned}
\sqrt{2\pi}\cdot\widehat{\frac{\partial^2 u}{\partial x^2}} &= \int_{-\infty}^{+\infty} e^{-i\xi x}\frac{\partial^2 u}{\partial x^2}\mathrm{d}x \\
&= e^{-i\xi x}\frac{\mathrm{d}u}{\mathrm{d}x}\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty}\frac{\mathrm{d}u}{\mathrm{d}x}(-i\xi)e^{-i\xi x}\mathrm{d}x \\
&= i\xi\int_{-\infty}^{+\infty}\frac{\mathrm{d}u}{\mathrm{d}x}e^{-i\xi x}\mathrm{d}x \\
&= i\xi(e^{-i\xi x}u)\Big|_{-\infty}^{+\infty} + (i\xi)^2\int_{-\infty}^{+\infty} ue^{-i\xi x}\mathrm{d}x \\
&= -\xi^2\int_{-\infty}^{+\infty} ue^{-i\xi x}\mathrm{d}x = -\xi^2\sqrt{2\pi}\hat{u},
\end{aligned}$$

where the first and last lines follow from Definition E.2, the second and fourth lines from the integration by parts formula, and the third line from (10.34). □

Theorem 10.33. The solution to the heat equation

$$u_t = \nu u_{xx} \text{ on } (-\infty,+\infty) \tag{10.36}$$

with the initial condition $\eta(x) = e^{-\beta x^2}$ is

$$u(x,t) = \frac{1}{\sqrt{4\beta\nu t+1}}e^{-\frac{x^2}{4\nu t+1/\beta}}. \tag{10.37}$$

Proof. By Lemma 10.32, the Fourier transform of (10.36) leads to the ODE

$$\hat{u}_t(\xi,t) = -\nu\xi^2\hat{u}(\xi,t),$$

the solution of which with the initial data $\hat{u}(\xi,0)=\hat{\eta}(\xi)$ yields

$$\hat{u}(\xi,t) = e^{-\nu\xi^2 t}\hat{\eta}(\xi).$$

Then

$$\begin{aligned}
u(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{u}(\xi,t)e^{i\xi x}\mathrm{d}\xi \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\nu\xi^2 t}\hat{\eta}(\xi)e^{i\xi x}\mathrm{d}\xi \\
&= \frac{1}{2\sqrt{\pi\beta}}\int_{-\infty}^{+\infty} e^{-\xi^2(\nu t+\frac{1}{4\beta})}e^{i\xi x}\mathrm{d}\xi.
\end{aligned}$$

Define $C = \frac{1}{4\nu t+1/\beta}$, then

$$\begin{aligned}
u(x,t) &= \frac{1}{2\sqrt{\pi\beta}}\int_{-\infty}^{+\infty} e^{\frac{\xi^2}{4C}}e^{i\xi x}\mathrm{d}\xi \\
&= \frac{1}{2\sqrt{\pi\beta}}\sqrt{4\pi C}\cdot e^{-x^2 C} \\
&= \frac{1}{\sqrt{4\beta\nu t+1}}e^{-\frac{x^2}{4\nu t+1/\beta}}.
\end{aligned}$$

As t increases this Gaussian becomes more spread out and the magnitude decreases. □

Corollary 10.34. A translation of the initial condition

$$\eta(x) = e^{-\beta(x-\bar{x})^2} \tag{10.38}$$

of the heat equation (10.36) leads to a translation of the solution, i.e.,

$$u(x,t) = \frac{1}{\sqrt{4\beta\nu t+1}}e^{-\frac{(x-\bar{x})^2}{4\nu t+1/\beta}}. \tag{10.39}$$

Corollary 10.35. For the heat equation (10.36) with the initial condition as

$$\omega_\beta(x,0;\bar{x}) = \sqrt{\frac{\beta}{\pi}}e^{-\beta(x-\bar{x})^2}, \tag{10.40}$$

its solution is

$$\omega_\beta(x,t;\bar{x}) = \frac{1}{\sqrt{4\pi\nu t+\pi/\beta}}e^{-\frac{(x-\bar{x})^2}{4\nu t+1/\beta}}. \tag{10.41}$$

