Chapter 11

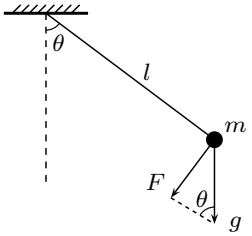
Initial Value Problems (IVPs)

Definition 11.1. A *system of ordinary differential equations* (ODEs) of dimension N is a set of differential equations of the form

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t), t), \tag{11.1}$$

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, (11.1) is an ODE for $N = 1$.

Definition 11.2. A system of ODEs is *linear* if its RHS function can be expressed as $\mathbf{f}(\mathbf{u}, t) = A(t)\mathbf{u} + \boldsymbol{\beta}(t)$, 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. A linear ODE system is further said to have *constant coefficient* if the matrix A is independent on time.



Example 11.3. 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

$$\theta''(t) = -\frac{g}{\ell} \sin\theta, \tag{11.2}$$

which admits a unique solution if we impose two initial conditions

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

Alternatively, (11.2) can be derived by the fact that the total energy remains a constant with respect to time.

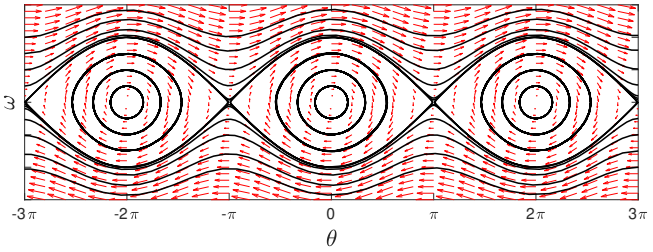
$$L = \frac{1}{2} m(\ell\theta')^2 + mg\ell(1 - \cos\theta);$$

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

The ODE in (11.2) is autonomous, nonlinear, and second-order; it can be reduced to a first-order ODE system,

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

where $\omega = \theta'$ and $\mathbf{u} = (\theta, \omega)^T$. Its solutions are illustrated as follows.



In the above plot, any point in the set

$$\{(\theta, \omega) : \omega = 0; \ \theta = 2n\pi, \ n \in \mathbb{Z}\}$$

is a *stable fixed point* to which a small perturbation changes the solution from the fixed piont to a small circle close to the point. In contrast, any point in the set

$$\{(\theta, \omega) : \omega = 0; \ \theta = (2n + 1)\pi, \ n \in \mathbb{Z}\}$$

is an *unstable fixed point* to which a small perturbation changes the solution from the fixed piont to another unstable fixed point far away from the original one.

If the initial kinetic energy is large enough, the pendulum will forever circle around the fixture; accordingly, the solution will follow the wavy curve with θ increasing indefinitely. On the other hand, if the initial kinetic energy is not large enough, the pendulum will oscillate in a manner symmetric to the vertical line of $\theta = 0$; accordingly, the solution will follow the simple closed curves. An open curve that separates these two cases connects two unstable fixed points. For more on dynamical systems, see Guckenheimer and Holmes [1983] or Wiggins [2003].

Definition 11.4. 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

$$\begin{cases} \mathbf{u}(0) &= \mathbf{u}_0, \\ \mathbf{u}'(t) &= \mathbf{f}(\mathbf{u}(t), t), \ \forall t \in [0, T]. \end{cases} \tag{11.3}$$

11.1 Mathematical foundation

11.1.1 Operator norm

Lemma 11.5. The length-scaling factor of a linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ upon any vector is bounded, i.e.,

$$\forall T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m),\ \exists M \in \mathbb{F} \text{ s.t. } \forall \mathbf{x} \in \mathbb{F}^n,\ \|T\mathbf{x}\| \le M\|\mathbf{x}\|, \tag{11.4}$$

where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$ and the *vector norm* $\|\cdot\|$ is the Euclidean 2-norm in (B.56).

Proof. Since $\mathbf{x} = \sum_j x_j \mathbf{e}_j$ and T is a linear map, we have

$$\begin{aligned}\|T\mathbf{x}\| &= \left\| \sum_j x_j T\mathbf{e}_j \right\| \le \sum_j \|x_j T\mathbf{e}_j\| = \sum_j |x_j| \|T\mathbf{e}_j\| \\ &\le \|\mathbf{x}\| \sum_j \|T\mathbf{e}_j\|,\end{aligned}$$

where the second step follows from the triangle inequality in Definition B.159, the third step from the absolute homogeneity in Definition B.159, and the last step from $|x_j| \le \|\mathbf{x}\|$. The proof is completed by setting $M := \sum_j \|T\mathbf{e}_j\|$. □

Corollary 11.6. Any linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is uniformly continuous on $\mathbb{F}^n$.

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, we have

$$\|T\mathbf{x} - T\mathbf{y}\| = \|T(\mathbf{x} - \mathbf{y})\| \le M\|\mathbf{x} - \mathbf{y}\|.$$

The rest follows from setting $\delta = \frac{\epsilon}{M}$ in Definition D.195. □

Definition 11.7. The *operator norm* of a linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is the non-negative number

$$\|T\| := \inf\{M \ge 0 : \forall \mathbf{x} \in \mathbb{F}^n, \|T\mathbf{x}\| \le M\|\mathbf{x}\|\}. \tag{11.5}$$

Corollary 11.8. $\forall \mathbf{x} \in \mathbb{F}^n,\ \|T\mathbf{x}\| \le \|T\|\|\mathbf{x}\|$.

Proof. For any given T, Lemma 11.5 implies that the set

$$\mathcal{M} := \{M \ge 0 : \forall \mathbf{x} \in \mathbb{F}^n, \|T\mathbf{x}\| \le M\|\mathbf{x}\|\}$$

is nonempty and bounded from below. For any Cauchy sequence $\{M_n\}_{n=1}^{\infty}$ in $\mathcal{M}$ that converges to c, we can take the limit to the sequence of inequalities $\|T\mathbf{x}\| \le M_n\|\mathbf{x}\|$ to obtain $\|T\mathbf{x}\| \le c\|\mathbf{x}\|$. Hence $c \in \mathcal{M}$ and $\mathcal{M}$ is closed. It follows that the infimum of $\mathcal{M}$ is contained in $\mathcal{M}$. □

Exercise 11.9. Verify that (11.5) is indeed a norm in the sense of Definition B.159.

Corollary 11.10. $\|T\| = \sup_{\|\mathbf{x}\|\le 1} \|T\mathbf{x}\| = \sup_{\|\mathbf{x}\|=1} \|T\mathbf{x}\|$.

Proof. Since T is a linear map and $\|\cdot\|$ is a norm, we have

$$\|T(c\mathbf{x})\| = \|cT\mathbf{x}\| = |c|\|T\mathbf{x}\|.$$

Hence the inequality $\|T\mathbf{x}\| \le M\|\mathbf{x}\|$ in (11.5) holds for all $\mathbf{x} \ne 0$ if and only if it holds for all $\mathbf{x}$ with $\|\mathbf{x}\| \in (0, 1]$, if and only if it holds for all $\mathbf{x}$ with $\|\mathbf{x}\| = 1$. □

Corollary 11.11. The composition of two linear maps $S \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ and $T \in \mathcal{L}(\mathbb{F}^m, \mathbb{F}^k)$ satisfies

$$\|TS\| \le \|T\|\|S\|. \tag{11.6}$$

Proof. By Corollary 11.8, we have

$$\|(TS)(\mathbf{x})\| = \|T(S\mathbf{x})\| \le \|T\|\|S\mathbf{x}\| \le \|T\|\|S\|\|\mathbf{x}\|.$$

Taking supremum of the above for $\|\mathbf{x}\| \le 1$ and applying Corollary 11.10 yield (11.6). □

Corollary 11.12. The identity function $I \in \mathcal{L}(\mathbb{F}^n)$ satisfies $\|I\| = 1$.

Proof. This follows directly from (11.5). □

Exercise 11.13. Verify that the space $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ becomes a metric space if we define the metric as $d(T, S) = \|T - S\|$.

Lemma 11.14. For $T \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^m)$, suppose for each standard basis vectors $\mathbf{e}_j$ in Definition B.33 we have $T\mathbf{e}_j \in \mathbb{R}^m$. Then T carries $\mathbb{R}^n$ into $\mathbb{R}^m$ and $\|T\|$ is consistently defined in the sense that

$$\|T\| = \sup_{\mathbf{x}\in\mathbb{R}^n;\|\mathbf{x}\|\le 1} \|T\mathbf{x}\| = \sup_{\mathbf{z}\in\mathbb{C}^n;\|\mathbf{z}\|\le 1} \|T\mathbf{z}\|. \tag{11.7}$$

Proof. First, $T(\mathbb{R}^n) \subset \mathbb{R}^m$ is trivial because the matrix of T is a real matrix. Thus, $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. From this fact, we can define an operator norm of T as

$$\|T\| = \sup_{\mathbf{x}\in\mathbb{R}^n;\|\mathbf{x}\|\le 1} \|T\mathbf{x}\|. \tag{11.8}$$

On the one hand, we have

$$\sup_{\mathbf{x}\in\mathbb{R}^n;\|\mathbf{x}\|\le 1} \|T\mathbf{x}\| \le \sup_{\mathbf{z}\in\mathbb{C}^n;\|\mathbf{z}\|\le 1} \|T\mathbf{z}\|$$

since $\mathbb{R}^n \subset \mathbb{C}^n$. On the other hand, $\forall \mathbf{z} \in \mathbb{C}^n$, $\|\mathbf{z}\| \le 1$, $\exists \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, s.t. $\mathbf{z} = \mathbf{x} + \mathbf{i}\mathbf{y}$ and $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \le 1$, and we have

$$\begin{aligned}\|T\mathbf{z}\| &= \|T\mathbf{x} + \mathbf{i}T\mathbf{y}\| = \sqrt{\|T\mathbf{x}\|^2 + \|T\mathbf{y}\|^2} \\ &\le \sqrt{\|T\|^2\|\mathbf{x}\|^2 + \|T\|^2\|\mathbf{y}\|^2} \\ &= \|T\|\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2} \le \|T\| = \sup_{\mathbf{x}\in\mathbb{R}^n;\|\mathbf{x}\|\le 1} \|T\mathbf{x}\|\end{aligned}$$

where the third step follows from Corollary 11.8, the fifth step from $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \le 1$, and the last step from (11.8). Thus we have

$$\sup_{\mathbf{x}\in\mathbb{R}^n;\|\mathbf{x}\|\le 1} \|T\mathbf{x}\| \ge \sup_{\mathbf{z}\in\mathbb{C}^n;\|\mathbf{z}\|\le 1} \|T\mathbf{z}\|,$$

which completes the proof. □

Definition 11.15. The *Hilbert-Schmidt norm* or the *Frobenius norm* of a linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is the non-negative number

$$|T| := \left(\sum_{j=1}^{n} \|T\mathbf{e}_j\|^2 \right)^{\frac{1}{2}} \tag{11.9}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is the standard basis in Definition B.33.

Exercise 11.16. Verify that (11.9) is indeed a norm in the sense of Definition B.159.

Corollary 11.17. The matrix A of $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ satisfies

$$|A| = \left(\sum_{i,j} |a_{ij}|^2 \right)^{\frac{1}{2}} \tag{11.10}$$

Proof. Since A is a linear map, we rewrite (11.9) as

$$|A|^2 = |A\mathbf{e}_1|^2 + |A\mathbf{e}_2|^2 + \cdots + |A\mathbf{e}_n|^2.$$

The proof is completed by the fact that $|\cdot|$ is the Euclidean 2-norm. □

Corollary 11.18. $\forall \mathbf{x} \in \mathbb{F}^n,\ \|T\mathbf{x}\| \le |T|\|\mathbf{x}\|$.

Proof. We have

$$\begin{aligned} \|T\mathbf{x}\| &= \left\| \sum_j x_j T\mathbf{e}_j \right\| \le \sum_j |x_j| \, \|T\mathbf{e}_j\| \\ &\le \left(\sum_j |x_j|^2 \right)^{\frac{1}{2}} \left(\sum_j \|T\mathbf{e}_j\|^2 \right)^{\frac{1}{2}} = |T|\|\mathbf{x}\|, \end{aligned}$$

where the first inequality follows from (NRM-3,4) in Definition B.159, the second inequality from the Cauchy-Schwarz inequality (B.64), and the last step from Definition 11.15. □

Corollary 11.19. The composition of two linear maps $S \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ and $T \in \mathcal{L}(\mathbb{F}^m, \mathbb{F}^k)$ satisfies

$$|TS| \le |T||S|. \tag{11.11}$$

Exercise 11.20. Prove Corollary 11.19.

Corollary 11.21. The identity function $I \in \mathcal{L}(\mathbb{F}^n)$ satisfies $|I| = \sqrt{n}$.

Proof. This follows directly from (11.9). □

Theorem 11.22. The operator norm and the Hilbert-Schmidt norm on $\mathcal{L}(\mathbb{F}^n)$ are related by

$$\|T\| \le |T| \le \sqrt{n}\|T\|. \tag{11.12}$$

Proof. Take supremum of Corollary 11.18, apply Corollary 11.10, and we have $\|T\| \le |T|$. Then $|T| \le \sqrt{n}\|T\|$ is given by

$$|T|^2 = \sum_j \|T\mathbf{e}_j\|^2 \le \sum_j \|T\|^2 \|\mathbf{e}_j\|^2 = n\|T\|^2,$$

where the inequality follows from Corollary 11.8. □

Corollary 11.23. Let d_1 and d_2 denote the metrics induced from the operator norm and the Hilbert-Schmidt norm, respectively. The identity map

$$I : (\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m), d_1) \to (\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m), d_2)$$

is uniformly continuous and has a uniformly continuous inverse.

Proof. This follows directly from Theorem 11.22 and Definition D.195. □

11.1.2 Matrix exponential

Definition 11.24. The *matrix exponential* e^A of a complex matrix $A \in \mathbb{C}^{n\times n}$ is

$$e^A := \sum_{N=0}^{\infty} \frac{1}{N!} A^N = I + A + \frac{1}{2!}A^2 + \cdots \tag{11.13}$$

Lemma 11.25. The series in (11.13) is *entry-by-entry* convergent.

Proof. Apply the Hilbert-Schmidt norm in Definition 11.15 to (11.13) and we have

$$\left|e^A\right| = \left| \sum_{N=0}^{\infty} \frac{1}{N!} A^N \right| \le \sum_{N=0}^{\infty} \frac{1}{N!} \left|A^N\right| \le \sum_{N=0}^{\infty} \frac{1}{N!} |A|^N,$$

where the second inequality follows from Corollary 11.19. Since the size of A is fixed, $|A|$ is a bounded nonnegative number. Then the ratio test (Theorem C.32) implies the convergence of the last series in the Hilbert-Schmidt norm, which, by (11.10), is equivalent to the entry-by-entry convergence. □

Theorem 11.26. Interpreted as a map $f : \mathbb{R}^{2n^2} \to \mathbb{R}^{2n^2}$, the matrix exponential $A \mapsto e^A$ is $\mathcal{C}^\infty$. In other words, every partial derivative of f to any order is entry-by-entry uniformly convergent to some continuous function.

Proof. Denote by E_j, $j = 1, \ldots, 2n^2$, an n-by-n complex matrix that has 1 or $\mathbf{i}$ in one entry and 0 in all other entries. By Definition C.99, the partial derivative of f in the direction of E_j is

$$\frac{\partial f}{\partial E_j}(A) = \frac{\mathrm{d}}{\mathrm{d}t} f(A + tE_j)\bigg|_{t=0}.$$

By (11.13) and Definition C.24, the sequence associated with the series $f(A)$ is $\left\{\frac{1}{N!}A^N\right\}$. Hence $\frac{\partial f}{\partial E_j}(A)$ is the sum of derivatives of all terms in the sequence; by the chain rule, each derivative is of the form

$$\frac{1}{N!} \sum_{i=1}^{N} g_1(A) \cdots g_{i-1}(A) \, \frac{\mathrm{d}}{\mathrm{d}t} g_i(A + tE_j)\bigg|_{t=0} g_{i+1}(A) \cdots g_N(A),$$

where each g_i is A. Taking further partial derivatives preserves this general form, except that g_i is either A or E_j and that the number of products to be summed up is increased.

The kth-order partial derivative of the Nth term $\frac{1}{N!}A^N$ in the sequence is a sum of N^k products, each product consisting of N terms and each term is either A or E_j satisfying

$$\max(|A|, |E_j|) \le M := \max(|A|, 1).$$

Hence we have, for any fixed $k \in \mathbb{N}$,

$$\left| \frac{\partial^k}{\partial E_{j_1} \cdots \partial E_{j_k}} \left(\frac{1}{N!} A^N \right) \right| \le \frac{N^k M^N}{N!}.$$

By the ratio test, the series $\frac{\partial^k f}{\partial E_{j_1} \cdots \partial E_{j_k}}(A)$ uniformly converges entry-by-entry to some function. The rest of the proof follows from Theorem C.88 and Lemma 11.25. □

Example 11.27. For the real skew-symmetric matrix

$$A = \begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix},$$

we have

$$e^A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Indeed, define

$$I_2 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad J_2 := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

and we have

$$A^{4n} = \theta^{4n} I_2,\ A^{4n+2} = -\theta^{4n+2} I_2,$$
$$A^{4n+1} = \theta^{4n+1} J_2,\ A^{4n+3} = -\theta^{4n+3} J_2.$$

It follows that

$$e^A = \cos\theta I_2 + \sin\theta J_2.$$

Lemma 11.28. If two matrices X and Y commute, then

$$e^X e^Y = e^{X+Y}. \tag{11.14}$$

Proof. By rearranging double summations, we have

$$\begin{aligned} e^X e^Y &= \left(\sum_{r=0}^{\infty} \frac{1}{r!} X^r\right)\left(\sum_{s=0}^{\infty} \frac{1}{s!} Y^s\right) = \sum_{r,s\in\mathbb{N}} \frac{1}{r!s!} X^r Y^s \\ &= \sum_{N=0}^{\infty} \sum_{k=0}^{N} \frac{X^k Y^{N-k}}{k!(N-k)!} = \sum_{N=0}^{\infty} \frac{1}{N!} \sum_{k=0}^{N} \binom{N}{k} X^k Y^{N-k} \\ &= \sum_{N=0}^{\infty} \frac{1}{N!} (X+Y)^N = e^{X+Y}, \end{aligned}$$

where the commutativity of X and Y ensures the validity of the last two steps. □

Example 11.29. If two matrices X and Y do not commute, then Lemma 11.28 does not hold, e.g.,

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},\ Y = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Corollary 11.30. The matrix e^X is nonsingular for any $X \in \mathbb{C}^{n\times n}$.

Proof. This follows from setting $Y = -X$ in (11.14) and taking determinant of both sides. □

Lemma 11.31. $\frac{\mathrm{d}}{\mathrm{d}t}(e^{tX}) = Xe^{tX}$.

Proof. By (11.13) and Theorem 11.26, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{tX}\right) &= \frac{\mathrm{d}}{\mathrm{d}t} \sum_{N=0}^{\infty} \frac{1}{N!} (tX)^N = X \sum_{N=1}^{\infty} \frac{1}{(N-1)!} (tX)^{N-1} \\ &= Xe^{tX}. \end{aligned}$$

□

Lemma 11.32. For any nonsingular matrix W, we have

$$e^{W^{-1}XW} = W^{-1} e^X W. \tag{11.15}$$

Proof. By (11.13), we have

$$\begin{aligned} e^{W^{-1}XW} &= \sum_{N=0}^{\infty} \frac{1}{N!} (W^{-1}XW)^N = \sum_{N=0}^{\infty} \frac{1}{N!} W^{-1} X^N W \\ &= W^{-1} e^X W. \end{aligned}$$

□

Corollary 11.33. For a diagonalizable matrix $A = R\Lambda R^{-1}$, we have

$$e^A = Re^{\Lambda} R^{-1}. \tag{11.16}$$

Proof. This follows directly from Lemma 11.32. □

Lemma 11.34. If $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $A \in \mathbb{C}^{n\times n}$, then $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are eigenvalues of e^A. Furthermore, if $\mathbf{u}$ is an eigenvector of A for λ_i, then $\mathbf{u}$ is an eigenvector of e^A for e^{λ_i}.

Proof. By the Schur Theorem B.179, there exist an invertible matrix P and an upper triangular matrix T such that

$$A = P^{-1} T P.$$

Then Lemma 11.32 yields

$$e^A = e^{P^{-1}TP} = P^{-1} e^T P,$$

where e^T, by Definition 11.24, is an upper triangular matrix with its diagonal entries as $e^{\lambda_1}, \ldots, e^{\lambda_n}$. If $\mathbf{u}$ is an eigenvector of A for the eigenvalue λ, then $\mathbf{u}$ is an eigenvector of A^n for the eigenvalue λ^n, the rest follows from Definition 11.24. □

Theorem 11.35. $\det e^X = e^{\text{Trace } X}$.

Exercise 11.36. Prove Theorem 11.35.

