Chapter 8

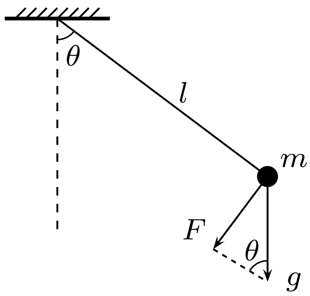
Initial Value Problems

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,

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

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

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

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

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

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

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

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 $S = \max_{(\mathbf{u},t)\in\mathcal{D}} \|\mathbf{f}(\mathbf{u}, t)\|$.

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.

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

$$\lim_{u\to 0} f'(u) = \lim_{u\to 0}\frac{1}{2\sqrt{u}} = +\infty.$$

Hence $f(u)$ is not Lipschitz near $u = 0$. However, $u(t) \equiv 0$ and $u(t) = \frac{1}{4}t^2$ 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 $\min_{a>0}(\eta + a)^2/a$:

$$(\eta + a)^2/a = 2\eta + a + \eta^2\frac{1}{a} \ge 2\eta + 2\sqrt{\eta^2} = 4\eta.$$

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 $|\omega - \omega^*| \le \|\mathbf{u} - \mathbf{u}^*\|_\infty$.

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

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

$$B = S^{-1}AS, \quad (8.10)$$

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.

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.,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

is a solution to (8.16).

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

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

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

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

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

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

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

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

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 $k = 9.76\text{e-}4$ and $k = 9.50\text{e-}4$. What's going on?

Definition 8.43. The Euler's method

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

for solving the scalar IVP

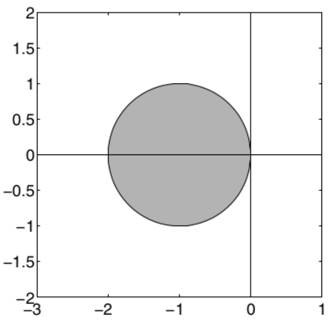
$$u'(t) = \lambda u(t) \tag{8.29}$$

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

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

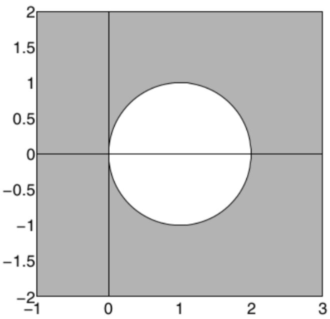


Example 8.45. The backward Euler's method applied to (8.29) reads

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

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).

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

$$\mathrm{A} \underset{c_2}{\overset{c_1}{\rightleftharpoons}} \mathrm{B},$$

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

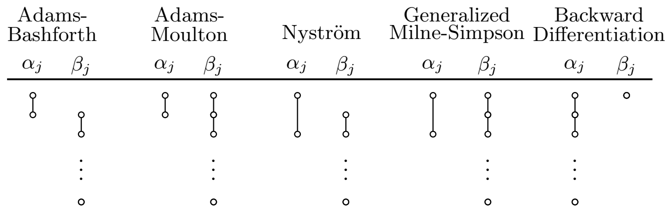
Definition 8.64. For solving the IVP (8.3), an *s-step linear multistep method* (LMM) has the form

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

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

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

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

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

where α_j's are chosen to give order s.

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

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

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

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

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 $C_{p+1} \neq 0$. 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

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

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

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

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

as $z \to 1$.

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.

8.6.4 Linear difference equations

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

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

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

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

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

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

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

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

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

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

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.

Lemma 8.103. A convergent LMM is preconsistent.

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

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

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

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.

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

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

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

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

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

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

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

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.

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),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa = \zeta - (1 + \kappa), \qquad (8.80)$$

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

$$\mathcal{R} = \{\kappa : |1 + \kappa| \le 1\}. \qquad (8.81)$$

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

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa\zeta = (1 - \kappa)\zeta - 1, \qquad (8.82)$$

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

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

Example 8.117. For the trapezoidal method (8.20),

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

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

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

Example 8.118. For the midpoint method (8.21),

$$\pi_\kappa(\zeta) = \zeta^2 - 2\kappa\zeta - 1. \qquad (8.86)$$

$\pi_z(\zeta) = 0$ 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:

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

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.