Definition 10.36. The *Green's function*

$$G(x,t;\bar{x}) := \lim_{\beta\to+\infty}\omega_\beta(x,t;\bar{t}) = \frac{1}{\sqrt{4\pi\nu t}}e^{-\frac{(x-\bar{x})^2}{4\nu}} \tag{10.42}$$

is the solution of the heat equation (10.36) with its initial condition as the delta function

$$\delta(x-\bar{x}) := \lim_{\beta\to+\infty}\omega_\beta(x,0;\bar{x}). \tag{10.43}$$

Chapter 11

Hyperbolic Problems

Definition 11.1. A second-order, constant-coefficient, linear partial differential equation (PDE) of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \tag{11.1}$$

is called a *hyperbolic PDE* if its coefficients satisfy

$$B^2 - 4AC > 0. \tag{11.2}$$

Definition 11.2. The *one-dimensional wave equation* is a hyperbolic PDE of the form

$$u_{tt} = a^2 u_{xx}, \tag{11.3}$$

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

Definition 11.3. The *one-dimensional advection equation* is

$$u_t = -au_x \text{ in } \Omega := (0,1) \times (0,T), \tag{11.4}$$

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*

$$u(x,0) = \eta(x), \text{ on } (0,1) \times \{0\} \tag{11.5}$$

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

Theorem 11.4. The exact solution of the Cauchy problem (11.4) is

$$u(x,t) = \eta(x - at). \tag{11.6}$$

Proof. It is straightforward to verify that

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

□

Definition 11.5. A system of PDEs of the form

$$\mathbf{u}_t + A\mathbf{u}_x = \mathbf{0} \tag{11.7}$$

is *hyperbolic* if A is diagonalizable and its eigenvalues are all real.

Example 11.6. The Euler equations are

$$\frac{\partial}{\partial t}\begin{bmatrix} p \\ u \end{bmatrix} + \begin{bmatrix} 0 & \kappa_0 \\ \frac{1}{\rho_0} & 0 \end{bmatrix} \frac{\partial}{\partial x}\begin{bmatrix} p \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{11.8}$$

The equation for the pressure p can be further written as

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

11.1 Classical MOLs

Example 11.7. Discretize the advection equation (11.4) in space at grid point x_j by

$$U_j'(t) = -\frac{a}{2h}\left(U_{j+1}(t) - U_{j-1}(t)\right), \quad 2 \le j \le m, \tag{11.9}$$

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

$$u(0,t) = u(0,t) = g_0(t), \tag{11.10}$$

the discretizations of (11.4) at $j = 1$ and $j = m+1$ are

$$U_1'(t) = -\frac{a}{2h}\left(U_2(t) - U_{m+1}(t)\right), \tag{11.11}$$

$$U_{m+1}'(t) = -\frac{a}{2h}\left(U_1(t) - U_m(t)\right). \tag{11.12}$$

Then the semi-discrete system can be written as

$$\mathbf{U}'(t) = A\mathbf{U}(t), \tag{11.13}$$

where

$$A = -\frac{a}{2h}\begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{bmatrix}, \tag{11.14}$$

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

Lemma 11.8. The eigenvalues of A in (11.13) are

$$\lambda_p = -\frac{ia}{h}\sin(2\pi ph) \text{ for } p = 1, 2, \ldots, m+1. \tag{11.15}$$

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

$$w_j^p = e^{2\pi ipjh} \text{ for } j = 1, 2, \ldots, m+1. \tag{11.16}$$

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

$$\begin{aligned}(A\mathbf{w}^p)_j &= -\frac{a}{2h}\left(w_j^{p+1} - w_j^{p-1}\right) \\ &= -\frac{a}{2h}e^{2\pi ipjh}\left(e^{2\pi iph} - e^{-2\pi iph}\right) \\ &= -\frac{ia}{h}\sin(2\pi ph)e^{2\pi ipjh} \\ &= \lambda_p w_j^p.\end{aligned}$$

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

Notation 12. Hereafter we define the Courant number as

$$\mu := \frac{ak}{h}, \tag{11.17}$$

where k is the uniform time-step size.