11.1.3 Lipschitz continuity

Definition 11.37. A function $\mathbf{f} : \mathbb{R}^N \times [0, +\infty) \to \mathbb{R}^N$ is *Lipschitz continuous* in its first variable over some domain

$$\mathcal{D} = \{(\mathbf{u}, t) : \|\mathbf{u} - \mathbf{u}_0\| \le a, t \in [0, T]\} \tag{11.17}$$

iff

$$\exists L \ge 0 \text{ s.t. } \forall (\mathbf{u}, t), (\mathbf{v}, t) \in \mathcal{D},\ \|\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t)\| \le L\|\mathbf{u} - \mathbf{v}\|. \tag{11.18}$$

Example 11.38. If $\mathbf{f}(\mathbf{u}, t) = \mathbf{f}(t)$, then $L = 0$.

Example 11.39. If $\mathbf{f} \notin \mathcal{C}^0$, then $\mathbf{f}$ is not Lipschitz continuous.

Definition 11.40. 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 11.41. 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 $\mathbf{f} : S \to \mathbb{R}^m$, there exist continuously differentiable functions $\mathbf{g}_1(\mathbf{x}), \mathbf{g}_2(\mathbf{x}), \ldots, \mathbf{g}_n(\mathbf{x})$ such that

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(p) + \sum_{j=1}^{n} (x_j - p_j)\mathbf{g}_j(\mathbf{x}),\quad \mathbf{g}_j(p) = \frac{\partial \mathbf{f}}{\partial x_j}(p). \tag{11.19}$$

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

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

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

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(p) = \sum_i (y_i - p_i)\mathbf{g}_i(\mathbf{y}),$$
$$\mathbf{g}_i(\mathbf{y}) = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_i}(p + t(\mathbf{y} - p))\mathrm{d}t,$$

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

Lemma 11.42. If $\mathbf{f}(\mathbf{u}, t) : \mathcal{D} \to \mathbb{R}^m$ is continuously differentiable on some compact convex set $\mathcal{D} \subseteq \mathbb{R}^{n+1}$, then $\mathbf{f}$ is Lipschitz continuous in $\mathbf{u}$ on $\mathcal{D}$.

Proof. For a fixed t, the matrix form of Theorem 11.41 yields

$$\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t) = G(\mathbf{u}, \mathbf{v})(\mathbf{u} - \mathbf{v}).$$

Take the Euclidean 2-norm and we have

$$|\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t)| = |G(\mathbf{u}, \mathbf{v})(\mathbf{u} - \mathbf{v})| \le |G(\mathbf{u}, \mathbf{v})||\mathbf{u} - \mathbf{v}|,$$

where the last step follows from Corollary 11.18. Each entry in G is a continuous function $\mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and is bounded since the compactness of $\mathcal{D}$ implies that $\mathcal{D} \times \mathcal{D}$ is compact. Hence $|G(\mathbf{u}, \mathbf{v})|$ is bounded and this completes the proof. □

11.1.4 Existence and uniqueness of solution

Lemma 11.43. Let (M, ρ) denote a complete metric space and $\phi : M \to M$ a contractive mapping in the sense that

$$\exists c \in [0,1) \text{ s.t. } \forall \eta, \zeta \in M,\ \rho(\phi(\eta), \phi(\zeta)) \le c\rho(\eta, \zeta). \quad (11.20)$$

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

Theorem 11.44 (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 in Definition 11.4 at least up to time

$$T^* = \min\left(T, \frac{a}{S}\right), \quad \text{where} \quad S = \max_{(\mathbf{u},t)\in\mathcal{D}} \|\mathbf{f}(\mathbf{u}, t)\|. \quad (11.21)$$

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

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

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 11.4, 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 $\forall \mathbf{u}, \mathbf{v} \in M$,

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

where the third step follows from Definition 11.37 and the fifth from $(*)$. Then Lemma 11.43 completes the proof. □

Example 11.45. 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 continuous 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 the existence of solutions.

Example 11.46. Consider the IVP $u'(t) = u^2$, $u_0 = \eta > 0$. The slope $f'(u) = 2u$ goes to $+\infty$ as $u \to +\infty$. So there is no guarantee for a 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 11.44, $f(u) = u^2$ 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 11.44 is thus quite conservative for this case.

Example 11.47. For the simple pendulum in Example 11.3, we have

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

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

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

Therefore, $\mathbf{f}$ is Lipschitz continuous with $L = \max\left(\frac{g}{\ell}, 1\right)$.

11.1.5 Well-posedness

Definition 11.48. The IVP in Definition 11.4 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

$$\mathbf{v}' = \mathbf{f}(\mathbf{v}, t) + \boldsymbol{\delta}(t), \qquad \mathbf{v}(0) = \mathbf{u}_0 + \boldsymbol{\epsilon}_0 \tag{11.22}$$

satisfies

$$\forall t \in [0, T], \begin{cases} \|\boldsymbol{\epsilon}_0\| < \epsilon \\ \|\boldsymbol{\delta}(t)\| < \epsilon \end{cases} \Rightarrow \|\mathbf{u}(t) - \mathbf{v}(t)\| \le c\epsilon. \tag{11.23}$$

Theorem 11.49. If $\mathbf{f}(\mathbf{u}, t)$ is Lipschitz continuous 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 11.4 is well-posed.

Proof. Theorem 11.44 has already established the existence and uniqueness of the solution of the IVP. It remains to prove that the solution is overly sensitive neither to the initial condition nor to the RHS $\mathbf{f}(\mathbf{u}, t)$. To this end, we consider two IVPs

$$\begin{cases} \mathbf{v}' = \mathbf{f}(\mathbf{v}, t) \\ \mathbf{v}(0) = \mathbf{v}_0 \end{cases} \quad \text{and} \quad \begin{cases} \mathbf{w}' = \mathbf{f}(\mathbf{w}, t) + \boldsymbol{\delta}(t) \\ \mathbf{w}(0) = \mathbf{w}_0 \end{cases}$$

where $\mathbf{v}_0$, $\mathbf{w}_0$ and $\boldsymbol{\delta}(t)$ satisfy the conditions

$$(*): \begin{cases} \forall t \in [0, T] \ \ \|\boldsymbol{\delta}(t)\| < \epsilon \\ \|\mathbf{v}_0 - \mathbf{w}_0\| < \epsilon \end{cases},$$

with ϵ being a small positive constant. Obviously, the two IVPs are respectively solved by

$$\mathbf{v}(t) = \mathbf{v}_0 + \int_0^t \mathbf{f}(\mathbf{v}(s), s)\mathrm{d}s,$$
$$\mathbf{w}(t) = \mathbf{w}_0 + \int_0^t [\mathbf{f}(\mathbf{w}(s), s) + \boldsymbol{\delta}(s)]\mathrm{d}s.$$

Thus we have

$$\mathbf{v}(t) - \mathbf{w}(t) = \mathbf{v}_0 - \mathbf{w}_0 + \int_0^t [\mathbf{f}(\mathbf{v}(s), s) - \mathbf{f}(\mathbf{w}(s), s) - \boldsymbol{\delta}(s)]\mathrm{d}s.$$

Take the Euclidean 2-norm on both sides and we have

$$\begin{aligned} & \|\mathbf{v}(t) - \mathbf{w}(t)\| \\ = \ & \left\|\mathbf{v}_0 - \mathbf{w}_0 + \textstyle\int_0^t [\mathbf{f}(\mathbf{v}(s), s) - \mathbf{f}(\mathbf{w}(s), s) - \boldsymbol{\delta}(s)]\mathrm{d}s\right\| \\ \le \ & \|\mathbf{v}_0 - \mathbf{w}_0\| + \left\|\textstyle\int_0^t [\mathbf{f}(\mathbf{v}(s), s) - \mathbf{f}(\mathbf{w}(s), s) - \boldsymbol{\delta}(s)]\mathrm{d}s\right\| \\ \le \ & \|\mathbf{v}_0 - \mathbf{w}_0\| + \textstyle\int_0^t \|\mathbf{f}(\mathbf{v}(s), s) - \mathbf{f}(\mathbf{w}(s), s) - \boldsymbol{\delta}(s)\|\mathrm{d}s \\ \le \ & \|\mathbf{v}_0 - \mathbf{w}_0\| + \textstyle\int_0^t \|\mathbf{f}(\mathbf{v}(s), s) - \mathbf{f}(\mathbf{w}(s), s)\| + \|\boldsymbol{\delta}(s)\|\mathrm{d}s \\ \le \ & \epsilon + \textstyle\int_0^t L\|\mathbf{v}(s) - \mathbf{w}(s)\|\mathrm{d}s + \int_0^t \epsilon \mathrm{d}s \\ = \ & (1 + t)\epsilon + \textstyle\int_0^t L\|\mathbf{v}(s) - \mathbf{w}(s)\|\mathrm{d}s, \end{aligned}$$

where the fifth step follows from $(*)$. To proceed with

$$(**): \ \|\mathbf{v}(t) - \mathbf{w}(t)\| \le (1 + t)\epsilon + \int_0^t L\|\mathbf{v}(s) - \mathbf{w}(s)\|\mathrm{d}s,$$

we define a function $h : [0, T] \to \mathbb{R}$,

$$h(s) = e^{-sL} \int_0^s L\|\mathbf{v}(r) - \mathbf{w}(r)\|\mathrm{d}r,$$

of which the derivative is

$$h'(s) = Le^{-sL} \left(\|\mathbf{v}(s) - \mathbf{w}(s)\| - \int_0^s L\|\mathbf{v}(r) - \mathbf{w}(r)\|\mathrm{d}r \right).$$

Consequently, we have

$$\begin{aligned} & e^{-tL} \textstyle\int_0^t L\|\mathbf{v}(r) - \mathbf{w}(r)\|\mathrm{d}r \\ = \ & h(t) = h(t) - h(0) = \textstyle\int_0^t h'(s)\mathrm{d}s \\ = \ & \textstyle\int_0^t Le^{-sL} \left(\|\mathbf{v}(s) - \mathbf{w}(s)\| - \int_0^s L\|\mathbf{v}(r) - \mathbf{w}(r)\|\mathrm{d}r \right) \mathrm{d}s \\ \le \ & \textstyle\int_0^t Le^{-sL}(1 + s)\epsilon \mathrm{d}s \\ \Longrightarrow \ & \textstyle\int_0^t L\|\mathbf{v}(r) - \mathbf{w}(r)\|\mathrm{d}r \le \int_0^t Le^{L(t-s)}(1 + s)\epsilon \mathrm{d}s, \end{aligned}$$

where the first step follows from the definition of $h(t)$, the second from $h(0) = 0$ and the fifth from $(**)$. Substitute the above inequality to $(**)$ and we have

$$\begin{aligned} \|\mathbf{v}(t) - \mathbf{w}(t)\| &\le (1 + t)\epsilon + \int_0^t Le^{L(t-s)}(1 + s)\epsilon \mathrm{d}s \\ &\le \left(1 + T + TLe^{LT}(1 + T)\right)\epsilon =: C\epsilon. \end{aligned}$$ □

Exercise 11.50. Show that the solution of IVP is not overly sensitive to the initial condition of problems that satisfy a Lipschitz condition. In other words, if both $\mathbf{v}$ and $\mathbf{w}$ satisfy the *same* IVP with initial condition $\mathbf{v}(a) = \mathbf{v}_0$ and $\mathbf{w}(a) = \mathbf{w}_0$, then we have

$$\|\mathbf{v}(t) - \mathbf{w}(t)\| \le \|\mathbf{v}_0 - \mathbf{w}_0\| \exp(L(t - a)).$$

11.1.6 Linear IVPs with constant coefficients

Theorem 11.51 (Duhamel's principle)**.** For a linear IVP

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{g}(t) \tag{11.24}$$

with a time-independent matrix A, the solution is

$$\mathbf{u}(t) = e^{tA}\mathbf{u}_0 + \int_0^t e^{(t-\tau)A}\mathbf{g}(\tau)\mathrm{d}\tau. \tag{11.25}$$

Proof. This solution follows from Lemma 11.31 and 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 11.52. Many linear problems are naturally formulated in the form of a single high-order ODE

$$v^{(m)}(t) - \sum_{j=1}^{m} c_j v^{(m-j)}(t) = \phi(t). \tag{11.26}$$

By setting $u_j(t) = v^{(j-1)}(t)$ and $\mathbf{u} = [u_1, u_2, \ldots, u_m]^T$, we can rewrite (11.26) as

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{g}(t), \tag{11.27}$$

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

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

If A is diagonalizable, say $A = X^{-1}\Lambda X$, then we can define $\mathbf{v} = X\mathbf{u}$ and rewrite (11.24) as

$$\mathbf{v}'(t) = \Lambda\mathbf{v} + X\mathbf{g}(t).$$

Thus the linear system can be decoupled into a number of scalar IVPs, each of which has its solution from Theorem 11.51.

Example 11.53. The matrix of the linear IVP system

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \tag{11.28}$$

is not diagonalizable. For v we have $v(t) = v(0)e^{\lambda t}$. For u we consider the form of the solution

$$u = (E + Ft)e^{\lambda t};$$

then (11.28) yields $E = u(0)$ and $F = v(0)$, i.e.,

$$u(t) = u(0)e^{\lambda t} + tv(t).$$

11.2 Basic numerical methods

Notation 9. To numerically solve the IVP (11.3), we are given initial condition $\mathbf{U}^0 = \mathbf{u}_0$, and want to compute approximations $\mathbf{U}^1, \mathbf{U}^2, \ldots$ such that

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

where k is the uniform time-step size and $t_n = nk$.

Definition 11.54. The *(forward) Euler's method* solves the IVP (11.3) by

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^n, t_n), \tag{11.29}$$

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 11.55. The *backward Euler's method* solves the IVP (11.3) by

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^{n+1}, t_{n+1}), \tag{11.30}$$

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 11.56. The *trapezoidal method* is

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \frac{k}{2}\left(\mathbf{f}(\mathbf{U}^n, t_n) + \mathbf{f}(\mathbf{U}^{n+1}, t_{n+1})\right). \tag{11.31}$$

Definition 11.57. The *midpoint (or leapfrog) method* is

$$\mathbf{U}^{n+1} = \mathbf{U}^{n-1} + 2k\mathbf{f}(\mathbf{U}^n, t_n). \tag{11.32}$$

Example 11.58. Applying Euler's method (11.29) 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	

11.2.1 Truncation and solution errors

Definition 11.59. The *local truncation error* (LTE) is the error caused by replacing continuous derivatives with finite difference formulas.

Example 11.60. The LTE of the leapfrog method 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 11.61. 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

$$\mathcal{L}^n := \mathbf{u}(t_{n+1}) - \boldsymbol{\phi}(\mathbf{u}(t_{n+1}), \mathbf{u}(t_n), \ldots, \mathbf{u}(t_{n-m})). \tag{11.33}$$

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

Example 11.62. The one-step error of the leapfrog method is

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

Definition 11.63. The *solution error* of a numerical method for solving the IVP in Definition 11.4 is

$$\mathbf{E}^N := \mathbf{U}^{T/k} - \mathbf{u}(T); \qquad \mathbf{E}^n = \mathbf{U}^n - \mathbf{u}(t_n). \tag{11.34}$$

11.2.2 Convergence of Euler's method

Definition 11.64. A numerical method is *convergent* iff its application to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\forall T > 0, \qquad \lim_{\substack{k \to 0 \\ Nk = T}} \mathbf{U}^N = \mathbf{u}(T). \tag{11.35}$$

Lemma 11.65. Consider a linear IVP (11.36) of the form

$$\begin{cases} \mathbf{u}'(t) = \lambda \mathbf{u}(t) + \mathbf{g}(t), \\ \mathbf{u}(0) = \mathbf{u}_0, \end{cases} \tag{11.36}$$

where λ is either a scalar or a diagonal matrix. The solution error and the LTE of Euler's method satisfy

$$\mathbf{E}^{n+1} = (1 + k\lambda)\mathbf{E}^n - k\boldsymbol{\tau}^n. \tag{11.37}$$

Proof. By Definition 11.59, 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 (11.36) 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 (11.37). □

Lemma 11.66. For the linear IVP (11.36), the solution errors and the LTEs of Euler's method satisfy

$$\mathbf{E}^n = (1 + k\lambda)^n \mathbf{E}^0 - k \sum_{m=1}^{n} (1 + k\lambda)^{n-m} \boldsymbol{\tau}^{m-1}. \tag{11.38}$$

Proof. We proceed by induction on n.

The induction basis holds because of (11.37). Suppose (11.38) 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 (11.37) and the second from the induction hypothesis. □

Theorem 11.67. Euler's method is convergent for solving the linear IVP (11.36).

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 11.66 yields

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

The LTE of Euler's method 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 $\|\mathbf{E}^0\| = O(k)$. □

Lemma 11.68. 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

$$\|\mathbf{E}^{n+1}\| \le (1 + kL)\|\mathbf{E}^n\| + k\|\boldsymbol{\tau}^n\|. \tag{11.39}$$

Proof. The definition of the LTE 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 11.69. For the nonlinear IVP in Lemma 11.68, Euler's method is convergent.

Proof. Follow the procedures of linear IVPs to show that

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

11.2.3 Zero stability and absolute stability

Definition 11.70. A numerical method is *stable* or *zero-stable* iff its application to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\forall T > 0, \qquad \lim_{\substack{k \to 0 \\ Nk = T}} \|\mathbf{U}^N\| < \infty. \tag{11.40}$$

Example 11.71. 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 11.72. The Euler's method

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

for solving the scalar IVP

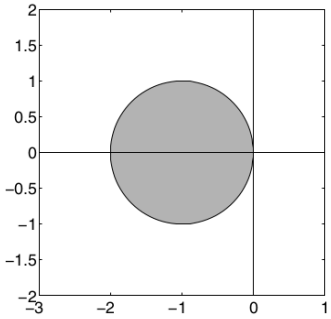
$$u'(t) = \lambda u(t) \tag{11.41}$$

is *absolutely stable* or has *eigenvalue stability* if

$$|1 + k\lambda| \le 1. \tag{11.42}$$

Definition 11.73. The *region of absolute stability* for Euler's method applied to (11.41) is the set of all points

$$\{z \in \mathbb{C} : |1 + z| \le 1\}. \tag{11.43}$$

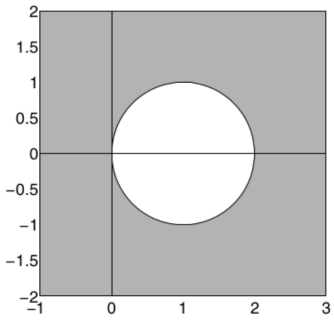


Example 11.74. The backward Euler's method applied to (11.41) reads

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

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

$$\{z \in \mathbb{C} : |1 - z| \ge 1\}. \tag{11.44}$$



Lemma 11.75. Consider an autonomous, homogeneous, and linear system of IVPs

$$\mathbf{u}'(t) = A\mathbf{u} \tag{11.45}$$

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 (11.43).

Proof. Applying Euler's method to (11.45) 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 (11.45). □

Definition 11.76. 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 11.77. 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 11.78. Consider

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

with $c_1, c_2 > 0$. Let

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

Then we have

$$\begin{aligned} u_1' &= -c_1 u_1 u_2 + c_2 u_3; \\ u_2' &= -c_1 u_1 u_2 + c_2 u_3; \\ u_3' &= c_1 u_1 u_2 - c_2 u_3, \end{aligned}$$

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 of 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 matrix, 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 $\lambda_2 = \lambda_3 = 0$. As the total concentration of species A and B, the time-dependent quantity $u_1 + u_2 \in \mathbb{R}^+$ is bounded by the constant $M_0 := u_1(0)+u_2(0)+2u_3(0)$. Therefore, the condition of absolute stability for Euler's method

$$|1 + \lambda_1 k| \le 1$$

can always be satisfied by setting $k < \frac{2}{c_1 M_0 + c_2}$, c.f. $\lambda_1 < 0$.

Example 11.79. For the reaction

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

we obtain the linear IVPs

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

Formula 11.80. 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 a diagonal matrix.

11.3 Linear multistep methods

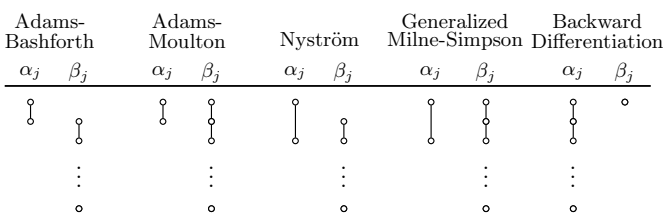
Definition 11.81. For solving the IVP (11.3), an *s-step linear multistep method* (LMM) has the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = k \sum_{j=0}^{s} \beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{11.46}$$

where $\alpha_s = 1$ is assumed WLOG.

Definition 11.82. An LMM (11.46) is *explicit* if $\beta_s = 0$; otherwise it is *implicit*.

11.3.1 Classical formulas



Definition 11.83. An *Adams formula* is an LMM of the form

$$\mathbf{U}^{n+s} = \mathbf{U}^{n+s-1} + k \sum_{j=0}^{s} \beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{11.47}$$

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

Definition 11.84. 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 11.85. 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 11.86. 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 11.87. A *Nyström formula* is an LMM of the form

$$\mathbf{U}^{n+s} = \mathbf{U}^{n+s-2} + k \sum_{j=0}^{s-1} \beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{11.48}$$

where β_j's are chosen to give order s.