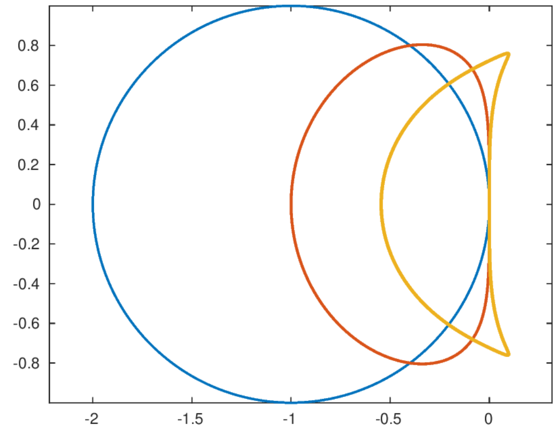
(a) compute the *root locus curve*

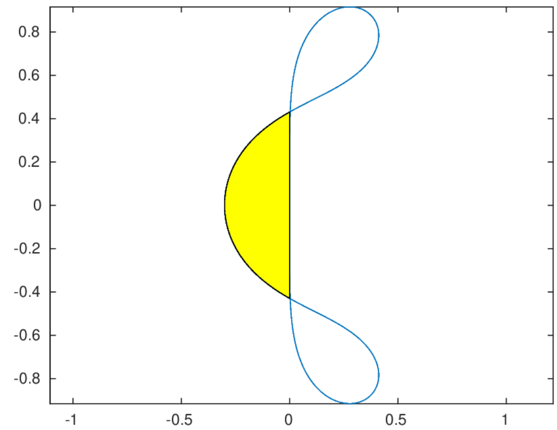
$$\gamma(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \qquad \theta = [0, 2\pi]; \tag{8.88}$$

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

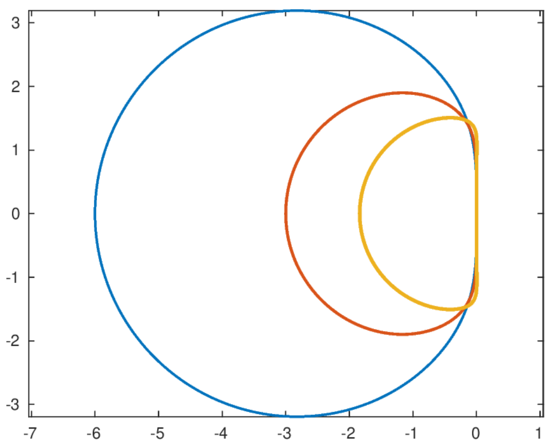
(c) for each connected region $S \subset \mathbb{C}$, choose a convenient interior point $\kappa_p \in S$ and solve the equation $\rho(\zeta) - \kappa_p \sigma(\zeta) = 0$: S is part of the RAS if all roots are in the unit disk; otherwise S is not.

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.

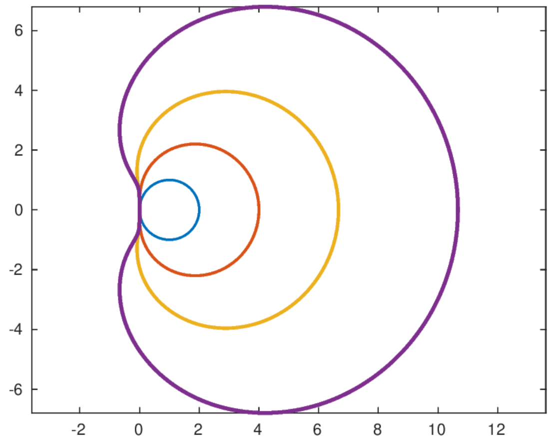




Example 8.121. The RASs of Adams-Moulton formulas with $p = 3, 4, 5$ are shown below. Each RAS is bounded.