11.1.1 The FTCS method

Definition 11.9. The FTCS method for the advection equation (11.4) is

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right), \tag{11.18}$$

or in matrix form

$$\mathbf{U}^{n+1} = (I + kA)\mathbf{U}^n. \tag{11.19}$$

Corollary 11.10. The FTCS method for the advection equation (11.4) is unconditionally unstable for $k = O(h)$.

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

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

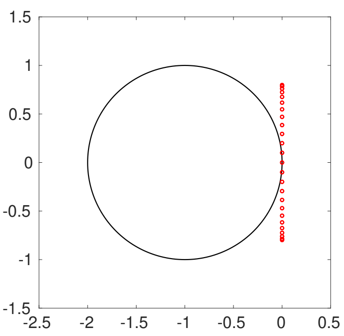
For (11.19), we have

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

which lies on the imaginary axis between $-i\mu$ and $i\mu$, and thus if $k = O(h)$, then

$$\forall p = 1, 2, \ldots, m+1, \quad z_p \notin \mathcal{S},$$

which implies the instability, as shown below. □



Lemma 11.11. The FTCS method for the advection equation has Lax-Richtmyer stability for $k = O(h^2)$.

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 some $\alpha = O(1)$, hence the skew-symmetry of A implies

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

which shows the uniform boundedness of the iteration matrix needed for Lax-Richtmyer stability. □

11.1.2 The leapfrog method

Definition 11.12. The *leapfrog method* for the advection equation (11.4) 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

$$U_j^{n+1} = U_j^{n-1} - \mu\left(U_{j+1}^n - U_{j-1}^n\right). \tag{11.20}$$

11.1.3 Lax-Friedrichs

Definition 11.13. The *Lax-Friedrichs method* for the advection equation (11.4) is

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right). \tag{11.21}$$

Lemma 11.14. Consider the IVP system

$$\mathbf{U}'(t) = A_\epsilon \mathbf{U}(t), \tag{11.22}$$

where

$$A_\epsilon = A + \frac{\epsilon}{h^2}\begin{bmatrix} -2 & 1 & & & & 1 \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ 1 & & & & 1 & -2 \end{bmatrix} \tag{11.23}$$

with A defined in (11.14). The eigenvalues of A_ϵ are

$$\lambda_p = -\frac{ia}{h}\sin(2\pi ph) - \frac{2\epsilon}{h^2}[1 - \cos(2\pi ph)] \tag{11.24}$$

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

$$w_j^p = e^{2\pi ipjh} \text{ for } j = 1, 2, \ldots, m+1. \tag{11.25}$$

Proof. This follows from Lemma 11.8 and the result on the eigenpair of the second-order discrete Laplacian. □

Lemma 11.15. The Lax-Friedrichs method can be considered as the MOL obtained by applying the forward Euler to the semidiscrete system (11.22) 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

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

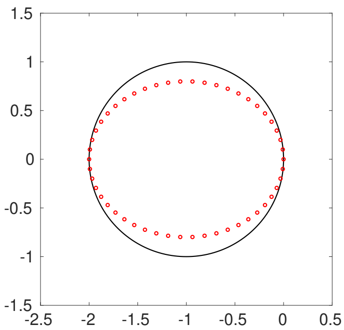
and this show the standard discretization from the advection-diffusion equation. □

Theorem 11.16. The Lax-Friedrichs method (11.21) is convergent provided that $|\mu| \le 1$.