Example 11.88. 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 11.89. A *backward differentiation formula* (BDF) is an LMM of the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = k\beta_s \mathbf{f}(\mathbf{U}^{n+s}, t_{n+s}), \tag{11.49}$$

where α_j's and β_s are chosen to give order s.

Example 11.90. The backward Euler's method is the 1-step BDF with

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

11.3.2 Consistency and accuracy

Definition 11.91. The *characteristic polynomials* or *generating polynomials* for the LMM (11.46) are

$$\rho(\zeta) = \sum_{j=0}^{s} \alpha_j \zeta^j; \qquad \sigma(\zeta) = \sum_{j=0}^{s} \beta_j \zeta^j. \tag{11.50}$$

Example 11.92. The forward Euler's method (11.29) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = 1, \tag{11.51}$$

while the backward Euler's method (11.30) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = \zeta. \tag{11.52}$$

Example 11.93. The trapezoidal method (11.31) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = \frac{1}{2}(\zeta + 1), \tag{11.53}$$

and the midpoint method (11.32) has

$$\rho(\zeta) = \zeta^2 - 1, \qquad \sigma(\zeta) = 2\zeta. \tag{11.54}$$

Notation 10. Denote by Z a *time shift operator* that acts on both discrete functions according to

$$Z\mathbf{U}^n = \mathbf{U}^{n+1} \tag{11.55}$$

and on continuous functions according to

$$Z\mathbf{u}(t) = \mathbf{u}(t + k). \tag{11.56}$$

Definition 11.94. The *difference operator of an LMM* is an operator $\mathcal{L}$ on the linear space of continuously differentiable functions given by

$$\mathcal{L} = \rho(Z) - k\mathcal{D}\sigma(Z), \tag{11.57}$$

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

Lemma 11.95. The one-step error of the LMM (11.46) is

$$\mathcal{L}\mathbf{u}(t_n) = C_0\mathbf{u}(t_n) + C_1 k\mathbf{u}_t(t_n) + C_2 k^2 \mathbf{u}_{tt}(t_n) + \cdots, \tag{11.58}$$

where

$$\begin{aligned} C_0 &= \textstyle\sum_{j=0}^{s} \alpha_j \\ C_1 &= \textstyle\sum_{j=0}^{s} (j\alpha_j - \beta_j) \\ C_2 &= \textstyle\sum_{j=0}^{s} \left(\frac{1}{2} j^2 \alpha_j - j\beta_j\right) \\ &\vdots \\ C_q &= \textstyle\sum_{j=0}^{s} \left(\frac{1}{q!} j^q \alpha_j - \frac{1}{(q-1)!} j^{q-1} \beta_j\right). \end{aligned} \tag{11.59}$$

Proof. Definitions 11.61, 11.94, and 11.91 yield

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

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 (11.58). □

Definition 11.96. An LMM has *order of accuracy* p if

$$\forall \mathbf{u} \in \mathcal{C}^{p+1}, \qquad \mathcal{L}\mathbf{u}(t_n) = \Theta(k^{p+1}) \text{ as } k \to 0, \tag{11.60}$$

i.e., if in (11.59) 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 11.97. 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 11.98. An LMM is *consistent* if it has order of accuracy $p \geq 1$.

Example 11.99. For Euler's method, the coefficients C_j's in (11.59) can be computed directly from Example 11.85 as $C_0 = C_1 = 0, C_2 = \frac{1}{2}, C_3 = \frac{1}{6}$.

Exercise 11.100. Compute the first five coefficients C_j's of the trapezoidal rule and the midpoint rule from Examples 11.86 and 11.88.

Example 11.101. 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 (11.58) should be zero, i.e.,

$$\sum_{j=0}^{s} \alpha_j = 0, \qquad \sum_{j=0}^{s} j\alpha_j = \sum_{j=0}^{s} \beta_j, \tag{11.61}$$

which is equivalent to

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1). \tag{11.62}$$

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 (11.61) and

$$\forall q = 2, \ldots, p, \quad \sum_{j=0}^{s} \frac{1}{q!} j^q \alpha_j = \sum_{j=0}^{s} \frac{1}{(q-1)!} j^{q-1} \beta_j. \tag{11.63}$$

Exercise 11.102. Express conditions of $\|\mathcal{L}\mathbf{u}(t_n)\| = O(k^3)$ using characteristic polynomials.

Exercise 11.103. 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 11.84

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 11.84

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	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

Backward differentiation formulas in Definition 11.89

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 11.104. 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 11.105. An LMM with $\sigma(1) \neq 0$ has order of accuracy p if and only if

$$\frac{\rho(e^\kappa)}{\sigma(e^\kappa)} = \kappa + \Theta(\kappa^{p+1}) \qquad \text{as } \kappa \to \mathbf{0}. \tag{11.64}$$

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 10, 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 11.24. Set $\kappa = k\mathcal{D}$ and we have from Definitions 11.96 and 11.94,

$$\mathcal{L} = \rho(e^\kappa) - \kappa\sigma(e^\kappa) = \Theta(k^{p+1}),$$

Since $\sigma(e^\kappa)$ is an analytic function of k and is nonzero at $k = 0$, we divide it on both sides to get (11.64). □

Theorem 11.106. An LMM with $\sigma(1) \neq 0$ has order of accuracy p if and only if

$$\begin{aligned} \tfrac{\rho(z)}{\sigma(z)} &= \log z + \Theta\left((z-1)^{p+1}\right) \\ &= \left[(z-1) - \tfrac{1}{2}(z-1)^2 + \tfrac{1}{3}(z-1)^3 - \cdots\right] \\ &\quad + \Theta((z-1)^{p+1}). \end{aligned} \tag{11.65}$$

as $z \to 1$.

Proof. We only prove the case of scalar IVPs. To get from (11.64) 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 11.107. The trapezoidal rule has $\rho(z) = z - 1$ and $\sigma(z) = \frac{1}{2}(z+1)$. A comparison of (11.65) 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 11.108. For the third-order BDF in Definition 11.89 and Exercise 11.103, derive its characteristic polynomials and apply Theorem 11.106 to verify that the order of accuracy is indeed 3.

Exercise 11.109. 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 condition u_0 and exact numerical initial data $v^0, \cdots, v^{s-1}$.

11.3.3 Zero stability

Example 11.110 (A consistent but unstable LMM)**.** The LMM

$$\mathbf{U}^{n+2} - 3\mathbf{U}^{n+1} + 2\mathbf{U}^n = -k\mathbf{f}(\mathbf{U}^n, t_n) \tag{11.66}$$

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 (11.66) 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 11.111. An s-step LMM is *stable* or *zero-stable* if all solutions $\{\mathbf{U}^n\}$ of the recurrence

$$\rho(Z)\mathbf{U}^n = \sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = \mathbf{0} \tag{11.67}$$

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

Theorem 11.112. λ is an eigenvalue of $A \in \mathbb{C}^{m\times m}$ if and only if λ is a root of the characteristic polynomial of A,

$$p_A(z) = \det(zI - A). \tag{11.68}$$

Exercise 11.113. Show that

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

is the characteristic polynomial of

$$M = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{s-2} & -\alpha_{s-1} \end{bmatrix} \in \mathbb{C}^{s\times s}. \tag{11.69}$$

Theorem 11.114. 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. WLOG, we only prove the case of scalar IVPs; see Hairer et al. [1993] for the vector case. For a scalar IVP, we write (11.67) as $U^{n+s} + \sum_{j=0}^{s-1} \alpha_j U^{n+j} = 0$, 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* (11.69) and

$$\mathbf{V}^n = (U^n, U^{n+1}, \cdots, U^{n+s-1})^T.$$

Hence

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

By Exercise 11.113, 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 $\{U^n\}_{n=0}^{+\infty}$ 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 are 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 8.19, 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 B.108, 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 8.19, 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 $m_i \ge 2$. □

11.3.4 Linear difference equations

Definition 11.115. A *system of linear difference equations* is a set of equations of the form

$$X_n = A_n X_{n-1} + \phi_n, \tag{11.70}$$

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 11.116. 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 (11.70) by writing

$$A_n = \begin{bmatrix} \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{n,s-1} & \alpha_{ns} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, X_{n-1} = \begin{bmatrix} y_{n-1} \\ y_{n-2} \\ \vdots \\ y_{n-s} \end{bmatrix},$$

$$\phi_n = \begin{bmatrix} \psi_n & 0 & 0 & \cdots & 0 \end{bmatrix}^T.$$

Theorem 11.117. The problem (11.70) with initial value X_0 has the unique solution

$$\begin{aligned} X_n = &\left(\prod_{i=1}^{n} A_i\right) X_0 \\ &+ \left(\prod_{i=2}^{n} A_i\right)\phi_1 + \left(\prod_{i=3}^{n} A_i\right)\phi_2 + \cdots + A_n\phi_{n-1} + \phi_n, \end{aligned} \tag{11.71}$$

where

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

Proof. For $n = 1$, (11.71) reduces to (11.70). The rest of the proof is a straightforward induction. □

Theorem 11.118. Let θ_n be the solution to the homogeneous linear difference equation

$$\theta_{n+s} + \sum_{i=0}^{s-1} \alpha_i \theta_{n+i} = 0 \tag{11.72}$$

with constant coefficients α_i's and the initial values

$$\begin{bmatrix} \theta_0 \\ \theta_{-1} \\ \vdots \\ \theta_{-s+2} \\ \theta_{-s+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{11.73}$$

Then the inhomogeneous equation

$$y_{n+s} + \sum_{i=0}^{s-1} \alpha_i y_{n+i} = \psi_{n+s} \tag{11.74}$$

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

$$y_n = \sum_{i=0}^{s-1} \theta_{n-i} \tilde{y}_i + \sum_{i=s}^{n} \theta_{n-i} \psi_i \tag{11.75}$$

where

$$\begin{bmatrix} \tilde{y}_{s-1} \\ \tilde{y}_{s-2} \\ \tilde{y}_{s-3} \\ \vdots \\ \tilde{y}_1 \\ \tilde{y}_0 \end{bmatrix} = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_{s-2} & \theta_{s-1} \\ 0 & 1 & \theta_1 & \cdots & \theta_{s-3} & \theta_{s-2} \\ 0 & 0 & 1 & \cdots & \theta_{s-4} & \theta_{s-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \theta_1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} y_{s-1} \\ y_{s-2} \\ y_{s-3} \\ \vdots \\ y_1 \\ y_0 \end{bmatrix}. \tag{11.76}$$

Exercise 11.119. Prove Theorem 11.118 by induction.

11.3.5 Convergence

Definition 11.120. Given initial values

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

satisfying

$$\forall i = 0, 1, \ldots, s-1, \quad \lim_{k \to 0} \|\phi^i(\mathbf{u}(0), k) - \mathbf{u}(0)\| = 0, \tag{11.77}$$

an LMM is *convergent* if it yields

$$\lim_{\substack{k \to 0 \\ Nk = T}} \mathbf{U}^N = \mathbf{u}(T) \tag{11.78}$$

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

Lemma 11.121. A convergent LMM is zero-stable.

Proof. Apply the convergent LMM to the trivial IVP

$$u'(t) = 0; \qquad u(0) = 0 \tag{11.79}$$

and we have (11.67). Suppose the LMM is not zero-stable. Then Theorem 11.114 implies that the characteristic polynomial $\rho(Z)$ either has a root ζ_1 with $|\zeta_1| > 1$ or has a multiple root ζ_2 with $|\zeta_2| = 1$. In the former case, we have

$$\sum_{j=0}^{s} \alpha_j \zeta_1^j = 0 \;\Rightarrow\; \zeta_1^m \sum_{j=0}^{s} \alpha_j \zeta_1^j = 0 \;\Rightarrow\; \sum_{j=0}^{s} \alpha_j Z^j \zeta_1^m = 0,$$

i.e., $U^n = \zeta_1^n$ is a solution of (11.67). In the latter case, we have $\rho(\zeta_2)\zeta_2^m = 0$ for any $m \in \mathbb{N}^+$. In addition, define

$$\chi(z) := (\rho(z) z^m)' = \left(\sum_{i=0}^{s} \alpha_i z^{m+i} \right)' = \sum_{i=0}^{s} \alpha_i (m+i) z^{m+i-1}$$

and we know from $\chi(\zeta_2) = 0$ that

$$\forall n \in \mathbb{N}, \; V^n = n \zeta_2^{n-1}$$

is a solution of (11.67). For (11.79) and any fixed $T > 0$,

$$U_k(T) = \sqrt{k} \zeta_1^{T/k}, \qquad V_k(T) = \frac{T}{\sqrt{k}} \zeta_2^{T/k-1}$$

satisfy condition (11.77) for initial values, but diverge as $k \to 0$, i.e. $n \to \infty$. This contradicts the convergence of the LMM and completes the proof. □

Lemma 11.122. A convergent LMM is preconsistent.

Proof. By (11.78) and the continuity of $\mathbf{u}$ in time, we have

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

where $N = T/k$. Substituting this equation into the limit of the LMM equation (11.46) yields preconsistency as in Definition 11.97. □

Lemma 11.123. A convergent LMM is consistent.

Exercise 11.124. Prove Lemma 11.123 by the approach similar with that for Lemma 11.121, i.e., by considering the particular IVP problems

$$u'(t) = f(t) = 0, u(0) = 1; \tag{11.80a}$$
$$u'(t) = f(t) = 1, u(0) = 0. \tag{11.80b}$$

Lemma 11.125. For an autonomous IVP, the one-step error of a consistent LMM satisfies

$$\|\mathcal{L}\mathbf{u}(t_n)\| \le \sum_{j=0}^{s-1} \Big(\frac{1}{2}(s-j)^2 |\alpha_j| + (s-j)|\beta_j| \Big) L M k^2, \tag{11.81}$$

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 (11.33), 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 (11.61), i.e.,

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

Integration of the automonous ODE $\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t))$ yields 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 (11.81). □

Lemma 11.126. For an autonomous IVP, the solution errors of a consistent LMM with $k < k_0$ and $k_0|\beta_s|L < 1$ satisfy

$$\left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| \le Ck \max_{i=0}^{s-1} \|\mathbf{E}^{n+i}\| + Dk^2, \quad (11.82)$$

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

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

where the third step follows from the triangular inequality

$$\left\|\mathbf{E}^{n+s}\right\| \le \left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| + \sum_{i=0}^{s-1} \left\|\alpha_i \mathbf{E}^{n+i}\right\|.$$

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 11.125 yield (11.82). □

Theorem 11.127. An LMM is convergent if and only if it is consistent and stable.

Proof. We only prove the sufficiency since the necessity has been stated in Lemmas 11.121 and 11.123. By Lemma 11.126, we have

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

where $\|\psi_{n+s}\| \le Ck \max_{i=1}^{s} \|\mathbf{E}^{n+i}\| + Dk^2$ for any k sufficiently small. Then the zero-stability of the LMM and Theorem 11.118 imply the existence of bounded constants θ_i's such that

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

where each $\widetilde{\mathbf{E}^i}$ is a linear combination of $\mathbf{E}^j$'s for $i, j = 0, 1, \ldots, s-1$; see (11.76). Note that, in order to apply Theorem 11.118, 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 appear in at most s adjacent terms. Construct a sequence (v_i) as an upper bound sequence of $(\|\mathbf{E}^i\|)$:

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

where $\lim_{k\to 0} v_0 = 0$ because Definition 11.120 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

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

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 11.128. 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

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

its numerical solution of the IVP satisfies

$$\left\|\mathbf{U}^{t/k} - \mathbf{u}(t)\right\| = O(k^p) \tag{11.83}$$

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

Proof. This proof is similar to that of Theorem 11.127. □

11.3.6 Absolute stability

Definition 11.129. The *stability polynomial* of an LMM is

$$\pi_\kappa(\zeta) := \rho(\zeta) - \kappa\sigma(\zeta) = \sum_{j=0}^{s} (\alpha_j - \kappa\beta_j)\zeta^j. \tag{11.84}$$

Definition 11.130. An LMM is *absolutely stable* for some κ if all solutions $\{\mathbf{U}^n\}$ of

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

are bounded as $n \to +\infty$, where Z is the time-shift operator in Notation 10.

Theorem 11.131 (*Root condition* for absolute stability)**.** An LMM is absolutely stable for $\kappa := k\lambda$ if and only if all the roots of $\pi_\kappa(\zeta)$ satisfy $|\zeta| \le 1$, and any root satisfying $|\zeta| = 1$ is simple.

Proof. This proof is the same as that of Theorem 11.114. □

Definition 11.132. 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 11.133. For Euler's method (11.29),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa = \zeta - (1 + \kappa), \tag{11.85}$$

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

$$\mathcal{R} = \{\kappa : |1 + \kappa| \le 1\}. \tag{11.86}$$

Example 11.134. For backward Euler's method (11.30),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa\zeta = (1 - \kappa)\zeta - 1, \tag{11.87}$$

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

$$\mathcal{R} = \{\kappa : |(1 - \kappa)^{-1}| \le 1\} = \{\kappa : |1 - \kappa| \ge 1\}. \tag{11.88}$$

Example 11.135. For the trapezoidal method (11.31),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \frac{1}{2}\kappa(\zeta + 1) = \left(1 - \frac{1}{2}\kappa\right)\zeta - \left(1 + \frac{1}{2}\kappa\right). \tag{11.89}$$

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

$$\begin{aligned} \mathcal{R} &= \left\{\kappa \in \mathbb{C} : \left|\frac{2+\kappa}{2-\kappa}\right| \le 1\right\} \\ &= \{\kappa \in \mathbb{C} : \operatorname{Re}\kappa \le 0\}. \end{aligned} \tag{11.90}$$

Example 11.136. For the midpoint method (11.32),

$$\pi_\kappa(\zeta) = \zeta^2 - 2\kappa\zeta - 1. \tag{11.91}$$

$\pi_\kappa(\zeta) = 0$ 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:

$$\mathcal{R} = \{\kappa \in \mathbb{C} : \kappa = i\alpha \text{ with } |\alpha| < 1\}. \tag{11.92}$$

Definition 11.137. The *boundary locus* method finds the RAS of an LMM (ρ, σ) with $\sigma(e^{i\theta}) \neq 0$ by steps as follows:

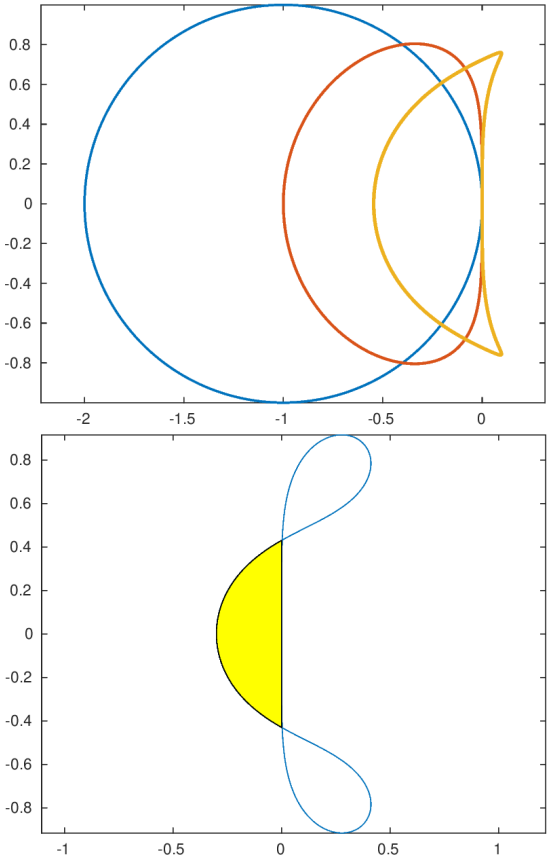
(a) compute the *root locus curve*

$$\gamma(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \qquad \theta \in [0, 2\pi]; \tag{11.93}$$

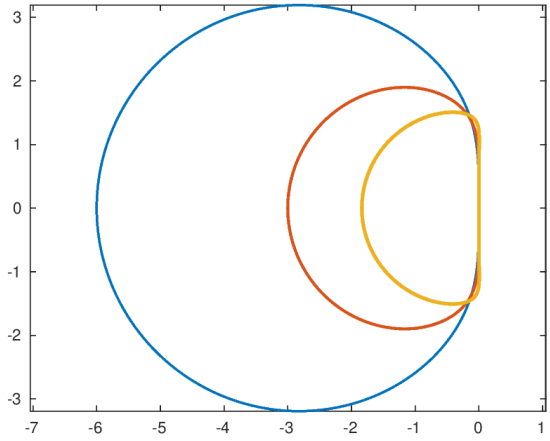
(b) the closed curve γ divides the complex plane $\mathbb{C}$ into a number of connected regions;

(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. Note that one can exit the loop upon the first finding of such a path-connected region.

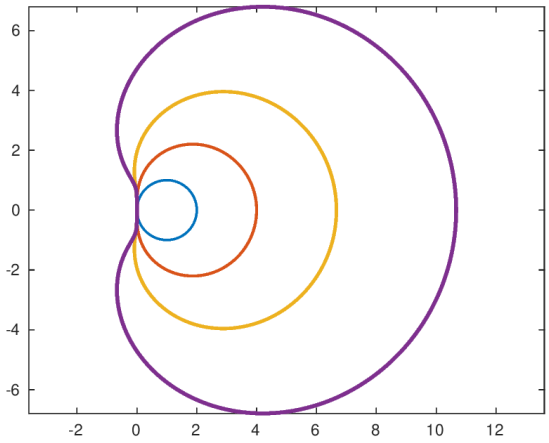
Example 11.138. The RASs of Adams-Bashforth formulas are shown below, with the first plot as those of $p = 1, 2, 3$ and the second as that of $p = 4$. Each RAS is bounded.