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

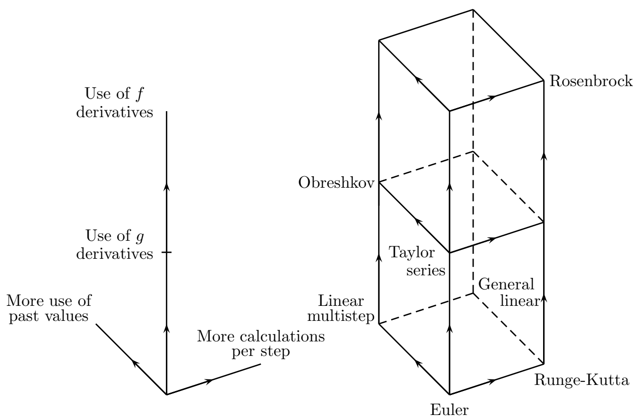
The proofs of conclusions in this subsection can be found in *Hairer et. al. 1993 Solving Ordinary Differential Equations I, Springer 2nd ed.*

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 are real coefficients for $i, j = 1, 2, \ldots, s$ and

$$c_i = \sum_{j=1}^{i-1} a_{i,j}. \tag{8.96}$$

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 $i = 1, 2, \ldots, s$, the cooefficients $a_{i,j}$, b_j are real, and c_i satisfies (8.96).

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, two-stage diagonally implicit Runge-Kutta 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

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

It is *consistent* if it has order of accuracy $p \geq 1$.

Example 8.142. For the modified Euler method, we have

$$\frac{U^{n+1}-U^n}{k} = f\left(U^n + \frac{k}{2} f(U^n, t_n), t_n + \frac{k}{2}\right) \qquad (8.104)$$

and thus the one-step error is

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

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 8.143. Derive the $O(k^3)$ term in Example 8.142 to verify that it does not valish.

Theorem 8.144. A multistage method is consistent if and only if

$$\lim_{k\to 0} \Phi(u, t; k) = f(u, t) \qquad (8.105)$$

for any $(u, t) \in \mathcal{D}$ where $\mathcal{D}$ is the domain of f,

$$\mathcal{D} = \{(u, t) : |u - u_0| \le a, t \in [0, T]\}.$$

Corollary 8.145. The Euler method is consistent.

Definition 8.146. A multistage method is *convergent* if its solution error tends to zero as $k \to 0$ for any $T > 0$, i.e.,

$$\lim_{k\to 0; Nk=T} U^N = u(T). \qquad (8.106)$$

Lemma 8.147. Let (ξ_n) be a sequence in $\mathbb{R}$ such that

$$|\xi_{n+1}| \le (1 + C)|\xi_n| + D, \quad n \in \mathbb{N} \qquad (8.107)$$

for some positive constants C and D. Then we have

$$|\xi_n| \le e^{nC}|\xi_0| + \frac{D}{C}(e^{nC} - 1), \quad n \in \mathbb{N}. \qquad (8.108)$$

Theorem 8.148. Suppose the increment function Φ that describes a multistage method is continuous and satisfies a Lipschitz condition

$$|\Phi(u, t; k) - \Phi(v, t; k)| \le M|u - v| \qquad (8.109)$$

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 $|E^0| = O(k^p)$, then its solution error can be bounded as

$$|E^n| \le \frac{K}{M}\left(e^{MT} - 1\right) k^p. \qquad (8.110)$$

Corollary 8.149. 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.

Lemma 8.150. The one-step error of the classical Runge-Kutta method (8.94) is

$$\mathcal{L}u(t_n) = O(k^5). \qquad (8.111)$$

Exercise 8.151. Prove Lemma 8.150.

Corollary 8.152. 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.

8.7.3 Absolute stability

Definition 8.153. The *stability function of a one-step method* is a ratio of two polynomials

$$R(z) = \frac{P(z)}{Q(z)} \qquad (8.112)$$

that satisfies

$$U^{n+1} = R(z)U^n \qquad (8.113)$$

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

Example 8.154. The fourth-order Runge-Kutta method has its stability function as

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4. \qquad (8.114)$$

Example 8.155. The trapezoidal rule, when viewed as a one-step method has its stability function as