Proof. By Lemma 11.15, we have

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

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 $|\mu| \le 1$. □



11.1.4 Lax-Wendroff

Definition 11.17. The *Lax-Wendroff method* for the advection equation (11.4) is

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right). \end{aligned} \tag{11.26}$$

Lemma 11.18. The Lax-Wendroff method (11.26) 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 $u_{tt} = -au_{tx} = a^2 u_{xx}$. □

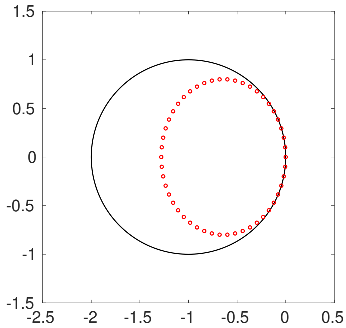
Lemma 11.19. The Lax-Wendroff method (11.26) can be considered as the MOL obtained by applying the forward Euler to the semidiscrete system (11.22) with $\epsilon = \frac{1}{2}ka^2$.

Theorem 11.20. The Lax-Wendroff method (11.26) is convergent provided $|\mu| \le 1$.

Proof. By Lemma 11.19, we have

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

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



11.1.5 The Upwind method

Definition 11.21. The *upwind method* for the advection equation (11.4) is

$$U_j^{n+1} = \begin{cases} U_j^n - \mu\left(U_j^n - U_{j-1}^n\right) & \text{if } a \ge 0; \\ U_j^n - \mu\left(U_{j+1}^n - U_j^n\right) & \text{if } a < 0. \end{cases} \tag{11.27}$$

Lemma 11.22. The upwind method (11.26) can be considered as the MOL obtained by applying the forward Euler to the semidiscrete system (11.22) with $\epsilon = \frac{ah}{2}$.

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

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

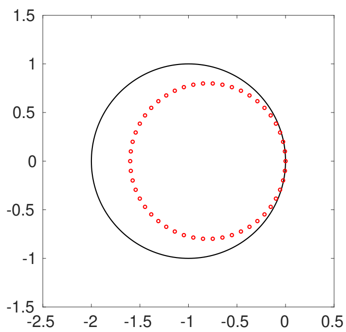
which is the forward Euler's method applied to (11.22) with $\epsilon = \frac{ah}{2}$. □

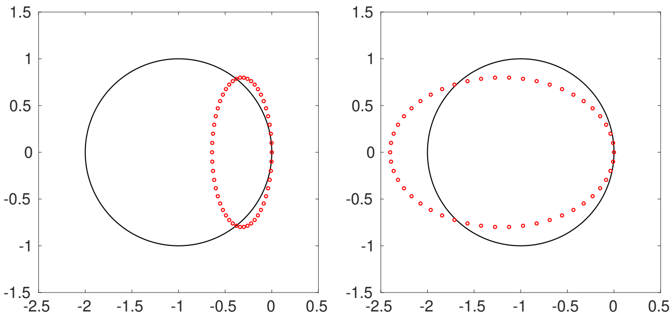
Theorem 11.23. For $a > 0$, the upwind method is convergent if and only if $\mu \le 1$; for $a < 0$, the upwind method is convergent if and only if $\mu \ge -1$.

Proof. We only prove the case of $a > 0$. By Lemma 11.22, we have

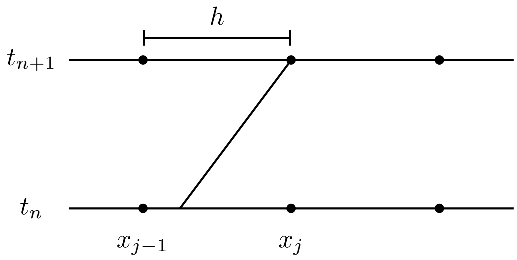
$$z_p = k\lambda_p = -i\mu\sin(2\pi ph) + \mu[\cos(2\pi ph) - 1].$$

These values all lie on a circle centered at $(-\mu, 0)$ with radius μ. If $\mu \le 1$, 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. □





Corollary 11.24. The upwind method is equivalent to characteristic tracing followed by a linear interpolation.