Example 11.139. The RASs of Adams-Moulton formulas with $p = 3, 4, 5$ are shown below. Each RAS is bounded.



Example 11.140. The RASs of backward differentiation formulas with $p = 1, 2, 3, 4$ are shown below. Each RAS is unbounded.



Exercise 11.141. Write a program to reproduce the RAS plots in Examples 11.138, 11.139, and 11.140.

Theorem 11.142. 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 $s \geq 7$.

Proof. See Hairer et al. [1993]. □

Theorem 11.143 (The first Dahlquist's barrier)**.** The order of accuracy p of a stable s-step LMM satisfies

$$p \leq \begin{cases} s & \text{if the LMM is explicit,} \\ s+1 & \text{else if } s \text{ is odd,} \\ s+2 & \text{else if } s \text{ is even.} \end{cases} \tag{11.94}$$

Proof. See Hairer et al. [1993]. □

11.4 Stiff IVPs

Example 11.144. Consider the IVP

$$u'(t) = \lambda(u - \cos t) - \sin t, \quad u(0) = \eta. \tag{11.95}$$

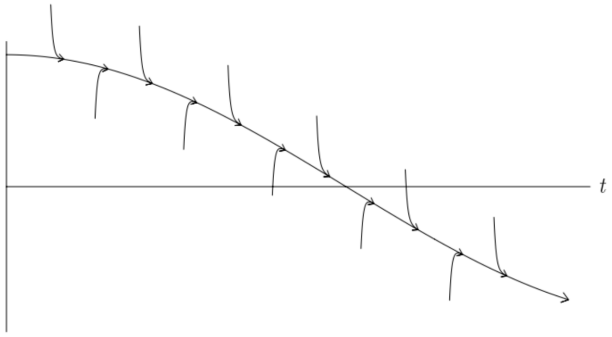
By Duhamel's principle (11.25), the exact solution is

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

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
⋮	⋮	⋮
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.

11.4.1 The notion of stiffness

Definition 11.145. An IVP is said to be *stiff in an interval* if for some initial condition any numerical method with a bounded RAS is forced to use a time-step size that is unnecessarily and excessively smaller than the time scale of the true solution of the IVP.

Definition 11.146. For an IVP

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{b}(t), \tag{11.96}$$

where $\mathbf{u}, \mathbf{b} \in \mathbb{R}^n$ 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

$$\frac{\max\limits_{\lambda \in \Lambda(A)} |\mathrm{Re}\,\lambda|}{\min\limits_{\lambda \in \Lambda(A)} |\mathrm{Re}\,\lambda|}. \tag{11.97}$$

Example 11.147. Consider the linear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -1000 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \qquad t \in [0, 1] \tag{11.98}$$

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 (11.98) is solved by a p-th order LMM with time-step size k. For U_2^N to be sufficiently accurate, we need $k = O(\epsilon^{1/p})$. But for U_1^N to be sufficiently accurate, 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 11.148. Consider the nonlinear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -u_1 u_2 \\ \cos(u_1) - \exp(u_2) \end{pmatrix}. \tag{11.99}$$

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 11.149. Let A, B and C represent chemical compounds and consider reactions of the form

$$A \xrightarrow{K_1} B \xrightarrow{K_2} C.$$

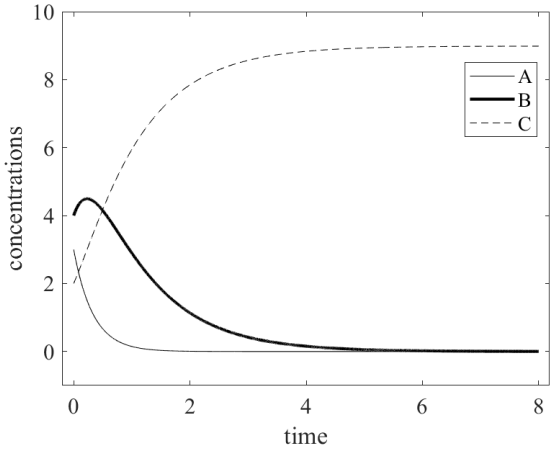
This represents a reaction in which A is transformed into B with rate $K_1 > 0$ and B is transformed into C with rate $K_2 > 0$ simultaneously. If we let u_1, u_2 and u_3 represent the concentration of A, B and C respectively, then the ODEs for u_1, u_2 and u_3 are

$$\begin{bmatrix} u_1' \\ u_2' \\ u_3' \end{bmatrix} = \begin{bmatrix} -K_1 & 0 & 0 \\ K_1 & -K_2 & 0 \\ 0 & K_2 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

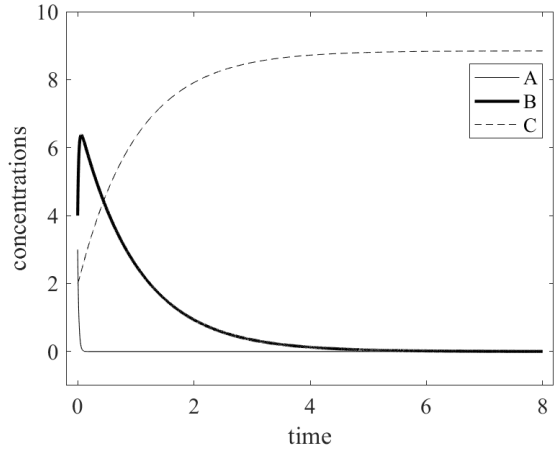
Note that the eigenvalues are $-K_1$, $-K_2$ and 0. With the assumption of $K_1 \neq K_2$, the general solution has the form

$$u_j(t) = c_{j_1} e^{-K_1 t} + c_{j_2} e^{-K_2 t} + c_{j_3}.$$

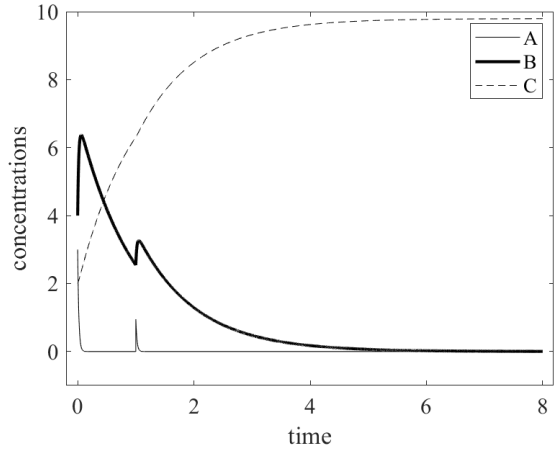
Since A decays into B which decays into C, we expect that u_1 simply decays to 0 exponentially, and also that u_2 ultimately decays to 0 (although it may first grow if K_1 is larger than K_2), while u_3 grows and asymptotically approaches the value $u_1(0) + u_2(0) + u_3(0)$ as $t \to \infty$. A typical solution for $K_1 = 3$ and $K_2 = 1$ with $u_1(0) = 3$, $u_2(0) = 4$ and $u_3(0) = 2$ is shown below.



For $K_1 \gg K_2$, e.g. $K_1 = 50$, $K_2 = 1$, we get the solution curves shown below.



Now suppose at time $t = 1$ we perturb the system by adding more of species A, then the solution curves are shown below.



The sudden increment of A is rapidly converted into B and then slowly from B into C. The different time scales of these two procedures limit the choice of time-step size.

11.4.2 A-stability

Definition 11.150. An LMM is *A-stable* if its RAS $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \operatorname{Re} z \le 0\} \subseteq \mathcal{R}. \tag{11.100}$$

Example 11.151. Both the backward Euler's method and trapezoidal method are A-stable.

Theorem 11.152 (The second Dahlquist's barrier)**.** The order of accuracy of an implicit A-stable LMM satisfies $p \le 2$. An explicit LMM cannot be A-stable.

Definition 11.153. An IVP method is *A(α)-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \pi - \alpha \le \arg(z) \le \pi + \alpha\} \subseteq \mathcal{R}. \tag{11.101}$$

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

Example 11.154. As shown in Example 11.140, 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 $p = 6$.

11.5 One-step methods

Definition 11.155. A *one-step method* constructs numerical solutions of an IVP (11.3) at each time step $n = 0, 1, \dots$ by a formula of the form

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\boldsymbol{\Phi}(\mathbf{U}^n, t_n; k), \tag{11.102}$$

where $\boldsymbol{\Phi} : \mathbb{R}^N \times [0, T] \times (0, +\infty) \to \mathbb{R}^N$ is the *increment function* given in terms of the RHS function $\mathbf{f} : \mathbb{R}^N \times [0, T] \to \mathbb{R}^N$ in (11.3).

Example 11.156. For a scalar ODE $u' = f(u, t)$, the *Taylor series method* is obtained by replacing derivatives of $u(t)$ at t_n in the truncated Taylor expansion

$$U^{n+1} = U^n + ku'_n + \frac{k^2}{2}u''_n + \cdots + \frac{k^p}{p!}u_n^{(p)} \tag{11.103}$$

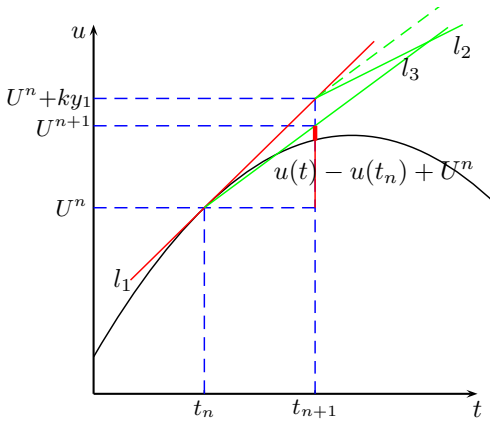
with $f(u, t)$ and its derivatives evaluated at (U^n, t_n),

$$\begin{aligned} u'_n &= f, \\ u''_n &= f_u f + f_t, \\ u'''_n &= f_u^2 f + f_{uu} f^2 + f_u f_t + 2f_{tu} f + f_{tt}, \\ &\cdots \end{aligned}$$

The one-step error of (11.103) is $O(k^{p+1})$. This method also applies to ODE systems.

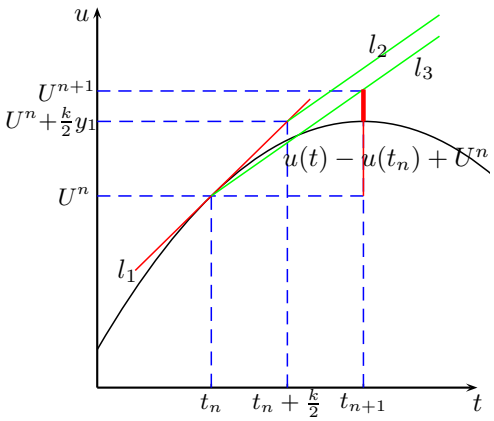
Definition 11.157. The *improved Euler method* or the *explicit trapezoidal method* is a one-step method of the form

$$\begin{cases} \mathbf{y}_1 = \mathbf{f}(\mathbf{U}^n, t_n), \\ \mathbf{y}_2 = \mathbf{f}(\mathbf{U}^n + k\mathbf{y}_1, t_n + k), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + \frac{k}{2}(\mathbf{y}_1 + \mathbf{y}_2). \end{cases} \tag{11.104}$$



Definition 11.158. The *modified Euler method* or the *improved polygon method* or the *explicit midpoint method* is a one-step method of the form

$$\begin{cases} \mathbf{y}_1 = \mathbf{f}(\mathbf{U}^n, t_n), \\ \mathbf{y}_2 = \mathbf{f}(\mathbf{U}^n + \frac{k}{2}\mathbf{y}_1, t_n + \frac{k}{2}), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{y}_2. \end{cases} \tag{11.105}$$



Exercise 11.159. Does the length of the thick red line segment in the above figure represent the one-step error in Definition 11.162? If so, prove it; otherwise derive an expression of the represented quantity.

Definition 11.160. The *TR-BDF2 method* is a one-step method of the form

$$\begin{cases} \mathbf{U}^* = \mathbf{U}^n + \frac{k}{4}\left(\mathbf{f}(\mathbf{U}^n, t_n) + \mathbf{f}(\mathbf{U}^*, t_n + \frac{k}{2})\right), \\ \mathbf{U}^{n+1} = \frac{1}{3}\left(4\mathbf{U}^* - \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^{n+1}, t_{n+1})\right). \end{cases} \tag{11.106}$$

Exercise 11.161. Give a geometric interpretation of TR-BDF2 by drawing a figure similar to those for the improved Euler method and the modified Euler method.

11.5.1 Consistency and convergence

Definition 11.162. The *one-step error of a one-step method* (11.102) is

$$\mathcal{L}\mathbf{u}(t_n) := \mathbf{u}(t_{n+1}) - \mathbf{u}(t_n) - k\mathbf{\Phi}(\mathbf{u}(t_n), t_n; k). \tag{11.107}$$

Definition 11.163. A one-step method is said to have *order of accuracy* p or to have *pth-order accuracy* or to be *pth-order accurate* if, as $k \to 0$,

$$\mathcal{L}\mathbf{u}(t_n) = \Theta(k^{p+1}). \tag{11.108}$$

Definition 11.164. A one-step method is *consistent* if

$$\lim_{k\to 0} \frac{1}{k}\mathcal{L}\mathbf{u}(t_n) = \mathbf{0}. \tag{11.109}$$

Exercise 11.165. Use recursive Taylor expansions to derive the k^3 term in the one-step error $\mathcal{L}\mathbf{u}(t_n)$ of the explicit midpoint method, verifying $\mathcal{L}\mathbf{u}(t_n) = \Theta(k^3)$, i.e., the explicit midpoint method is second-order accurate.

Theorem 11.166. A one-step method is consistent if and only if

$$\lim_{k\to 0} \mathbf{\Phi}(\mathbf{u}, t; k) = \mathbf{f}(\mathbf{u}, t) \tag{11.110}$$

for any $(\mathbf{u}, t)$ in the domain of $\mathbf{f}$.

Proof. Definition 11.155 and a Taylor expansion of $\mathbf{u}(t_{n+1})$ at t_n yield

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

The proof is completed by taking limit of the above equation in the asymptotic range of $k \to 0$, c.f. Definition 11.164. □

Corollary 11.167. The Euler method is consistent.

Proof. This follows from Theorem 11.166 and the fact that $\mathbf{\Phi}(\mathbf{u}, t; 0) = \mathbf{f}(\mathbf{u}, t)$ for Euler's method. □

Definition 11.168. A one-step method is *convergent* if its solution error tends to zero as $k \to 0$ for any $T > 0$ and for any initial condition $\mathbf{U}^0 = \mathbf{u}(0) + o(1)$, i.e.,

$$\lim_{k\to 0; Nk=T} \mathbf{U}^N = \mathbf{u}(T). \tag{11.111}$$

Lemma 11.169. Let (ξ_n) be a sequence in $\mathbb{R}$ such that

$$|\xi_{n+1}| \le (1 + C)|\xi_n| + D, \quad n \in \mathbb{N} \tag{11.112}$$

for some positive constants C and D. Then we have

$$|\xi_n| \le e^{nC}|\xi_0| + \frac{D}{C}(e^{nC} - 1), \quad n \in \mathbb{N}. \tag{11.113}$$

Proof. The induction basis $n = 0$ clearly holds. Now suppose (11.113) 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 (11.113) holds for $n + 1$ as well. □

Theorem 11.170. Suppose the increment function $\mathbf{\Phi}$ that describes a one-step method is continuous (in $\mathbf{u}$, t, and k) and satisfies a Lipschitz condition

$$\|\mathbf{\Phi}(\mathbf{u}, t; k) - \mathbf{\Phi}(\mathbf{v}, t; k)\| \le M\|\mathbf{u} - \mathbf{v}\| \tag{11.114}$$

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

$$\|\mathbf{E}^n\| \le \frac{K}{M}\left(e^{MT} - 1\right)k^p. \tag{11.115}$$

Proof. For sufficiency, we assume that the one-step method is consistent and compute

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

where the last step follows from the Lipschitz condition (11.114) and $\lim_{k\to 0} c(k) = \lim_{k\to 0} \frac{1}{k}\max\|\mathcal{L}\mathbf{u}(t)\| = 0$. Hence we have

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

Applying Lemma 11.169 with $C = kM$ and $D = kc(k)$ yields

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

which establishes the convergence since $\|\mathbf{E}^0\|$ and $c(k)$ both tend to 0 as $k \to 0$. In particular, (11.115) follows from this inequality and the condition of $c(k) \le Kk^p$.

For necessity, we assume that the one-step method is convergent, i.e., the one-step method (11.102) converges to the solution of

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}, t), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

for all final time $T > 0$. Consider

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

and observe that by Theorem 11.166 the one-step method is consistent with the new IVP

$$\mathbf{u}'(t) = \mathbf{g}(\mathbf{u}, t), \quad \mathbf{u}(0) = \mathbf{u}_0.$$

Since we have already shown that consistency implies convergence, the one-step method also converges to this new IVP. Hence the solutions of the two IVPs coincide and we have $\mathbf{f}(\mathbf{u}(\tau), \tau) = \mathbf{g}(\mathbf{u}(\tau), \tau)$ for all $(\mathbf{u}(\tau), \tau)$ in the domain of $\mathbf{f}$. Then the continuity of $\mathbf{\Phi}$ in k at $k = 0$ implies

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

which implies uniform convergence of $\mathbf{\Phi}(\mathbf{u}, t; k)$ to $\mathbf{f}$, c.f. Definition C.83. Then the proof is completed by Theorem 11.166. □

Corollary 11.171. Both the modified Euler method and the improved Euler method are convergent. If $\mathbf{f}$ in the IVP is twice continuously differentiable, then both methods are second-order accurate.

Proof. The increment function

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

describing the modified Euler method (11.105) clearly satisfies the consistency condition (11.110) and hence by Theorem 11.170, it only remains to verify the Lipschitz condition of $\mathbf{\Phi}$. From the Lipschitz condition for $\mathbf{f}$ we obtain

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

hence $\mathbf{\Phi}$ also satisfies a Lipschitz condition.

If $\mathbf{f}$ is twice continuously differentiable, then by Exercise 11.165, the one-step error of the modified Euler method satisfies

$$\|\mathcal{L}\mathbf{u}(t_n)\| \le Kk^3.$$

Therefore the modified Euler method (11.105) has order of accuracy two by Theorem 11.170.

The same result concerning the improved Euler method (11.104) can be proved in a similar manner. □

11.5.2 Absolute stability

Definition 11.172. The *stability function of a one-step method* is a function $R : \mathbb{C} \to \mathbb{C}$ that satisfies

$$U^{n+1} = R(z)U^n \tag{11.116}$$

for the test problem $u'(t) = \lambda u$ where $\operatorname{Re}\lambda \le 0$ and $z := k\lambda$.

Example 11.173. The trapezoidal rule, when viewed as a one-step method, has its stability function as

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, \tag{11.117}$$

which is also the root of the LMM stability polynomial in Example 11.135. Indeed, the stability function of the one-step method and the LMM stability polynomial have to coincide because they characaterize the same concept from different viewpoints.

Exercise 11.174. Show that the TR-BDF2 method (11.106) has

$$R(z) = \frac{1 + \frac{5}{12}z}{1 - \frac{7}{12}z + \frac{1}{12}z^2}, \tag{11.118}$$

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

Definition 11.175. The *RAS of a one-step method* is a subset of the complex plane

$$\mathcal{R} := \{z \in \mathbb{C} : |R(z)| \le 1\}. \tag{11.119}$$

11.5.3 A-stability and L-stability

Definition 11.176. A one-step method is *A-stable* if its RAS $\mathcal{R}$ satisfies

$$\mathbb{C}^- := \{z \in \mathbb{C} : \operatorname{Re} z \le 0\} \subseteq \mathcal{R}. \tag{11.120}$$

Definition 11.177. A one-step method is *L-stable* if it is A-stable and its stability function satisfies

$$\lim_{z \to \infty} |R(z)| = 0. \tag{11.121}$$

Example 11.178. We use the trapezoidal and backward Euler's methods to solve the IVP (11.95) with $\lambda = -10^6$. By Example 11.151, both methods are A-stable. However, the following table shows different patterns of their errors at $T = 3$ with various values of k and the initial condition $u(0) = \eta$.

	k	Backward Euler	Trapezoidal
$\eta = 1$	0.2	9.7731e-08	4.7229e-10
	0.1	4.9223e-08	1.1772e-10
	0.05	2.4686e-08	2.9406e-11
$\eta = 1.5$	0.2	9.7731e-08	4.9985e-01
	0.1	4.9223e-08	4.9940e-01
	0.05	2.4686e-08	4.9761e-01

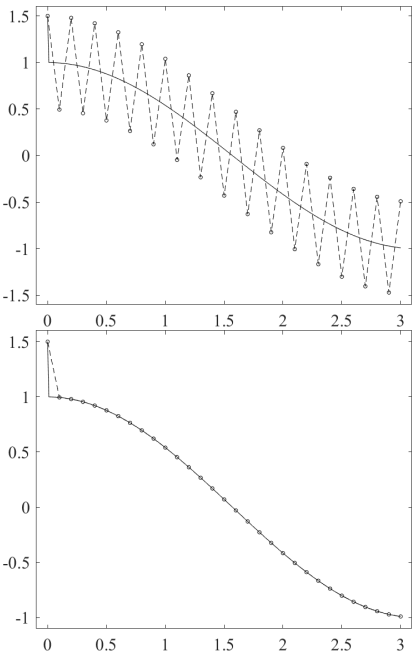
The results are caused by the fact that the backward Euler's method is L-stable while the trapezoidal method is not. More precisely, their stability functions are

$$R(z) = \frac{1}{1-z} \text{ for backward Euler,}$$
$$R(z) = \frac{1+z/2}{1-z/2} \text{ for trapezoidal.}$$

Although the trapezoidal method is stable and the results stay bounded, but since

$$\frac{1+k\lambda/2}{1-k\lambda/2} \approx -1$$

when k is small, the initial deviation from the smooth curve $\cos t$ is essentially negated in each time step. For $k = 0.1$, we have $\frac{1+k\lambda/2}{1-k\lambda/2} = -0.99996 \approx -1$ for trapezoidal method while for backward Euler's method $(1-k\lambda)^{-1} \approx 10^{-5}$. This leads to dramatically different results as below.