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, \qquad (8.115)$$

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

Exercise 8.156. Show that the TR-BDF2 method (8.101) has

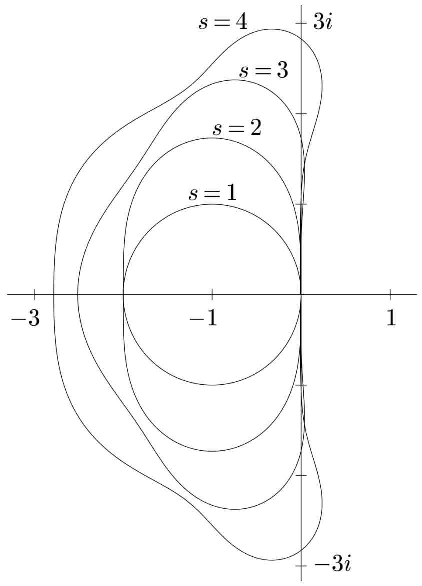
$$R(z) = \frac{1 + \frac{5}{12}z}{1 - \frac{7}{12}z + \frac{1}{12}z^2}, \qquad (8.116)$$

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

Definition 8.157. The *region of absolute stability (RAS) of a one-step method* is a subset of the complex plane

$$\mathcal{R} := \{z \in \mathbb{C} : |R(z)| \le 1\}. \qquad (8.117)$$

Example 8.158. The boundaries of RASs for ERKs with $s = 1, 2, 3, 4$ are shown below.



8.8 Stiff IVPs

Example 8.159. Consider the IVP

$$u'(t) = \lambda(u - \cos t) - \sin t, \quad u(0) = \eta. \tag{8.118}$$

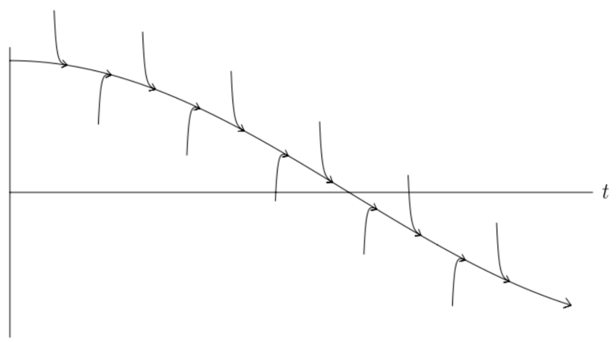
By Duhamel's principle (8.13), the exact solution is

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

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 $\eta \neq 1$ 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 $E(T) \leq \epsilon = 4 \times 10^{-5}$, the BDF2 method may use $k = 0.1$, the AB2 method has to use $k \leq 0.01$ while the time scale of the IVP is 1.

8.8.1 The notion of stiffness

Definition 8.160. 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 8.161. 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 8.162. For an IVP

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{b}(t) \tag{8.119}$$

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

$$\frac{\max_{\lambda\in\Lambda(A)} |\mathrm{Re}\,\lambda|}{\min_{\lambda\in\Lambda(A)} |\mathrm{Re}\,\lambda|}. \tag{8.120}$$

Example 8.163. Consider the linear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -1000 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad t \in [0, 1] \tag{8.121}$$

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

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

that is

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

If (8.121) 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 8.164. Consider the nonlinear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -u_1 u_2 \\ \cos(u_1) - \exp(u_2) \end{pmatrix}. \tag{8.122}$$

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

8.8.2 A-stability and L-stability

Definition 8.166. An ODE method is *A-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \mathrm{Re}\, z \leq 0\} \subseteq \mathcal{R}. \tag{8.123}$$

Example 8.167. The backward Euler's method and trapezoidal method are A-stable.

Theorem 8.168 (Dahlquist's Second Barrier)**.** The order of accuracy of an implicit A-stable LMM satisfies $p \leq 2$. An explicit LMM cannot be A-stable.

Definition 8.169. An ODE method is *$A(\alpha)$-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \pi - \alpha \leq \arg(z) \leq \pi + \alpha\} \subseteq \mathcal{R}. \tag{8.124}$$

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

Example 8.170. 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 8.171. A one-step method is *L-stable* if it is A-stable and

$$\lim_{z\to\infty} |R(z)| = 0, \tag{8.125}$$

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

Example 8.172. We use the trapezoidal and backward Euler's methods to solve the IVP (8.118) 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
$\eta = 1$	0.4	4.7770e-02	4.7770e-02
	0.2	9.7731e-08	4.7229e-10
	0.1	4.9223e-08	1.1772e-10
$\eta = 1.5$	0.4	4.7770e-02	4.5219e-01
	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 8.173. Reproduce the results in Example 8.172.