Proof. If we take $\mu = 1$, then the upwind (11.27) method 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

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

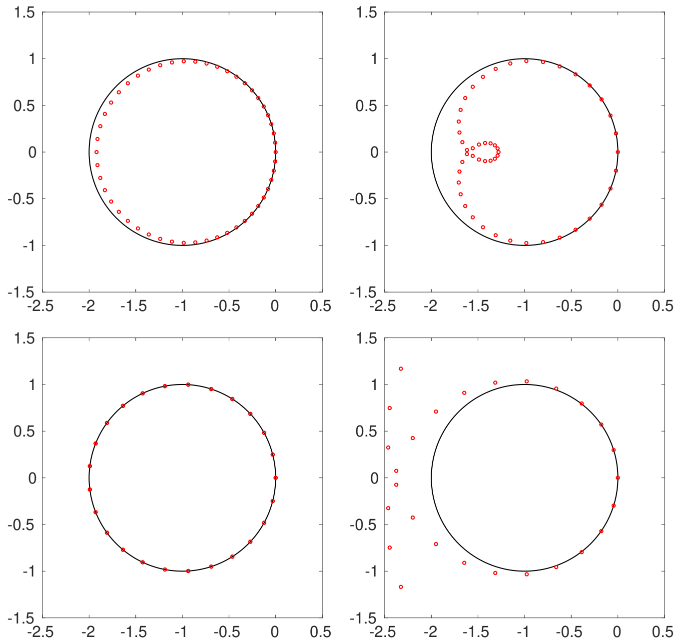
11.1.6 The Beam-Warming method

Definition 11.25. The *Beam-Warming method* solves the advection equation (11.4) by

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(3U_j^n - 4U_{j-1}^n + U_{j-2}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_j^n - 2U_{j-1}^n + U_{j-2}^n\right) \quad \text{if } a \geq 0; \qquad (11.28) \\ U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(-3U_j^n + 4U_{j+1}^n - U_{j+2}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_j^n - 2U_{j+1}^n + U_{j+2}^n\right) \quad \text{if } a < 0. \qquad (11.29) \end{aligned}$$

Exercise 11.26. Show that the Beam-Warming method is second-order accurate both in time and in space.

Exercise 11.27. Show that the Beam-Warming methods (11.28) and (11.29) 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.



11.2 The CFL condition

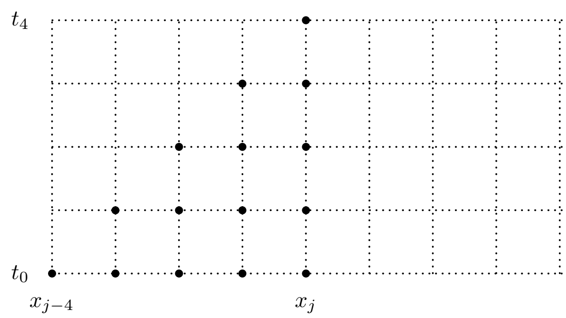
Definition 11.28. For the advection equation (11.4), the *domain of dependence* of a point $(X, T) \in \Omega$ is

$$\mathcal{D}_{\text{ADV}}(X, T) = \{X - aT\}. \qquad (11.30)$$

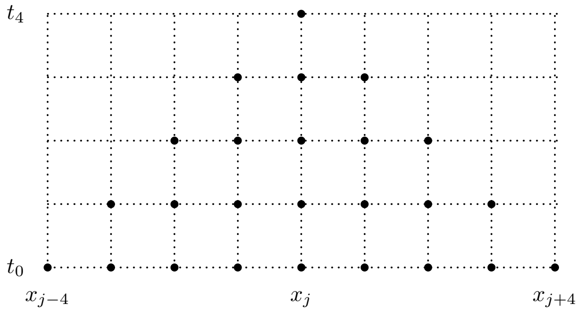
Definition 11.29. The *numerical domain of dependence* of a grid point (x_j, t_n) is the set of all grid points x_i such that U_i^0 at x_i has an effect on U_j^n.

$$\mathcal{D}_N(x_j, t_n) = \{x_i : U_i^0 \text{ affects } U_j^n\}. \qquad (11.31)$$

Example 11.30. Numerical domain dependence of a grid point using the upwind method.