Exercise 11.179. Reproduce the results in Example 11.178 and explain in your own language why the first-order backward Euler method is superior to the second-order trapezoidal method.

11.6 Runge-Kutta (RK) methods

Definition 11.180. An *s-stage Runge-Kutta (RK) method* is a one-step method of the form

$$\begin{cases} \mathbf{y}_i = \mathbf{f}(\mathbf{U}^n + k\sum_{j=1}^s a_{i,j}\mathbf{y}_j, t_n + c_i k), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + k\sum_{j=1}^s b_j \mathbf{y}_j, \end{cases} \tag{11.122}$$

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

Definition 11.181. The *Butcher tableau* is one way to organize the coefficients of an RK method as follows.

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \cdots & a_{1,s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & \cdots & a_{s,s} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{11.123}$$

The matrix $A = (a_{i,j})_{i,j=1,\ldots,s}$ is called the *RK matrix* while

$$\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_s \end{bmatrix}^T \text{ and } \mathbf{c} = \begin{bmatrix} c_1 & c_2 & \cdots & c_s \end{bmatrix}^T$$

are called the *RK weights* and the *RK nodes*, respectively.

Exercise 11.182. Write down the Butcher tableaux of the modified Euler method, the improved Euler method, and Heun's third-order method in Definition 11.189.

Exercise 11.183. Verify that the RK method (11.122) can be rewritten as

$$\begin{cases} \boldsymbol{\xi}_i = \mathbf{U}^n + k\sum_{j=1}^s a_{i,j}\mathbf{f}(\boldsymbol{\xi}_j, t_n + c_j k), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + k\sum_{j=1}^s b_j \mathbf{f}(\boldsymbol{\xi}_j, t_n + c_j k), \end{cases} \tag{11.124}$$

where $i = 1, 2, \ldots, s$.

Theorem 11.184. It suffices to only consider autonomous IVPs for the analysis of RK methods if we set

$$c_i = \sum_{j=1}^s a_{i,j}, \quad i = 1, 2, \ldots, s; \tag{11.125}$$

$$\sum_{i=1}^s b_i = 1. \tag{11.126}$$

Proof. A non-autonomous IVP can be converted to an equivalent autonomous IVP. For the IVP

$$\mathbf{u}' = \mathbf{f}(\mathbf{u}, t), \qquad \mathbf{u}(t_0) = \mathbf{u}_0, \tag{11.127}$$

define a new variable

$$\widetilde{\mathbf{u}}(t) = \begin{pmatrix} \mathbf{u}(t) \\ t \end{pmatrix} \in \mathbb{R}^{N+1}$$

and we have

$$\widetilde{\mathbf{u}}' = \widetilde{\mathbf{f}}(\widetilde{\mathbf{u}}), \quad \widetilde{\mathbf{u}}(t_0) = \begin{pmatrix} \mathbf{u}_0 \\ t_0 \end{pmatrix}, \tag{11.128}$$

where

$$\widetilde{\mathbf{f}}(\widetilde{\mathbf{u}}) = \begin{pmatrix} \mathbf{f}(\mathbf{u}, t) \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{f}\left((\widetilde{u}_1, \ldots, \widetilde{u}_N)^T, \widetilde{u}_{N+1}\right) \\ 1 \end{pmatrix},$$

$(\widetilde{u}_1, \ldots, \widetilde{u}_N)^T = \mathbf{u}$, and $\widetilde{u}_{N+1}$ is the time variable in the original formulation. The last equation of (11.128),

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{u}_{N+1} = 1, \quad \widetilde{u}_{N+1}(t_0) = t_0,$$

is solved by $\widetilde{u}_{N+1}(t) = t$; the other equations are equivalent to the original system (11.127). Therefore, if (11.128) has a unique solution, (11.127) will have the same one.

For an RK method (11.122) satisfying (11.125) and (11.126), the solution we obtain for (11.128) is the same as that for (11.127). To see this, consider the ith stage,

$$\mathbf{y}_i = \mathbf{f}\left(\mathbf{U}^n + k\sum_{j=1}^{s} a_{i,j}\mathbf{y}_j, t_n + c_i k\right)$$

for (11.127). For the autonomous system (11.128), we have

$$\begin{aligned}
\widetilde{\mathbf{y}}_i &= \widetilde{\mathbf{f}}\left(\widetilde{\mathbf{U}}^n + k\sum_{j=1}^{s} a_{i,j}\widetilde{\mathbf{y}}_j\right) \\
&= \begin{pmatrix} \mathbf{f}\left(\widetilde{\mathbf{U}}^n_* + k\sum_{j=1}^{s} a_{i,j}\widetilde{\mathbf{y}}_{j,*}, \widetilde{U}^n_{N+1} + k\sum_{j=1}^{s} a_{i,j}\widetilde{y}_{j,N+1}\right) \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{f}\left(\widetilde{\mathbf{U}}^n_* + k\sum_{j=1}^{s} a_{i,j}\widetilde{\mathbf{y}}_{j,*}, \widetilde{U}^n_{N+1} + k\sum_{j=1}^{s} a_{i,j}\right) \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{f}\left(\mathbf{U}^n + k\sum_{j=1}^{s} a_{i,j}\widetilde{\mathbf{y}}_{j,*}, t_n + c_i k\right) \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{y}_i \\ 1 \end{pmatrix},
\end{aligned}$$

where $\widetilde{\mathbf{U}}^n_*$ and $\widetilde{\mathbf{y}}_{j,*}$ respectively denote the first N components of $\widetilde{\mathbf{U}}_n$ and $\widetilde{\mathbf{y}}_j$, the third step follows from the fact that the $(N+1)$th component of each $\widetilde{\mathbf{y}}_i$ is 1, the fourth from $\widetilde{\mathbf{U}}^n_* = \mathbf{U}^n$, $\widetilde{U}^n_{N+1} = t_n$, and (11.125). Finally, (11.126) yields

$$\widetilde{\mathbf{U}}^{n+1}_* = \widetilde{\mathbf{U}}^n_* + k\sum_{j=1}^{s} b_j\widetilde{\mathbf{y}}_{j,*} = \mathbf{U}^n + k\sum_{j=1}^{s} b_j\mathbf{y}_j = \mathbf{U}^{n+1};$$

$$\widetilde{U}^{n+1}_{N+1} = \widetilde{U}^n_{N+1} + k\sum_{j=1}^{s} b_j\widetilde{y}_{j,N+1} = t_n + k\sum_{j=1}^{s} b_j = t_{n+1},$$

which complete the proof. □

11.6.1 Explicit RK (ERK) methods

Definition 11.185. An *s-stage explicit RK (ERK) method* for solving the IVP (11.3) is an RK method of the form

$$\begin{cases} \mathbf{y}_i = \mathbf{f}(\mathbf{U}^n + k\sum_{j=1}^{i-1} a_{i,j}\mathbf{y}_j, t_n + c_i k), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + k\sum_{j=1}^{s} b_j\mathbf{y}_j, \end{cases} \tag{11.129}$$

where $a_{i,j} = 0$ for $i \le j$,

$$\forall i = 1, 2, \ldots, s, \quad c_i = \sum_{j=1}^{i-1} a_{i,j}. \tag{11.130}$$

Example 11.186. The Butcher tableau of an s-stage ERK method is of the form

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a_{2,1} & 0 & & & \\ c_3 & a_{3,1} & a_{3,2} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array}. \tag{11.131}$$

Theorem 11.187. For an ERK method, the condition (11.130) is equivalent to

$$\mathbf{y}_i(\mathbf{u}(t_n), t_n; k) = \mathbf{u}'(t_n + c_i k) + O(k^2), \tag{11.132}$$

for $i = 1, 2, \ldots, s$.

Proof. We prove (11.132) by an induction on i. The induction basis of $i = 1$ holds because

$$\mathbf{y}_1(\mathbf{u}(t_n), t_n; k) = \mathbf{f}(\mathbf{u}(t_n), t_n) = \mathbf{u}'(t_n).$$

Suppose (11.132) holds for each $i = 1, 2, \ldots, m$. Then for $m+1$, we have

$$\begin{aligned}
&\mathbf{y}_{m+1}(\mathbf{u}(t_n), t_n; k) \\
=&\mathbf{f}\left(\mathbf{u}(t_n) + k\sum_{j=1}^{m} a_{m+1,j}\mathbf{y}_j, t_n + c_{m+1}k\right) \\
=&\mathbf{f} + k\left(\mathbf{f}_\mathbf{u}\sum_{j=1}^{m} a_{m+1,j}\mathbf{y}_j + c_{m+1}\mathbf{f}_t\right) + O(k^2) \\
=&\mathbf{f} + k\left(\mathbf{f}_\mathbf{u}\sum_{j=1}^{m} a_{m+1,j}\mathbf{u}'(t_n + c_j k) + c_{m+1}\mathbf{f}_t\right) + O(k^2) \\
=&\mathbf{f} + k\left(\mathbf{f}_\mathbf{u}\mathbf{u}'(t_n)\sum_{j=1}^{m} a_{m+1,j} + c_{m+1}\mathbf{f}_t\right) + O(k^2),
\end{aligned}$$

where all functions are evaluated at $(\mathbf{u}(t_n), t_n)$, the first step follows from the condition $a_{i,j} = 0$ for $i \le j$, the second and fourth steps follow from the Taylor expansion theorem, and the third from the induction hypothesis. By Taylor's expansion, we also have

$$\begin{aligned}
\mathbf{u}'(t_n + c_{m+1}k) &= \mathbf{u}'(t_n) + kc_{m+1}\mathbf{u}''(t_n) + O(k^2) \\
&= \mathbf{f} + kc_{m+1}\left(\mathbf{f}_\mathbf{u}\mathbf{u}'(t_n) + \mathbf{f}_t\right) + O(k^2).
\end{aligned}$$

On one hand, the condition

$$(*): \quad c_{m+1} = \sum_{j=1}^{m} a_{m+1,j}$$

yields (11.132) for $i = m+1$; on the other hand, (11.132) for $i = m+1$ implies (*) since c_{m+1} and $a_{m+1,j}$'s does not depend on k. □

Example 11.188. For the two-stage ERK method

$$\begin{cases} \mathbf{y}_1 = \mathbf{f}(\mathbf{U}^n, t_n), \\ \mathbf{y}_2 = \mathbf{f}(\mathbf{U}^n + ka_{2,1}\mathbf{y}_1, t_n + c_2 k), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + k(b_1\mathbf{y}_1 + b_2\mathbf{y}_2), \end{cases} \tag{11.133}$$

the one-step error is

$$\begin{aligned}
\mathcal{L}\mathbf{u}(t_n) =&\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n) - kb_1\mathbf{f}(\mathbf{u}(t_n), t_n) \\
&- kb_2\mathbf{f}(\mathbf{u}(t_n) + kc_2\mathbf{y}_1, t_n + c_2 k) \\
=&\left(\mathbf{u} + k\mathbf{u}' + \frac{k^2}{2}\mathbf{u}'' + O(k^3)\right) - \mathbf{u} - kb_1\mathbf{u}' \\
&- kb_2\left(\mathbf{f} + kc_2\left(\mathbf{f}_\mathbf{u}\mathbf{f} + \mathbf{f}_t\right) + O(k^2)\right) \\
=&k(1 - b_1 - b_2)\mathbf{u}' + k^2\left(\frac{1}{2} - b_2 c_2\right)\mathbf{u}'' + O(k^3),
\end{aligned}$$

where we have applied (11.130). Hence to maximize the order of accuracy of (11.133), we choose

$$b_1 + b_2 = 1, \quad b_2 c_2 = \frac{1}{2}$$

to get the family of two-stage, second-order ERK methods

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}.$$

The choice of $\alpha = 1$ yields the explicit trapezoidal method (11.104), and that of $\alpha = \frac{1}{2}$ yields the explicit midpoint method (11.105). We have united the two method under one derivation. Note that the above derivation is more general than Exercise 11.165.

Definition 11.189. *Heun's third-order formula* is an ERK method of the form

$$\begin{cases} \mathbf{y}_1 = \mathbf{f}(\mathbf{U}^n, t_n), \\ \mathbf{y}_2 = \mathbf{f}(\mathbf{U}^n + \frac{k}{3}\mathbf{y}_1, t_n + \frac{k}{3}), \\ \mathbf{y}_3 = \mathbf{f}(\mathbf{U}^n + \frac{2k}{3}\mathbf{y}_2, t_n + \frac{2k}{3}), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + \frac{k}{4}(\mathbf{y}_1 + 3\mathbf{y}_3). \end{cases} \tag{11.134}$$

Exercise 11.190. There are three one-parameter families of third-order three-stage ERK methods, one of which is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{2}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{2}{3}-\frac{1}{4\alpha} & \frac{1}{4\alpha} & 0 \\ \hline & \frac{1}{4} & \frac{3}{4}-\alpha & \alpha \end{array},$$

where α is the free parameter. Derive the above family. Does Heun's third-order formula belong to this family?

Definition 11.191. The *classical fourth-order RK method* is an ERK method of the form

$$\begin{cases} \mathbf{y}_1 = \mathbf{f}(\mathbf{U}^n, t_n), \\ \mathbf{y}_2 = \mathbf{f}\left(\mathbf{U}^n + \frac{k}{2}\mathbf{y}_1, t_n + \frac{k}{2}\right), \\ \mathbf{y}_3 = \mathbf{f}\left(\mathbf{U}^n + \frac{k}{2}\mathbf{y}_2, t_n + \frac{k}{2}\right), \\ \mathbf{y}_4 = \mathbf{f}(\mathbf{U}^n + k\mathbf{y}_3, t_n + k), \\ \mathbf{U}^{n+1} = \mathbf{U}^n + \frac{k}{6}(\mathbf{y}_1 + 2\mathbf{y}_2 + 2\mathbf{y}_3 + \mathbf{y}_4). \end{cases} \tag{11.135}$$

Example 11.192. The Butcher tableau of the classical fourth-order RK method (11.135) is

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{11.136}$$

11.6.2 Necessary order conditions

Theorem 11.193 (RK order conditions)**.** To be pth-order accurate, an s-stage RK method (11.122) must satisfy the order conditions

$$\begin{cases} \forall l = 1, 2, \ldots, p, \\ \forall m = 0, 1, \ldots, p-l, \end{cases} \quad \mathbf{b}^T A^m C^{l-1} \mathbf{1} = \frac{(l-1)!}{(l+m)!}, \tag{11.137}$$

where $C = \mathrm{diag}(c_1, c_2, \ldots, c_s)$ and $\mathbf{1} \in \mathbb{Z}^s$ is a column vector with all components equal to one.

Proof. For the scalar IVP

$$u' = u + t^{l-1}, \quad u(0) = 0, \tag{11.138}$$

where $l \in \mathbb{N}^+$, the first step of the RK method yields

$$\forall i = 1, 2, \ldots, s, \quad y_i = k \sum_{j=1}^{s} a_{i,j} y_j + (c_i k)^{l-1},$$

which can be rewritten in matrix form as

$$(I - kA)\mathbf{y} = k^{l-1} C^{l-1} \mathbf{1},$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_s)^T$. It follows that

$$\mathbf{y} = k^{l-1}(I - kA)^{-1} C^{l-1} \mathbf{1}$$

and we have $U^1 = U^0 + k\mathbf{b}^T\mathbf{y} = k^l \mathbf{b}^T (I - kA)^{-1} C^{l-1} \mathbf{1}$, i.e.,

$$U^1 = k^l \sum_{m=0}^{+\infty} k^m \left(\mathbf{b}^T A^m C^{l-1} \mathbf{1}\right). \tag{11.139}$$

By Theorem 11.51, the exact solution of (11.138) is

$$u(t) = \int_0^t e^{t-\tau} \tau^{l-1} \mathrm{d}\tau,$$

which, together with the chain rule or the Leibniz formula, implies

$$\begin{cases} u(0) = \cdots = u^{(l-1)}(0) = 0, \\ \forall j \geq 0, \ u^{(l+j)}(0) = (l-1)!. \end{cases}$$

The Taylor expansion of $u(t_1)$ at $t_0 = 0$ yields

$$u(t_1) = k^l \sum_{m=0}^{+\infty} k^m \frac{(l-1)!}{(l+m)!}. \tag{11.140}$$

Then the proof is completed by Definition 11.163, which implies that, in order for the RK method to be pth-order accurate, the series of U^1 and $u(t_1)$ in (11.139) and (11.140) must agree for the first $p - l + 1$ terms for each $l = 1, 2, \ldots, p$. □

Definition 11.194 (Pure quadrature order conditions)**.** An s-stage RK method is said to be $B(r)$ if it satisfies

$$\forall l = 1, 2, \ldots, r, \quad \mathbf{b}^T C^{l-1} \mathbf{1} = \sum_{j=1}^{s} b_j c_j^{l-1} = \frac{1}{l}. \tag{11.141}$$

Definition 11.195. The *quadrature formula of an s-stage RK method* is the weighted quadrature formula

$$I_s(f) := k \sum_{j=1}^{s} b_j f(t_n + c_j k) \tag{11.142}$$

that approximates $I(f) := \int_{t_n}^{t_n+k} f(t)\mathrm{d}t$.

Exercise 11.196. Show that the quadrature formula of a RK method is exact for all polynomials f of degree $< r$, i.e.,

$$\forall f \in \mathbb{P}_{r-1}, \quad I_s(f) = \int_{t_n}^{t_n+k} f(t)\mathrm{d}t, \tag{11.143}$$

if and only if the RK method is $B(r)$.

Example 11.197. Exercise 11.196 justifies the name "pure quadrature order conditions" since (11.141) is the special case of $m = 0$ in (11.137) that corresponds to the RHS of an IVP depending only on time t.

If $m \neq 0$, the order conditions in (11.137) can be put as

$$\sum_{i,j_1,\ldots,j_m} b_i a_{i,j_1} a_{j_1,j_2} \cdots a_{j_{m-1},j_m} c_{j_m}^{l-1} = \frac{(l-1)!}{(l+m)!}.$$

For $l = 1$, the order conditions in (11.137) reduce to

$$\forall m = 0, 1, \ldots, p-1, \quad \mathbf{b}^T A^m \mathbf{1} = \frac{1}{(m+1)!}. \tag{11.144}$$

In particular, $m = 0$ and $l = 1$ yields the condition (11.126) discussed in Theorem 11.184.

Corollary 11.198. An s-stage ERK method is at best sth-order accurate.

Proof. Since the RK matrix A of an s-stage ERK method is strictly lower triangular, A is nilpotent with $A^s = \mathbf{0}$. Suppose for the IVP (11.138) the order of accuracy of the ERK method is at least $s + 1$. Then Definition 11.163, (11.139) and (11.140) yield

$$\mathcal{L}u(t_0) = u(t_1) - U^1 = o(k^{s+1}),$$

which is impossible because the coefficient of k^{s+1} of U^1 is $\mathbf{b}^T A^s \mathbf{1} = 0$ whereas that of $u(t_1)$ is $\frac{1}{(s+1)!} \neq 0$. □

Example 11.199. The classical fourth-order RK method (11.135) can be obtained as follows.

(a) Simpson's rule (Definition 6.14) for quadrature is fourth-order accurate:

$$\int_0^1 f(t)\mathrm{d}t \approx \frac{1}{6} f(0) + \frac{2}{3} f\left(\frac{1}{2}\right) + \frac{1}{6} f(1).$$

(b) Simpson's rule has only three abscissas: 0, $\frac{1}{2}$, and 1. By Corollary 11.198, a fourth-order ERK method has at least four stages, so tentatively we settle for four stages with the abscissas as $c_i = 0, \frac{1}{2}, \frac{1}{2}, 1$.

(c) Simpson's rule indicates the choice $b_1 = b_4 = \frac{1}{6}$. Set $l = 1$ in (11.141) and we have (11.126), i.e. $\sum_j b_j = 1$. A simple choice is then $b_2 = b_3 = \frac{1}{3}$.

(d) Set $m = 3$ in (11.144) and we have

$$\mathbf{b}^T A^3 \mathbf{1} = b_4\, a_{43}\, a_{32}\, a_{21} = \frac{1}{24},$$

which implies $a_{i+1,i} \neq 0$.

(e) Choose $a_{i+1,i}$'s to be the only nonzero elements of A and deduce $a_{i+1,i} = c_{i+1}$ for each $i = 1, 2, 3$ from (11.130).

The Butcher tableau (11.136) of the classical fourth-order RK method is now fully determined.

Theorem 11.200. ERK methods have the following limits on the attainable order of accuracy as a function of the number of stages:

number of stages	1	2	3	4	5	6	7	8	9	10
max order of accuracy	1	2	3	4	4	5	6	6	7	7

.

11.6.3 Implicit RK (IRK) methods

Definition 11.201. An *implicit RK (IRK) method* is an RK method with at least one $a_{i,j} \neq 0$ for $i \leq j$. A *diagonal IRK (DIRK) method* is an IRK method with $a_{i,j} = 0$ whenever $i < j$. A *singly DIRK (SDIRK) method* is a DIRK method with $a_{1,1} = a_{2,2} = \cdots = a_{s,s} = \gamma \neq 0$.

Example 11.202. The TR-BDF2 method can be regarded as a second-order DIRK method since it can be expressed in the standard form of an IRK method:

$$\begin{cases} y_1 = \mathbf{f}(U^n, t_n) \\ y_2 = \mathbf{f}(U^*, t_n + \frac{k}{2}) \text{ where } U^* = U^n + \frac{k}{4}(y_1 + y_2), \\ y_3 = \mathbf{f}(U^n + \frac{k}{3}(y_1 + y_2 + y_3), t_n + k) \\ U^{n+1} = U^n + \frac{k}{3}(y_1 + y_2 + y_3), \end{cases}$$

which yields the Butcher tableau

$$\begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}.$$

Example 11.203. It can be shown that, under the condition (11.125), a two-stage RK method (11.122) is third-order accurate iff it satisfies the order conditions (11.137) with $p = 3$. This fact is utilized to construct two-stage third-order IRK methods.

(11.125) and the order conditions in (11.137) yield

$$i = 1, 2: \quad c_i = \sum_{j=1}^{2} a_{i,j}, \tag{11.145a}$$

$$l = 1, 2, 3: \quad \sum_{i=1}^{2} b_i c_i^{l-1} = \frac{1}{l}, \tag{11.145b}$$

$$\sum_{i=1}^{2} \sum_{j=1}^{2} b_i a_{i,j} c_j = \frac{1}{6}, \tag{11.145c}$$

where (11.145a) is the same as (11.125) and the four equations in (11.145b,c) come from the six equations in (11.137) with $p = 3$. The eight equations in (11.125) with $s = 2$ and (11.137) with $p = 3$ are dependent in the sense that some equations can be deduced from others. For example, the cases $l = 1, m = 2$ and $l = 2, m = 1$ are the same, i.e.,

$$\mathbf{b}^T A^2 \mathbf{1} = \mathbf{b}^T A\,(A\mathbf{1}) = \mathbf{b}^T A\mathbf{c} = \mathbf{b}^T AC\mathbf{1}.$$

By Definition 11.194 and Theorem 11.193, (11.145b) implies that the method is $B(3)$ and thus could have third-order accuracy. By Exercise 11.196, the quadrature formula of this

RK method is exact for all polynomials in $\mathbb{P}_2$. By Theorem 6.24, the node polynomial is orthorgonal to $\mathbb{P}_0$, i.e.,

$$\int_0^1 (x-c_1)(x-c_2)\mathrm{d}x = 0,$$

which yields

$$c_2 = \frac{2-3c_1}{3-6c_1} \quad \left(c_1 \neq \frac{1}{2}\right).$$

By Lemma 6.8, we determine the weights from the nodes as

$$b_1 = \int_0^1 \frac{c_2 - x}{c_2 - c_1}\mathrm{d}x = \frac{c_2 - \frac{1}{2}}{c_2 - c_1},$$
$$b_2 = \int_0^1 \frac{c_1 - x}{c_1 - c_2}\mathrm{d}x = \frac{c_1 - \frac{1}{2}}{c_1 - c_2}.$$

In (11.145c), we insert (11.145a) and consider $a_{1,2}$ and c_1 as free parameters. This gives

$$a_{2,2} = \frac{\frac{1}{6} - b_1 a_{1,2}(c_2 - c_1) - \frac{1}{2}c_1}{b_2(c_2 - c_1)}.$$

For $a_{1,2} = 0$ we get a one-parameter family of DIRK methods of order 3. If we further require $a_{1,1} = a_{2,2} = \gamma$, then we obtain two two-stage third-order SDIRK methods with Butcher tableau given by

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}, \quad \gamma = \frac{3 \pm \sqrt{3}}{6}. \tag{11.146}$$

Definition 11.204. An *explicit DIRK (EDIRK) method* is a DIRK method with an explicit first stage, i.e., $a_{1,1} = 0$. An *explicit SDIRK (ESDIRK) method* is an EDIRK method with $a_{2,2} = a_{3,3} = \cdots = a_{s,s} = \gamma \neq 0$.

Example 11.205. A commonly used ESDIRK method is the following six-stage fourth-order ESDIRK method

$$\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{83}{250} & \frac{8611}{62500} & \frac{-1743}{31250} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{31}{50} & \frac{5012029}{34652500} & \frac{-654441}{2922500} & \frac{174375}{388108} & \frac{1}{4} & 0 & 0 \\ \frac{17}{20} & \frac{15267082809}{155376265600} & \frac{-71443401}{120774400} & \frac{730878875}{902184768} & \frac{2285395}{8070912} & \frac{1}{4} & 0 \\ 1 & \frac{82889}{524892} & 0 & \frac{15625}{83664} & \frac{69875}{102672} & \frac{-2260}{8211} & \frac{1}{4} \\ \hline & \frac{82889}{524892} & 0 & \frac{15625}{83664} & \frac{69875}{102672} & \frac{-2260}{8211} & \frac{1}{4} \end{array}. \tag{11.147}$$

Theorem 11.206. The maximal attainable order of an s-stage DIRK method cannot exceed $s+1$.

11.6.4 Collocation methods

Definition 11.207. A numerical method for solving the IVP (11.3) is said to have the *dense output* property iff, in addition to the numerical results $U^{n+1} \approx \mathbf{u}(t_n + k)$, it provides cheap numerical approximations to $\mathbf{u}(t_n + \theta k)$ for the entire interval of the time step $\theta \in [0, 1]$. Here "cheap" means without or with only a few additional function evaluations.

Example 11.208. For fourth-order dense output, we can fit a cubic polynomial from the RK results $\mathbf{U}^n$, $\mathbf{f}(\mathbf{U}^n, t_n)$, $\mathbf{U}^{n+1}$, and $\mathbf{f}(\mathbf{U}^{n+1}, t_{n+1})$ by solving Hermite interpolation problems as in Definition 2.33.

Definition 11.209. An s-stage *collocation method* is a numerical method for solving the IVP (11.3), where we

(i) choose s distinct *collocation parameters* $c_1, c_2, \ldots, c_s$ (typically in $[0, 1]$),

(ii) seek s-degree *collocation polynomials* $\mathbf{p}$ satisfying

$$\begin{aligned} \mathbf{p}(t_n) &= \mathbf{U}^n, \\ \mathbf{p}'(t_n + c_i k) &= \mathbf{f}(\mathbf{p}(t_n + c_i k), t_n + c_i k), \end{aligned} \tag{11.148}$$

for each $i = 1, 2, \ldots, s$,

(iii) set

$$\mathbf{U}^{n+1} = \mathbf{p}(t_{n+1}). \tag{11.149}$$

Corollary 11.210. A collocation method has dense output.

Proof. This follows from Definitions 11.209 and 11.207: we get the dense-output property as a byproduct of polynomial fitting. □

Theorem 11.211. The collocation method in Definition 11.209 is an s-stage IRK method with

$$a_{i,j} = \int_0^{c_i} \ell_j(\tau)\mathrm{d}\tau, \tag{11.150a}$$
$$b_j = \int_0^1 \ell_j(\tau)\mathrm{d}\tau, \tag{11.150b}$$

where $i, j = 1, 2, \ldots, s$ and $\ell_j(\tau)$ is the elementary Lagrange interpolation polynomial, i.e.,

$$\ell_j(\tau) = \prod_{k \neq j; k=1}^{s} \frac{\tau - c_k}{c_j - c_k}. \tag{11.151}$$

Proof. The Lagrange interpolation polynomial

$$\mathbf{r}(t_n + \tau k) := \sum_{j=1}^{s} \mathbf{p}'(t_n + c_j k)\ell_j(\tau)$$

satisfies $\mathbf{r}(t_n + c_i k) = \mathbf{p}'(t_n + c_i k)$ for each $i = 1, 2, \ldots, s$. The coincidence of the polynomials $\mathbf{r}$ and $\mathbf{p}'$ at s points implies $\mathbf{r} \equiv \mathbf{p}'$ since their degrees are both $s-1$. Then (11.148) yields

$$\begin{aligned} \mathbf{p}'(t_n + \tau k) &= \textstyle\sum_{j=1}^{s} \mathbf{p}'(t_n + c_j k)\ell_j(\tau) \\ &= \textstyle\sum_{j=1}^{s} \mathbf{f}(\mathbf{p}(t_n + c_j k), t_n + c_j k)\ell_j(\tau). \end{aligned}$$

Integrate over $[t_n, t_n + ck]$, apply the fundamental theorem of calculus, notice $\mathbf{p}(t_n) = \mathbf{U}^n$, and we obtain

$$\begin{aligned} & \mathbf{p}(t_n + ck) = \mathbf{p}(t_n) + k\textstyle\int_0^c \mathbf{p}'(t_n + \tau k)\mathrm{d}\tau \\ = \; & \mathbf{U}^n + k\textstyle\int_0^c \sum_{j=1}^{s} \mathbf{f}(\mathbf{p}(t_n + c_j k), t_n + c_j k)\ell_j(\tau)\mathrm{d}\tau \\ = \; & \mathbf{U}^n + k\textstyle\sum_{j=1}^{s} \mathbf{f}(\mathbf{p}(t_n + c_j k), t_n + c_j k)\int_0^c \ell_j(\tau)\mathrm{d}\tau, \end{aligned}$$

which can be rewritten as

$$\mathbf{p}(t_n + ck) = \mathbf{U}^n + k\sum_{j=1}^{s} \mathbf{y}_j \int_0^c \ell_j(\tau)\mathrm{d}\tau \tag{11.152}$$

if we set, for each $i = 1, 2, \ldots, s$,

$$\mathbf{y}_i := \mathbf{f}(\mathbf{p}(t_n + c_i k), t_n + c_i k). \tag{11.153}$$

Choose $c = c_i$ in (11.152), substitute the resulting equation into (11.153), apply (11.150a), and we have

$$\mathbf{y}_i = \mathbf{f}\left(\mathbf{U}^n + k\sum_{j=1}^{s} a_{i,j}\mathbf{y}_j, t_n + c_i k\right).$$

Choose $c = 1$ in (11.152), apply (11.150b), and we have

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\sum_{j=1}^{s} b_j \mathbf{y}_j.$$

The above processes recover the formalism in (11.122) of an RK method. In general, we have $a_{i,j} \neq 0$ for $i \leq j$. Thus we conclude that the collocation method is an IRK method. □

Exercise 11.212. Show that an s-stage collocation method is at least s-order accurate.

Exercise 11.213. Prove that the collocation method viewed as an RK method satisfies (11.125) and (11.126).

Example 11.214. Not every IRK method originates in collocation. Let $s = 2, c_1 = 0, c_2 = \frac{2}{3}$, then the corresponding elementary Lagrange interpolation polynomials are

$$\ell_1(\tau) = \frac{\tau - \frac{2}{3}}{-\frac{2}{3}} = \frac{2 - 3\tau}{2}, \quad \ell_2(\tau) = \frac{3\tau}{2},$$

and (11.150) yields the IRK method with Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}.$$

By Definition 11.209, a collocation method is fully determined once the RK nodes c_i's are given. Therefore, we deduce from the above Butcher tableau that the IRK method

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}$$

is not a collocation method.

Exercise 11.215. Derive the three-stage IRK method that corresponds to the collocation points $c_1 = \frac{1}{4}, c_2 = \frac{1}{2}, c_3 = \frac{3}{4}$.

Definition 11.216. An s-stage RK method (11.122) is said to be $C(r)$ if

$$\begin{cases} \forall i = 1, 2, \ldots, s, \\ \forall m = 1, 2, \ldots, r, \end{cases} \quad \sum_{j=1}^{s} a_{i,j} c_j^{m-1} = \frac{c_i^m}{m}, \tag{11.154}$$

it is said to be $D(r)$ if

$$\begin{cases} \forall i = 1, 2, \ldots, s, \\ \forall m = 1, 2, \ldots, r, \end{cases} \quad \sum_{j=1}^{s} b_j c_j^{m-1} a_{j,i} = \frac{b_i}{m}(1 - c_i^m). \tag{11.155}$$

Theorem 11.217. An s-stage IRK method with all its RK nodes distinct and with its order of accuracy $p \geq s$ is an s-stage collocation method if and only if $C(s)$ is true.

Proof. The necessity holds because, for each $i = 1, 2, \ldots, s$,

$$\begin{aligned} \textstyle\sum_{j=1}^{s} a_{i,j} c_j^{m-1} &= \textstyle\sum_{j=1}^{s} c_j^{m-1} \int_0^{c_i} \ell_j(\tau)\mathrm{d}\tau \\ &= \textstyle\int_0^{c_i} \left(\sum_{j=1}^{s} c_j^{m-1} \ell_j(\tau)\right) \mathrm{d}\tau \\ &= \textstyle\int_0^{c_i} \tau^{m-1}\mathrm{d}\tau = \frac{c_i^m}{m}, \end{aligned}$$

where the first equality follows from (11.150a) and the third from interpolating a polynomial of degree $m - 1$ with the Lagrange formula, thanks to the fact of c_i's being distinct.

To prove the sufficiency, we rewrite the $C(s)$ condition

$$\forall m = 1, 2, \ldots, s, \qquad \sum_{j=1}^{s} c_j^{m-1} a_{i,j} = \frac{c_i^m}{m}$$

as a Vandermonde linear system

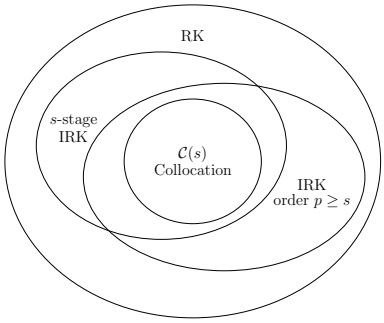
$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1 & c_2 & \cdots & c_s \\ \vdots & \vdots & \ddots & \vdots \\ c_1^{s-1} & c_2^{s-1} & \cdots & c_s^{s-1} \end{pmatrix} \begin{pmatrix} a_{i,1} \\ a_{i,2} \\ \vdots \\ a_{i,s} \end{pmatrix} = \begin{pmatrix} c_i \\ \frac{c_i^2}{2} \\ \vdots \\ \frac{c_i^s}{s} \end{pmatrix}$$

for each $i = 1, 2, \ldots, s$. By Lemma 2.4 and the distinctness of c_i's, each Vandermonde linear system has a unique solution, which can be readily verified to be given by (11.150a).

Since the IRK method has order at least s, the pure quadrature order conditions in (11.141) hold:

$$\forall m = 1, 2, \ldots, s, \quad \sum_{j=1}^{s} c_j^{m-1} b_j = \frac{1}{m};$$

this can also be considered as a Vandermonde linear system, and its solution is given by b_i's in (11.150b). The proof is then completed by Theorem 11.211. □



Exercise 11.218. Show $B(s + r)$ and $C(s)$ imply $D(r)$ via multiplying the two vectors $u_j := \sum_{i=1}^{s} b_i c_i^{m-1} a_{i,j}$ and $v_j := \frac{1}{m} b_j (1 - c_j^m)$ by the Vandermonde matrix $V(c_1, c_2, \ldots, c_s)$ in Definition 2.3.

Lemma 11.219. If an RK method satisfies $B(p)$, $C(\eta)$, and $D(\zeta)$ with $p \le 2\eta + 2$, and $p \le \zeta + \eta + 1$, then the method has order of accuracy at least p.

Proof. See [Hairer et al., 1993, p. 208]. □

Lemma 11.220. For each $r = 1, 2, \dots, s$, let $q_r \in \mathbb{P}_s$ be the polynomial that satisfies

$$\begin{cases} \forall p \in \mathbb{P}_{r-1}, & \langle q_r, p\rangle := \int_{t_n}^{t_n+k} q_r(x)p(x)\mathrm{d}x = 0, \\ & \langle q_r, t^r\rangle \neq 0. \end{cases} \tag{11.156}$$

If we design a weighted quadrature formula

$$I_s(f) := \sum_{j=1}^{s} b_j f(c_j) \tag{11.157}$$

by selecting c_j's as the zeros of the polynomial q_r and determine the weights b_j's by the Vandermonde linear system

$$\forall m = 0, 1, \dots, s-1, \quad \sum_{j=1}^{s} c_j^m b_j = \int_{t_n}^{t_n+k} \tau^m \mathrm{d}\tau,$$

then the degree of exactness for $I_s(f)$ is $s + r - 1$, i.e.

$$\begin{cases} \forall f \in \mathbb{P}_{s+r-1}, & E_s(f) = \int_{t_n}^{t_n+k} f(t)\mathrm{d}t - I_s(f) = 0, \\ \exists g \in \mathbb{P}_{s+r}, \text{ s.t.} & E_s(g) \neq 0, \end{cases} \tag{11.158}$$

c.f. Definition 6.6.

Proof. The weighted formula (11.157) has degree of exactness at least $s - 1$ since for any pairwise distinct nodes $c_1, \dots, c_s$ we can interpolate f with a polynomial and derive a Newton-Cotes formula to approximate $\int_{t_n}^{t_n+k} f(t)\mathrm{d}t$. Then the first statement of (11.158) follows from Theorem 6.24 with the weight function $\rho(x) = 1$. The proof is completed by setting $g = qt^r \in \mathbb{P}_{s+r}$ since

$$\begin{aligned} E_s(g) &= \int_{t_n}^{t_n+k} qt^r \mathrm{d}t - I_s(qt^r) \\ &= \langle q, t^r\rangle - \textstyle\sum_{j=1}^{s} b_j q(c_j) c_j^r = \langle q, t^r\rangle \neq 0, \end{aligned}$$

where the third step follows from c_j's being roots of q. □

Theorem 11.221. An s-stage collocation method with its nodes c_j and weights b_j determined as those in (11.157) is $(s + r)$th-order accurate.

Proof. By Lemma 11.219, it suffices to show that the collocation method is $B(s + r)$, $C(s)$ and $D(r)$.

It follows from Lemma 11.220 that the choices of c_j and b_j imply that the degree of exactness of the quadrature formula of the collocation method is $s+r-1$. Then by Exercise 11.196 the collocation method is $B(r + s)$.

By (11.156) and Definition 6.29, we can construct for the weight function $\rho \equiv 1$ a set of orthogonal polynomials, one for each degree $0, 1, \dots, r$, with q as the one with degree r.

By Theorem 6.31, the zeros of q are simple, hence the collocation parameters c_i's are pairwise distinct. By Theorem 11.217 and Exercise 11.212, the method is $C(s)$.

The proof is then completed by Exercise 11.218. □

Exercise 11.222. Determine the order of accuracy of the collocation method derived in Exercise 11.215.

Corollary 11.223. A collocation method with its RK nodes as zeros of q_s in (11.156) is $(2s)$th-order accurate.

Proof. This follows from Theorem 11.221 with $r = s$. □

Definition 11.224. A *Gauss-Legendre RK method* is an s-stage, $(2s)$th-order collocation method.

Lemma 11.225. The monic shifted Legendre polynomials, i.e., orthogonal polynomials with respect to the weight function $\rho(t) \equiv 1, 0 < t < 1$, are

$$\widetilde{P}_s(t) = \frac{(s!)^2}{(2s)!} \sum_{j=0}^{s} (-1)^{s-j} \binom{s}{j} \binom{s+j}{j} t^j. \tag{11.159}$$

Example 11.226. By Lemma 11.225, we have

$$\widetilde{P}_1(t) = t - \frac{1}{2}$$

and hence $c_1 = \frac{1}{2}$. Then (11.150) yields the 1-stage second-order Gauss-Legendre RK method as

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}. \tag{11.160}$$

Example 11.227. By Lemma 11.225, we have

$$\widetilde{P}_2(t) = t^2 - t + \frac{1}{6}$$

and hence $c_1 = \frac{3-\sqrt{3}}{6}$ and $c_2 = \frac{3+\sqrt{3}}{6}$. Then (11.150) yields the 2-stage fourth-order Gauss-Legendre RK method as

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{11.161}$$

Example 11.228. By Lemma 11.225, we have

$$\widetilde{P}_3(t) = t^3 - \frac{3}{2}t^2 + \frac{3}{5}t - \frac{1}{20}$$

and hence

$$c_1 = \frac{5 - \sqrt{15}}{10}, \quad c_2 = \frac{1}{2}, \quad c_3 = \frac{5 + \sqrt{15}}{10}.$$

Then (11.150) yields the 3-stage sixth-order Gauss-Legendre RK method as

$$\begin{array}{c|ccc} \frac{5-\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{5+\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}. \tag{11.162}$$

11.6.5 Practical error estimation and step size control

Definition 11.229. The *Richardson extrapolation* is an algorithm to construct a more accurate solution from the increment function $\mathbf{\Phi}$ of a pth-order one-step method (11.102).