Example 11.31. Numerical domain dependence of a grid point using a 3-point explicit method.



Theorem 11.32 (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 show that if some point p in the domain of dependence is not contained in the numerical domain of dependence, then the numerical method cannot be convergent. Because it has no control over the value of p that affects the true solution, the numerical method cannot be convergent. $\square$

Example 11.33. The heat equation

$$\begin{cases} u_t = \nu u_{xx} \\ u(x,0) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(x-\bar{x})^2}, \end{cases} \tag{11.32}$$

has its exact solution as

$$u(x,t) = \frac{1}{\sqrt{4\pi\nu t + \pi/\beta}} e^{-(x-\bar{x})^2/(4\nu t + 1/\beta)}. \tag{11.33}$$

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

$$\mathcal{D}_{\text{DIFF}}(X,T) = (-\infty, +\infty) \tag{11.34}$$

because an initial point source

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

instantaneously affect 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.

11.3 Modified equations

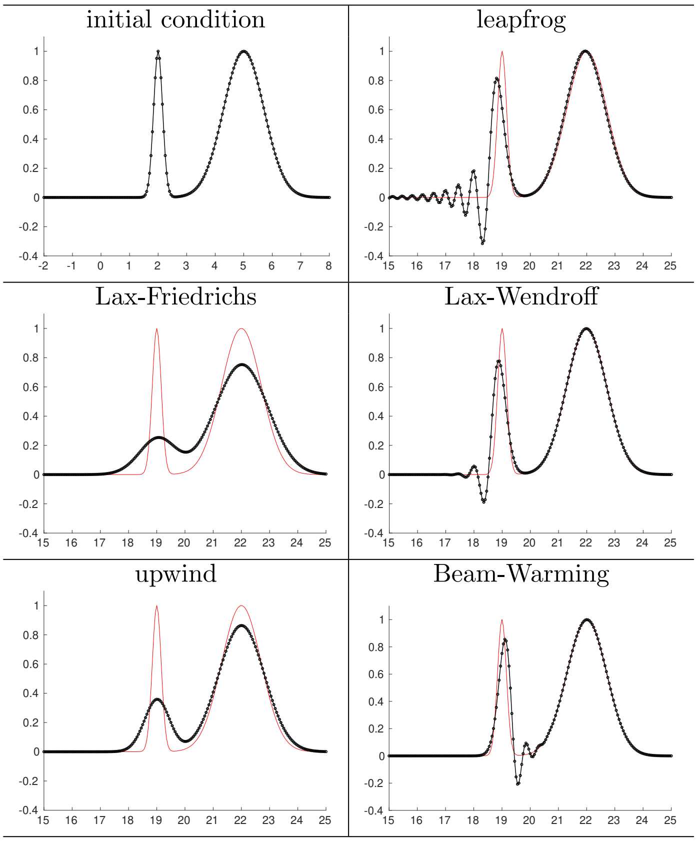
Example 11.34. For the advection equation

$$u_t + u_x = 0$$

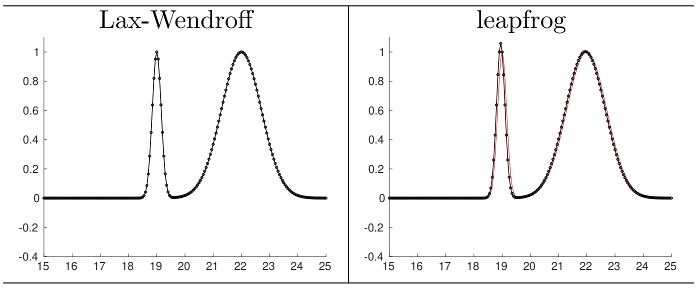
with initial condition

$$u(x,0) = \eta(x) = \exp(-20(x-2)^2) + \exp(-(x-5)^2), \tag{11.35}$$

the exact solution at $t = T$ is simply the initial data shifted by T. We solve this IVP 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 do the numerical solution of the leapfrog method contains 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.