(RET-1) Advance $\mathbf{u}_0$ by one step of size k:

$$\mathbf{U}^1 = \mathbf{u}_0 + k\mathbf{\Phi}(\mathbf{u}_0, t_0; k).$$

(RET-2) Advance $\mathbf{u}_0$ by two consecutive steps of size $\frac{k}{2}$:

$$\begin{cases} \mathbf{U}^{\frac{1}{2}} &= \mathbf{u}_0 + \frac{k}{2}\mathbf{\Phi}\left(\mathbf{u}_0, t_0; \frac{k}{2}\right), \\ \widetilde{\mathbf{U}}^1 &= \mathbf{U}^{\frac{1}{2}} + \frac{k}{2}\mathbf{\Phi}\left(\mathbf{U}^{\frac{1}{2}}, t_0 + \frac{k}{2}; \frac{k}{2}\right). \end{cases}$$

(RET-3) Set the more accurate solution to be

$$\widehat{\mathbf{U}}^1 = \widetilde{\mathbf{U}}^1 + \frac{1}{2^p - 1}\left(\widetilde{\mathbf{U}}^1 - \mathbf{U}^1\right). \tag{11.163}$$

Lemma 11.230. The error of the solution $\widehat{\mathbf{U}}^1$ given by the Richardson extrapolation for a consistent pth-order one-step method is

$$\widehat{\mathbf{E}}^1 = \widehat{\mathbf{U}}^1 - \mathbf{u}(t_0 + k) = O(k^{p+2}). \tag{11.164}$$

In addition, the error of the pth-order one-step method is

$$\mathbf{E}^1 = \mathbf{U}^1 - \mathbf{u}(t_0 + k) = \frac{2^p}{1 - 2^p}\left(\widetilde{\mathbf{U}}^1 - \mathbf{U}^1\right) + O(k^{p+2}). \tag{11.165}$$

Proof. The pth-order accuracy of the one-step method gives

$$\mathbf{E}^1 = \boldsymbol{\tau}(\mathbf{u}(t_0), t_0)k^{p+1} + O(k^{p+2}). \tag{11.166}$$

Similarly, we have

$$\mathbf{E}^{\frac{1}{2}} = \mathbf{U}^{\frac{1}{2}} - \mathbf{u}\left(t_0 + \frac{k}{2}\right) = \boldsymbol{\tau}(\mathbf{u}(t_0), t_0)\left(\frac{k}{2}\right)^{p+1} + O(k^{p+2}).$$

By (RET-2), the error $\widetilde{\mathbf{E}}^1 := \widetilde{\mathbf{U}}^1 - \mathbf{u}(t_0 + k)$ includes $\mathbf{E}^{\frac{1}{2}}$ and the local error of the second substep:

$$\begin{aligned} \widetilde{\mathbf{E}}^1 &= \mathbf{E}^{\frac{1}{2}} + \boldsymbol{\tau}\left(\mathbf{u}(t_0 + \tfrac{k}{2}), t_0 + \tfrac{k}{2}\right)\left(\tfrac{k}{2}\right)^{p+1} + O(k^{p+2}) \\ &= 2\boldsymbol{\tau}(\mathbf{u}(t_0), t_0)\left(\tfrac{k}{2}\right)^{p+1} + O(k^{p+2}), \end{aligned} \tag{11.167}$$

where the second step follows from the Taylor expansion of $\boldsymbol{\tau}\left(\mathbf{u}(t_0 + \frac{k}{2}), t_0 + \frac{k}{2}\right)$ at $(\mathbf{u}(t_0), t_0)$. Then (11.163), (11.166), and (11.167) yield

$$\widehat{\mathbf{E}}^1 = \frac{2^p}{2^p - 1}\widetilde{\mathbf{E}}^1 - \frac{1}{2^p - 1}\mathbf{E}^1 = O(k^{p+2}).$$

Similarly, (11.165) follows from (11.166) and (11.167). □

Definition 11.231. An *embedded RK method* is a pair of RK methods of orders p and $\hat{p}$ that share the same stage computations:

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array} \quad \text{and} \quad \begin{array}{c|c} \mathbf{c} & A \\ \hline & \widehat{\mathbf{b}}^T \end{array},$$

which, due to the same RK matrix A and the same RK node vector $\mathbf{c}$, often has the notation $p(\hat{p})$ and

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \\ \hline & \widehat{\mathbf{b}}^T \end{array}. \tag{11.168}$$

Example 11.232. The *Fehlberg 4(5) embedded RK method* has six stages and delivers a method of order 4 with an error estimate (or a method of order 5 without):

$$\begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{4} & \frac{1}{4} & & & & & \\ \frac{3}{8} & \frac{3}{32} & \frac{9}{32} & & & & \\ \frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & & & \\ 1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & & \\ \frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & \\ \hline & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\ \hline & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55} \end{array}. \tag{11.169}$$

The somewhat unintuitive coefficients in (11.169) arise not only from satisfying the order conditions but also from an attempt to minimize the one-step error.

Example 11.233. The best known general purpose RK method is the *Dormand-Prince 5(4) embedded RK method*

$$\begin{array}{c|ccccccc} 0 & & & & & & & \\ \frac{1}{5} & \frac{1}{5} & & & & & & \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\ \frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\ \frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\ 1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\ 1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\ \hline & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \hline & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \end{array} \tag{11.170}$$

which has seven stages and is a local extrapolation method, i.e., the 5th-order method is used to advance the solution. (11.170) is also optimized to make the coefficients in the leading term in the one-step error as small as possible.

(11.170) has the *first-same-as-last (FSAL)* property, i.e., $c_7 = 1$ and $\mathbf{b}^T$ is the same as the 7th row of A. When a step is accepted, $\mathbf{f}$ doesn't have to be re-evaluated at the start of the next step, and thus only six $\mathbf{f}$ evaluations are needed per step.

Formula 11.234. Given $\mathbf{U}^n$, k, $\mathbf{E}_{\text{abs}}$, and $\mathbf{E}_{\text{rel}}$ as the initial value, the initial step size, the absolute error tolerance, and the relative error tolerance, respectively, a $p(\hat{p})$ embedded RK method with *automatic step size control* advances the IVP (11.3) from $\mathbf{U}^n$ to $\mathbf{U}^{n+1}$ as follows.

(a) Advance $\mathbf{U}^n$ to $\left(\mathbf{U}^{n+1}, \widehat{\mathbf{U}}^{n+1}\right)$ by the two methods in the embedded RK method.

(b) Compute an error indicator based on the max-norm or the 2-norm

$$E_{\text{ind}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{U}_i^{n+1} - U_i^{n+1}}{\varepsilon_i}\right)^2}, \tag{11.171}$$

where N is the size of the IVP (11.3) and

$$\varepsilon_i = E_{\text{abs},i} + |U_i^n| E_{\text{rel},i}. \tag{11.172}$$

(c) Compute a new time step size $\tilde{k}$ by

$$\tilde{k} = k \min\left\{\rho_{\max}, \max\left\{\rho_{\min}, \rho\left(E_{\text{ind}}\right)^{\frac{-1}{q+1}}\right\}\right\}, \tag{11.173}$$

where $q = \min(p, \hat{p})$, usually $\rho_{\max} \in [1.5, 5]$, $\rho = 0.8, 0.9, (0.25)^{\frac{1}{q+1}}$ or $(0.38)^{\frac{1}{q+1}}$, and $\rho_{\min} \in [0.2, \rho]$.

(d) If $E_{\text{ind}} \le 1$, proceed to the next time step with $\mathbf{U}^{n+1}$, $\tilde{k}$, $\mathbf{E}_{\text{abs}}$, and $\mathbf{E}_{\text{rel}}$; otherwise reject $\left(\mathbf{U}^{n+1}, \widehat{\mathbf{U}}^{n+1}\right)$ and go back to step (a) with $\tilde{k}$ as the new initial step size.

Example 11.235. For the case $\hat{p} = p+1$ in Formula 11.234, we expect

$$\left\|\widehat{\mathbf{U}}^{n+1} - \mathbf{U}^{n+1}\right\| \approx Ck^{p+1}.$$

If we have $E_{\text{ind}} \le 1$, i.e.,

$$\left\|\widehat{\mathbf{U}}^{n+1} - \mathbf{U}^{n+1}\right\| \le \|\boldsymbol{\epsilon}\|,$$

we will increase the step size; otherwise we decrease it. The extent of increasing or decreasing is quantified as

$$\left(\frac{\tilde{k}}{k}\right)^{p+1} \approx \frac{\|\boldsymbol{\epsilon}\|}{\left\|\widehat{\mathbf{U}}^{n+1} - \mathbf{U}^{n+1}\right\|}$$

and the new step size can be obtained by one calculation

$$\tilde{k} = k \cdot \left(E_{\text{ind}}\right)^{\frac{-1}{q+1}}. \tag{11.174}$$

However, in practical computation k should not be allowed to increase nor to decrease too fast; (11.173) is designed to fulfill this purpose. As a price we pay for (11.173), the whole process in Formula 11.234 may be repeated multiple times until an acceptable step size is found. The values of the factor ρ are chosen such that the new time step size will be acceptable in the next time step with high probability. If $E_{\text{ind}} \le 1$, the computed step is accepted and the solution is advanced with $\mathbf{U}^{n+1}$ and a new step is tried with $\tilde{k}$ as the step size; otherwise, the step is rejected and the computations are repeated with the new step size $\tilde{k}$.

11.6.6 Consistency and convergence

Theorem 11.236. Consider an IVP (11.3) where the function $\mathbf{f} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ is continuous in t and Lipschitz continuous in $\mathbf{u}$ with Lipschitz constant L. The RK method (11.122) has a unique solution if its time step size k satisfies

$$k < \frac{1}{L\|A\|_\infty}, \tag{11.175}$$

where A is the RK matrix. Furthermore, $\mathbf{f}(\mathbf{u}, t) \in \mathcal{C}^p$ implies $\mathbf{y}_i(k) \in \mathcal{C}^p$.

Proof. Define $\mathbf{Y} := (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_s)^T$ and the first s equations in (11.122) state that $\mathbf{Y}$ is the fixed point of $\mathbf{f}$:

$$\mathbf{Y} = \mathbf{f}(\mathbf{Y}), \tag{11.176}$$

where, for each $i = 1, 2, \ldots, s$,

$$\mathbf{y}_i = \mathbf{f}\left(\mathbf{U}^n + k\sum_{j=1}^{s} a_{i,j}\mathbf{y}_j,\ t_n + c_i k\right).$$

For the norm

$$\|\mathbf{Y}\| = \max_{1\le i\le s} \|\mathbf{y}_i\|,$$

we have

$$\begin{aligned} & \|\mathbf{f}(\mathbf{y}_i) - \mathbf{f}(\mathbf{z}_i)\| \le kL\left\|\textstyle\sum_{j=1}^{s} a_{i,j}\left(\mathbf{y}_j - \mathbf{z}_j\right)\right\| \\ \le\ & kL\textstyle\sum_{j=1}^{s}|a_{i,j}|\,\|\mathbf{y}_j - \mathbf{z}_j\| \le kL\sum_{j=1}^{s}|a_{i,j}|\,\|\mathbf{Y} - \mathbf{Z}\| \\ \le\ & kL\|A\|_\infty\|\mathbf{Y} - \mathbf{Z}\|, \end{aligned}$$

where the first step follows from the Lipschitz condition of $\mathbf{f}$ and the second from the triangle inequality. Take the maximum over $i = 1, 2, \ldots, s$ and we have

$$\|\mathbf{f}(\mathbf{Y}) - \mathbf{f}(\mathbf{Z})\| \le kL\|A\|_\infty\|\mathbf{Y} - \mathbf{Z}\|.$$

Then (11.175) implies that $\mathbf{f}$ is a contractive mapping. The vector version of the contractive mapping theorem 1.38 ensures the existence and uniqueness of the solution.

For the last statement, we define $\mathbf{F} : \mathbb{R} \times \mathbb{R}^{sN} \to \mathbb{R}^{sN}$ as

$$\mathbf{F}(k, \mathbf{Y}) := \mathbf{Y} - \mathbf{f}(\mathbf{Y}),$$

and rewrite (11.176) as $\mathbf{F}(k, \mathbf{Y}) = 0$. The Jacobian matrix $\frac{\partial \mathbf{F}}{\partial \mathbf{Y}}(0, \mathbf{Y}_0)$ is the identity matrix for any $\mathbf{Y}_0$ and thus invertible, because $\mathbf{F}(0, \mathbf{Y}) = \mathbf{Y} - \mathbf{f}(\mathbf{U}^n, t_n)$. The proof is then completed by the implicit function theorem C.103. □

Corollary 11.237. Consider an IVP (11.3) where the function $\mathbf{f} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ is continuous in t and Lipschitz continuous in $\mathbf{u}$ with Lipschitz constant L. The existence and uniqueness of the collocation polynomial $\mathbf{p}(t)$ in (11.148) are guaranteed if the time step size k satisfies

$$k < \frac{1}{L\|A\|_\infty},$$

where A is the RK matrix of the collocation method in Definition 11.209.

Proof. This follows from Theorems 11.211 and 11.236. □

Lemma 11.238. The increment function $\boldsymbol{\Phi}$ of an RK method with order of accuracy $p \ge 1$ satisfies the Lipschitz condition (11.114) with

$$M = \frac{L\|\mathbf{b}\|_1}{1 - kL\|A\|_\infty} \tag{11.177}$$

if its time step size k satisfies (11.175).

Proof. The increment function $\mathbf{\Phi}$ that describes an RK method (11.122) is

$$\begin{aligned}\mathbf{\Phi} &= \textstyle\sum_{i=1}^s b_i \mathbf{y}_i,\\ \mathbf{y}_i(\mathbf{u},t;k) &= \mathbf{f}\left(\mathbf{u} + k\textstyle\sum_{j=1}^s a_{i,j}\mathbf{y}_j(\mathbf{u},t;k),\ t+c_ik\right).\end{aligned}$$

The Lipschitz continuity of $\mathbf{f}$ yields

$$\begin{aligned}&\|\mathbf{y}_i(\mathbf{u},t;k) - \mathbf{y}_i(\mathbf{v},t;k)\|\\ \le& L\|\mathbf{u}-\mathbf{v}\| + kL\left\|\sum_{j=1}^s a_{i,j}\left(\mathbf{y}_j(\mathbf{u},t;k) - \mathbf{y}_j(\mathbf{v},t;k)\right)\right\|\\ \le& L\|\mathbf{u}-\mathbf{v}\| + kL\sum_{j=1}^s |a_{i,j}| \max_{1\le\ell\le s}\|\mathbf{y}_\ell(\mathbf{u},t;k) - \mathbf{y}_\ell(\mathbf{v},t;k)\|\\ \le& L\|\mathbf{u}-\mathbf{v}\| + kL\|A\|_\infty \max_{1\le\ell\le s}\|\mathbf{y}_\ell(\mathbf{u},t;k) - \mathbf{y}_\ell(\mathbf{v},t;k)\|\end{aligned}$$

and taking maximum over $1 \le i \le s$ gives

$$\max_{1\le i\le s}\|\mathbf{y}_i(\mathbf{u},t;k) - \mathbf{y}_i(\mathbf{v},t;k)\| \le \frac{L}{1-kL\|A\|_\infty}\|\mathbf{u}-\mathbf{v}\|.$$

The proof is then completed by

$$\begin{aligned}&\|\mathbf{\Phi}(\mathbf{u},t;k) - \mathbf{\Phi}(\mathbf{v},t;k)\|\\ =\ & \left\|\textstyle\sum_{j=1}^s b_j\left(\mathbf{y}_j(\mathbf{u},t;k) - \mathbf{y}_j(\mathbf{v},t;k)\right)\right\|\\ \le\ & \textstyle\sum_{j=1}^s |b_j| \max_{1\le\ell\le s}\|\mathbf{y}_\ell(\mathbf{u},t;k) - \mathbf{y}_\ell(\mathbf{v},t;k)\|\\ \le\ & M\|\mathbf{u}-\mathbf{v}\|. \quad \square\end{aligned}$$

Theorem 11.239. An RK method with order of accuracy $p \ge 1$ and with initial error being $O(k^{p+1})$ is pth-order convergent for any IVP (11.3) with $\mathbf{f} \in \mathcal{C}^p$.

Proof. This follows directly from Lemma 11.238 and Theorem 11.170. □

Lemma 11.240. The one-step error of the classical fourth-order RK method (11.135) is

$$\mathcal{L}\mathbf{u}(t_n) = O(k^5). \tag{11.178}$$

Exercise 11.241. Prove Lemma 11.240 in the scalar case.

Corollary 11.242. The classical fourth-order RK method (11.135) is convergent. Furthermore, it is fourth-order convergent for any IVP (11.3) with $\mathbf{f} \in \mathcal{C}^p$.

Proof. This follows directly from Lemma 11.240 and Theorem 11.239. □

11.6.7 Absolute stability

Lemma 11.243. When applied to the scalar ODE

$$u' = \lambda u(t), \quad \operatorname{Re}\lambda \le 0,$$

the s-stage RK method (11.122) satisfies

$$U^{n+1} = \left[1 + k\lambda\mathbf{b}^T(I-k\lambda A)^{-1}\mathbf{1}\right]U^n, \tag{11.179}$$

where the matrix $I - k\lambda A$ is assumed to be nonsingular and $\mathbf{1} \in \mathbb{Z}^s$ is a column vector with all components equal to one.

Proof. Applying the RK method (11.122) to the linear ODE $u' = \lambda u$, we obtain

$$y_i = \lambda\left(U^n + k\sum_{j=1}^s a_{i,j}y_j\right), \quad i = 1,2,\dots,s.$$

Denote

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_s\end{bmatrix}^T,$$

then $\mathbf{y} = \lambda U^n\mathbf{1} + k\lambda A\mathbf{y}$ and the exact solution of this linear algebraic system is

$$\mathbf{y} = \lambda U^n\left(I - k\lambda A\right)^{-1}\mathbf{1}.$$

Therefore

$$\begin{aligned}U^{n+1} &= U^n + k\sum_{j=1}^s b_jy_j = U^n + k\mathbf{b}^T\mathbf{y}\\ &= \left[1 + k\lambda\mathbf{b}^T(I-k\lambda A)^{-1}\mathbf{1}\right]U^n. \quad \square\end{aligned}$$

Notation 11. Denote the set of all rational functions by

$$\mathbb{P}_{m/n} := \left\{\frac{P(z)}{Q(z)} :\ P(z)\in\mathbb{P}_m, Q(z)\in\mathbb{P}_n\right\}. \tag{11.180}$$

Lemma 11.244. When applied to the scalar IVP

$$u'(t) = \lambda u(t), \quad u(t_0) = u_0,$$

the s-stage RK method (11.122) yields

$$U^n = [R(k\lambda)]^n U^0, \tag{11.181}$$

where $R \in \mathbb{P}_{s/s}$. If the RK method is explicit, then $R \in \mathbb{P}_s$.

Proof. It follows from (11.179) that (11.181) holds with

$$R(z) = 1 + z\mathbf{b}^T(I-zA)^{-1}\mathbf{1}, \quad z\in\mathbb{C},$$

and it remains to verify $R \in \mathbb{P}_{s/s}$.

We represent the inverse of $I - zA$ using Cramer's rule,

$$(I-zA)^{-1} = \frac{\operatorname{adj}(I-zA)}{\det(I-zA)},$$

where $\operatorname{adj} C$ is the adjugate of the $s\times s$ matrix C: the (i,j)th entry of the adjugate is the determinant of the (i,j)th principal minor, multiplied by $(-1)^{i+j}$, c.f. Definition B.231. Since each entry of $I - zA$ is linear in z, we deduce that each element of $\operatorname{adj}(I-zA)$, being (up to a sign) a determinant of an $(s-1)\times(s-1)$ matrix, is in $\mathbb{P}_{s-1}$. We thus conclude that

$$\mathbf{b}^T\operatorname{adj}(I-zA)\mathbf{1} \in \mathbb{P}_{s-1},$$

therefore $\det(I-zA) \in \mathbb{P}_s$ implies $R \in \mathbb{P}_{s/s}$.