Exercise 11.35. Reproduce all results in Example 11.34.

Definition 11.36. The *modified equation of an MOL for solving a PDE* (the original equation) is a PDE obtained from the formula of the MOL by

(1) replacing U_j^n with a smooth grid function $v(x_j, t^n)$ in the MOL formula,
(2) expanding all terms in Taylor series at (x_j, t^n),
(3) neglecting potentially high-order terms.

Example 11.37. 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

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

Combining these gives

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

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

$$v_t + av_x = \frac{1}{2}ah\left(1-\mu\right)v_{xx} := \beta v_{xx}, \tag{11.36}$$

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

Exercise 11.38. Derive the modified equation of the Lax-Wendroff method for the advection equation as

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = 0. \tag{11.37}$$

Example 11.39. By Lemma E.16, The solution to the modified equation (11.37) 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, (11.37) and Example E.27 yield

$$\begin{aligned}C_p(\xi) &= a - \frac{ah^2}{6}\left(1-\mu^2\right)\xi^3,\\ C_g(\xi) &= a - \frac{ah^2}{2}\left(1-\mu^2\right)\xi^3.\end{aligned}$$

First, the phase velocity $C_p \neq a$ for $\mu \neq 1$, and its value depends on ξ; this answers Question (b) of Example 11.34. 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 11.34.

Exercise 11.40. Show that the modified equation of the leapfrog method is also (11.37). However, if one more term of higher-order derivative had been retained, the modified equation of the leapfrog method would have been

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = \epsilon_f v_{xxxxx} \tag{11.38}$$

while that of the Lax-Wendroff method would have been

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = \epsilon_w v_{xxxx}. \tag{11.39}$$

Exercise 11.41. Show that the modified equation of the Beam-Warming method is

$$v_t + av_x + \frac{ah^2}{6}\left(-2+3\mu-\mu^2\right)v_{xxx} = 0. \tag{11.40}$$

Thus we have

$$\begin{aligned}C_p(\xi) &= a + \frac{ah^2}{6}\left(\mu-1\right)\left(\mu-2\right)\xi^3,\\ C_g(\xi) &= a + \frac{ah^2}{2}\left(\mu-1\right)\left(\mu-2\right)\xi^3.\end{aligned}$$

How do these facts answer Question (e) of Example 11.34?

Exercise 11.42. What if $\mu = 1$? Discuss this case for both Lax-Wendroff and leapfrog methods to answer Question (f) of Example 11.34.

11.4 Von Neumann analysis

Exercise 11.43. Apply the von Neumann analysis to the upwind method to derive its amplification factor as

$$g(\xi) = (1-\mu) + \mu e^{-i\xi h}. \tag{11.41}$$

For which values of μ would the method be stable?

Exercise 11.44. Apply the von Neumann analysis to the Lax-Friedrichs method to derive its amplification factor as

$$g(\xi) = \cos(\xi h) - \mu i \sin(\xi h). \tag{11.42}$$

For which values of μ would the method be stable?

Exercise 11.45. Apply the von Neumann analysis to the Lax-Wendroff method to derive its amplification factor as

$$g(\xi) = 1 + 2\mu^2 \sin^2 \frac{\xi h}{2} - i\mu \sin(\xi h). \tag{11.43}$$

For which values of μ would the method be stable?

Example 11.46. 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

$$v(x_j, t_n) := e^{i(\xi x_j - \omega t_n)}. \tag{11.44}$$

For the leapfrog method, this form yields

$$\sin(\omega k) = \mu \sin(\xi h), \tag{11.45}$$

which yield the group velocity as

$$\frac{\mathrm{d}\omega}{\mathrm{d}\xi} = \pm \frac{a\cos(\xi h)}{\sqrt{1-\mu^2\sin^2(\xi h)}}, \tag{11.46}$$

where the $\pm$ follows from the multivalued dispersion relation (11.45). For high-frequency modes satisfying $\xi h \approx \pi$, the group velocity may have a sign different from that of a.