Finally, if the method is explicit, then A is strictly lower triangular and $I - zA$ is, regardless of $z \in \mathbb{C}$, a lower triangular matrix with ones along the diagonal. Therefore $\det(I-zA) \equiv 1$ and R is a polynomial. □

Corollary 11.245. The stability function of an s-stage RK method (11.122) is a rational function $R(z) \in \mathbb{P}_{s/s}$ given by

$$R(z) = 1 + z\mathbf{b}^T(I-zA)^{-1}\mathbf{1}, \quad z\in\mathbb{C}. \tag{11.182}$$

Proof. This follows immediately from Definition 11.172 and Lemmas 11.243 and 11.244. □

Example 11.246. The stability function of the modified Euler method (11.105) is

$$\begin{aligned} R(z) &= 1 + z\mathbf{b}^T(I - zA)^{-1}\mathbf{1} \\ &= 1 + z\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\frac{z}{2} & 1 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= 1 + z + \tfrac{z^2}{2}. \end{aligned}$$

Exercise 11.247. Show that the classical fourth-order RK method has its stability function as

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4. \tag{11.183}$$

Theorem 11.248. If the RK method (11.122) is at least pth-order accurate, then, as $z \to 0$,

$$R(z) = e^z + O(z^{p+1}). \tag{11.184}$$

Proof. For the scalar IVP $u' = \lambda u$, the one-step error of a pth- or higher-order RK method satisfies

$$O(k^{p+1}) = \mathcal{L}u(t_n) = u(t_{n+1}) - U^{n+1} = \left(e^{\lambda k} - R(\lambda k)\right) e^{\lambda t_n},$$

where the first equality follows from Definition 11.163, the second from Definitions 11.162 and 11.180, and the third from Definition 11.172. The proof is completed by dividing both sides of the above equation by $e^{\lambda t_n}$. □

Corollary 11.249. All s-stage, sth-order ERK methods have the same stability function

$$R(z) = \sum_{j=0}^{s} \frac{1}{j!} z^j, \quad z \in \mathbb{C}. \tag{11.185}$$

Proof. By Lemma 11.244, $R(z)$ is a polynomial of degree at most s. Theorem 11.248 states that as $z \to 0$,

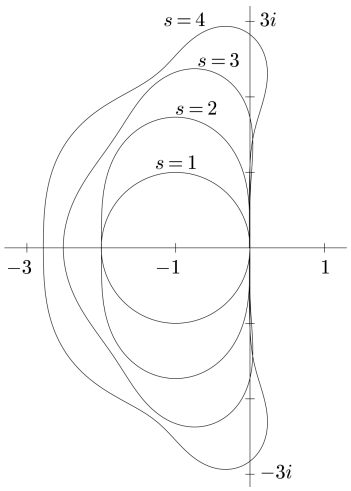
$$R(z) = e^z + O(z^{s+1}),$$

which implies that $R(z)$ is a polynomial of degree at least s. These two conditions uniquely determine $R(z)$. □

Corollary 11.250. All s-stage, sth-order ERK methods share the same RAS.

Proof. This follows immediately from Definition 11.175 and Corollary 11.249. □

Example 11.251. For each ERK method with $s = 1, 2, 3, 4$, its RAS boundary is shown below.



Exercise 11.252. Define $S_s := \{z : |R_s(z)| \le 1\}$ where $s = 1, 2, 3, 4$ and R_s is the stability function of the s-stage, sth-order ERK method. Show that

$$S_1 \subset S_2 \subset S_3. \tag{11.186}$$

Does this hold for ERK methods with a higher stage? Why?

11.6.8 I-stability and L-stability

Theorem 11.253. No ERK method is A-stable.

Proof. By Lemma 11.244, the stability function $R(z)$ of an ERK method is a polynomial, and hence

$$\lim_{z \to -\infty} |R(z)| = \infty,$$

which excludes A-stability. □

Definition 11.254. An RK method is *I-stable* iff its stability function $R(z)$ satisfies

$$\forall y \in \mathbb{R},\ |R(y\mathbf{i})| \le 1. \tag{11.187}$$

Theorem 11.255 (Maximum modulus principle)**.** Consider an analytic function $f : D \to \mathbb{C}$ over a path-connected nonempty open subset $D \subseteq \mathbb{C}$. If the modulus function $|f| : D \to \mathbb{R}^+ \cup \{0\}$ reaches its maximum on D, then f is a constant.

Proof. See [Freitag and Busam, 2009, page 129]. □

Theorem 11.256. An RK method is A-stable if and only if it is I-stable and all poles of its stability function $R(z)$ have positive real parts.

Proof. The necessity follows directly from Definitions 11.176 and 11.254. As for the sufficiency, Corollary 11.245 yields

$$R(z) = \frac{P(z)}{Q(z)}, \quad P(z) = \sum_{i=0}^{m} p_i z^i, \quad Q(z) = \sum_{i=0}^{n} q_i z^i,$$

where $p_m \neq 0, q_n \neq 0$. Then Definition 11.254 implies $m < n$ or $m = n$, the latter of which yields $|p_m| \le |q_n|$. Therefore the limit $\lim_{z\to\infty, z\in\mathbb{C}} |R(z)|$ exists and

$$\lim_{z \to \infty, z \in \mathbb{C}} |R(z)| \le 1. \tag{11.188}$$

Since the poles of $R(z)$ have positive real parts, the rational function $R(z)$ is analytic in

$$\Omega_\rho := \{z \in \mathbb{C} : \operatorname{Re} z < 0 \text{ and } |z| < \rho\}$$

for any $\rho > 0$. Suppose the RK method were not A-stable. Then Definitions 11.176 and 11.254 would imply

$$\exists! Z \subset \mathbb{C}^-, Z \neq \emptyset \text{ s.t. } \forall z_0 \in Z,\ \operatorname{Re} z_0 < 0 \text{ and } |R(z_0)| > 1.$$

By (11.188) and the continuity of the function $z \mapsto |R(z)|$, there exists ρ_0 sufficiently large such that $Z \subset \Omega_{\rho_0}$ and

$$\forall z \in \partial\Omega_{\rho_0}, \quad |R(z)| \leq |R(z_m)|,$$

where $z_m \in Z$ is a point that maximizes $|R(z)|$; the existence of z_m is guaranteed by the continuity of $z \mapsto |R(z)|$ and the closure of Ω_{ρ_0} being compact. Therefore, the non-constant function $z \mapsto |R(z)|$ attains its maximum in Ω_ρ, which contradicts Theorem 11.255. □

Example 11.257. The two-stage collocation method with $c_1 = \frac{1}{3}, c_2 = 1$ has its Butcher tableau as

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}.$$

This method is A-stable since it satisfies the two conditions in Theorem 11.256. Its stability function is

$$R(z) = 1 + z\mathbf{b}^T(I - zA)^{-1}\mathbf{1} = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}.$$

(i) I-stability (11.187) is equivalent to

$$\left|1 + \mathbf{i}\frac{1}{3}y\right|^2 \leq \left|1 - \frac{1}{6}y^2 - \mathbf{i}\frac{2}{3}y\right|^2, \quad \forall y \in \mathbb{R},$$

and hence to

$$1 + \frac{1}{9}y^2 \leq 1 + \frac{1}{9}y^2 + \frac{1}{36}y^4, \quad \forall y \in \mathbb{R}.$$

(ii) The poles of $R(z)$ are $2 \pm \mathbf{i}\sqrt{2}$, both reside at the open right half-plane.

This method is also L-stable, either by Exercise 11.258 or by Theorem 11.260.

Exercise 11.258. Prove that an A-stable RK method with its stability function as a rational polynomial $R(z) = \frac{P(z)}{Q(z)}$ is L-stable if and only if $\deg Q(z) > \deg P(z)$.

Definition 11.259. An RK method (11.122) is called *stiffly accurate* if the last row of the RK matrix A is the same as the RK weight $\mathbf{b}^T$, i.e.,

$$a_{s,j} = b_j, \quad j = 1, \ldots, s. \tag{11.189}$$

Theorem 11.260. If an A-stable RK method with a non-singular RK matrix A is stiffly accurate, then it is L-stable.

Proof. By Definition 11.172 of the stability function $R(z)$ and the continuity of the matrix inverse function, we have

$$\begin{aligned} \lim_{z\to\infty} R(z) &= \lim_{z\to\infty} \left(1 + z\mathbf{b}^T(I - zA)^{-1}\mathbf{1}\right) \\ &= 1 + \lim_{z\to\infty} \mathbf{b}^T \left(\tfrac{1}{z}I - A\right)^{-1} \mathbf{1} \\ &= 1 - \mathbf{b}^T A^{-1}\mathbf{1} \\ &= 1 - \mathbf{e}_s^T\mathbf{1} = 0, \end{aligned}$$

where $\mathbf{e}_s = (0, \ldots, 0, 1)^T$ and the fourth equality holds because $A^T\mathbf{e}_s = \mathbf{b}$ by (11.189). □

Exercise 11.261. Show that if an A-stable RK method with a nonsingular RK matrix A satisfies

$$a_{i,1} = b_1, \quad i = 1, \ldots, s, \tag{11.190}$$

then it is L-stable.

Example 11.262. The trapezoidal method (11.31) with Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

is stiffly accurate. But Example 11.178 shows that it is not L-stable, which is caused by the fact that its RK matrix is singular. Similarly, the ESDIRK method in Example 11.205 is stiffly accurate, but is not L-stable.

Example 11.263. The 5th-order RK method embedded in the Dormand-Prince pair (11.170) is stiffly accurate. However, Theorem 11.253 dictates that it can never be A-stable, and thus cannot be L-stable either.

Example 11.264. The collocation method in Example 11.257 is L-stable.

Example 11.265. The 3-stage 6th-order Gauss-Legendre RK method (11.162) is A-stable, but not L-stable since

$$\begin{aligned} \lim_{z\to\infty} R(z) &= \lim_{z\to\infty} \left(1 + z\mathbf{b}^T(I - zA)^{-1}\mathbf{1}\right) \\ &= 1 + \lim_{z\to\infty} \mathbf{b}^T \left(\tfrac{1}{z}I - A\right)^{-1} \mathbf{1} \\ &= 1 - \mathbf{b}^T A^{-1}\mathbf{1} = -1, \end{aligned}$$

where the first step follows from Corollary 11.245, the third from the continuity of the matrix inverse function, and the last from (11.162).

Exercise 11.266. Show that the collocation method

$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

is 5th-order accurate and L-stable.

11.6.9 Contractivity and B-stability

Definition 11.267. A function $\mathbf{f} : \mathbb{R}^N \times [0, +\infty) \to \mathbb{R}^N$ is said to satisfy a *one-sided Lipschitz condition* if

$$\forall t \geq 0, \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^N, \quad \langle \mathbf{u} - \mathbf{v},\ \mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t) \rangle \leq \mu \|\mathbf{u} - \mathbf{v}\|^2, \tag{11.191}$$

where $\langle \cdot, \cdot \rangle$ is an inner product in $\mathbb{R}^N$, $\|\cdot\|$ the corresponding norm, and μ the *one-sided Lipschitz constant* of $\mathbf{f}$.

Lemma 11.268. If $\mathbf{f} : \mathbb{R}^N \times [0, +\infty) \to \mathbb{R}^N$ is Lipschitz continuous with Lipschitz constant L, then it satisfies the one-sided Lipschitz condition (11.191) with $\mu = L$.

Proof. The Cauchy-Schwarz inequality and the Lipschitz continuity of $\mathbf{f}$ yield

$$\begin{aligned} & \langle \mathbf{u} - \mathbf{v}, \mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t) \rangle \\ \leq\ & \|\mathbf{u} - \mathbf{v}\| \cdot \|\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t)\| \leq L \|\mathbf{u} - \mathbf{v}\|^2. \end{aligned}$$

□

Lemma 11.269. If $\mathbf{f} : \mathbb{R}^N \times [0, +\infty) \to \mathbb{R}^N$ is continuous and satisfies a one-sided Lipschitz condition (11.191), then any two solutions $\mathbf{u}(t)$ and $\mathbf{v}(t)$ of the ODE (11.1) satisfy

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq e^{\mu t} \|\mathbf{u}(0) - \mathbf{v}(0)\|, \tag{11.192}$$

Proof. Let $\phi(t) = \|\mathbf{u}(t) - \mathbf{v}(t)\|^2$, then

$$\begin{aligned} \phi'(t) &= \tfrac{\mathrm{d}}{\mathrm{d}t} \langle \mathbf{u}(t) - \mathbf{v}(t), \mathbf{u}(t) - \mathbf{v}(t) \rangle \\ &= 2 \langle \mathbf{u}(t) - \mathbf{v}(t), \mathbf{u}'(t) - \mathbf{v}'(t) \rangle \\ &= 2 \langle \mathbf{u}(t) - \mathbf{v}(t), \mathbf{f}(\mathbf{u}(t), t) - \mathbf{f}(\mathbf{v}(t), t) \rangle \leq 2\mu\phi(t), \end{aligned}$$

where the second step follows from the product rule and the symmetry of an inner product, the third from the ODE, and the last from (11.191). Therefore, we have

$$\phi(t) \leq e^{2\mu t} \phi(0),$$

the square root of which completes the proof. □

Definition 11.270. An ODE system (11.1) is *contractive* or *monotone* if the RHS function $\mathbf{f}$ satisfies the one-sided Lipschitz condition (11.191) with $\mu = 0$, i.e.,

$$\forall t \geq 0, \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^N, \quad \langle \mathbf{u} - \mathbf{v}, \mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t) \rangle \leq 0. \tag{11.193}$$

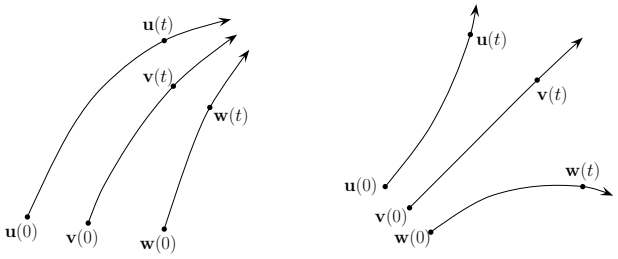
Example 11.271. The linear constant-coefficient system

$$\mathbf{u}'(t) = A\mathbf{u}(t) + \mathbf{b}(t), \quad A \in \mathbb{R}^{N \times N}$$

is contractive if and only if the matrix A is negative semi-definite; see Definition 8.73.

Lemma 11.272. A contractive ODE system is *dissipative*: for any two solutions $\mathbf{u}(t)$ and $\mathbf{v}(t)$ of the ODE (11.1), the norm $\|\mathbf{u}(t) - \mathbf{v}(t)\|$ decreases monotonically for $t \geq 0$.

Proof. Following the same procedure as that in the proof of Lemma 11.269, we can show $\phi'(t) \leq 0$, and therefore $\|\mathbf{u}(t) - \mathbf{v}(t)\| = \sqrt{\phi(t)}$ is monotonically decreasing. □



Definition 11.273 (Butcher 1975)**.** A one-step method is *B-stable* (or *contractive*) if, for any contractive ODE system, every pair of its numerical solutions $\mathbf{U}^n$ and $\mathbf{V}^n$ satisfy

$$\forall n = 0, 1, \ldots, \quad \|\mathbf{U}^{n+1} - \mathbf{V}^{n+1}\| \leq \|\mathbf{U}^n - \mathbf{V}^n\|. \tag{11.194}$$

Example 11.274. For the backward Euler method (11.30), consider $\mathbf{e}^n := \mathbf{U}^n - \mathbf{V}^n$ for an arbitrary pair of numerical solutions $\mathbf{U}^n$ and $\mathbf{V}^n$. The Cauchy-Schwarz inequality and the contractivity in Definition 11.270 yield

$$\begin{aligned} & \|\mathbf{e}^{n+1}\|^2 = \left\langle \mathbf{e}^{n+1}, \mathbf{e}^{n+1} \right\rangle \\ =\ & \left\langle \mathbf{e}^n + k \left[\mathbf{f}(\mathbf{U}^{n+1}, t_{n+1}) - \mathbf{f}(\mathbf{V}^{n+1}, t_{n+1}) \right], \mathbf{e}^{n+1} \right\rangle \\ \leq\ & \|\mathbf{e}^n\| \cdot \|\mathbf{e}^{n+1}\|, \end{aligned}$$

which implies $\|\mathbf{U}^{n+1} - \mathbf{V}^{n+1}\| \leq \|\mathbf{U}^n - \mathbf{V}^n\|$ and thus the backward Euler method is B-stable.

Exercise 11.275. Rewrite the implicit midpoint method

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

in the standard form (11.122) and derive its Butcher tableau. Show that it is B-stable.

Example 11.276. The trapezoidal method (11.31) is not B-stable. To show this, consider a scalar autonomous ODE with its RHS function given by

$$f(u) = \begin{cases} -u & \text{if } u \geq 0, \\ -\frac{u}{\epsilon} & \text{if } u < 0, \end{cases}$$

where $\epsilon \in (0, \frac{1}{5})$ is a constant. It is easy to verify that f satisfies the contractivity condition (11.193). But if we apply the trapezoidal method to the ODE with initial conditions $U^0 = 0$ and $V^0 = -\epsilon$ and time step size $k = 1$, then

$$U^1 = 0, \quad V^1 = \frac{1 - 2\epsilon}{3},$$

and so

$$|U^1 - V^1| > |U^0 - V^0|,$$

which violates (11.194).

Theorem 11.277. An s-stage Gauss-Legendre RK method is B-stable for all $s \geq 1$.

Proof. Let $\{\mathbf{U}^n\}$ and $\{\mathbf{V}^n\}$ be a pair of numerical solutions obtained by the Gauss-Legendre RK methods and denote by $\mathbf{p}_\mathbf{u}(t)$ and $\mathbf{p}_\mathbf{v}(t)$ the collocation polynomials for $\mathbf{U}^n$ and $\mathbf{V}^n$ in (11.148). Let $\phi(t) = \|\mathbf{p}_\mathbf{u}(t) - \mathbf{p}_\mathbf{v}(t)\|^2$, then at the collocation points $\xi_i = t_n + c_i k$ we have

$$\begin{aligned} & \phi'(\xi_i) = 2 \langle \mathbf{p}_\mathbf{u}(\xi_i) - \mathbf{p}_\mathbf{v}(\xi_i), \mathbf{p}'_\mathbf{u}(\xi_i) - \mathbf{p}'_\mathbf{v}(\xi_i) \rangle \\ =\ & 2 \langle \mathbf{p}_\mathbf{u}(\xi_i) - \mathbf{p}_\mathbf{v}(\xi_i), \mathbf{f}(\mathbf{p}_\mathbf{u}(\xi_i), \xi_i) - \mathbf{f}(\mathbf{p}_\mathbf{v}(\xi_i), \xi_i) \rangle \leq 0, \end{aligned}$$

where the second step follows from the definition (11.148) of the collocation method and the third from the contractivity condition (11.193). Therefore

$$\begin{aligned} & \|\mathbf{U}^{n+1}-\mathbf{V}^{n+1}\|^2=\phi(t_n+k)=\phi(t_n)+\textstyle\int_{t_n}^{t_n+k}\phi'(t)\mathrm{d}t \\ = \;& \phi(t_n)+k\textstyle\sum_{j=1}^{s}b_j\phi'(t_n+c_jk)\le\phi(t_n)=\|\mathbf{U}^n-\mathbf{V}^n\|^2, \end{aligned}$$

where the second equality follows from the fundamental theorem of calculus, the third from Definition 11.224 that Gaussian quadrature integrates the polynomial $\phi'(t)$ (which is of degree $2s-1$) exactly, and the fourth from the quadrature weights b_j being positive, c.f. Lemma 6.33. □

Theorem 11.278. B-stable one-step methods are A-stable.

Proof. It suffices to show that, for any contractive complex-valued scalar ODE

$$y'=\lambda y,\quad \lambda\in\mathbb{C}\text{ and }\operatorname{Re}\lambda\le 0, \tag{11.195}$$

any B-stable method with any time step size $k>0$ has a stability function $R:\mathbb{C}\to\mathbb{C}$ that satisfies $|R(k\lambda)|\le 1$. After all, the conditions on λ and k ensure that the left half complex plane is covered by the set $\{k\lambda\}$. Let

$$y(t)=y_1(t)+\mathbf{i}y_2(t),\quad \lambda=\alpha+\mathbf{i}\beta.$$

Then (11.195) is equivalent to the real-valued ODE system

$$\begin{pmatrix}y_1'\\ y_2'\end{pmatrix}=\begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix} \tag{11.196}$$

with $\alpha=\operatorname{Re}\lambda\le 0$. By Definition 8.73, the constant-coefficient matrix in (11.196) is negative semi-definite, and hence (11.196) is contractive. Let $U^n=U_1^n+\mathbf{i}U_2^n$ be a numerical solution of (11.195), then $\mathbf{U}^n=(U_1^n,U_2^n)^T$ is a numerical solution of (11.196). Taking $\mathbf{V}^n\equiv\mathbf{0}$ in Definition 11.273 yields

$$\forall k>0,\forall n=0,1,\ldots,\quad \left\|\mathbf{U}^{n+1}\right\|\le\|\mathbf{U}^n\|.$$

The fact $|U^n|^2=|U_1^n|^2+|U_2^n|^2=\|\mathbf{U}^n\|^2$ implies

$$\forall k>0,\forall n=0,1,\ldots,\quad \left|U^{n+1}\right|\le|U^n|.$$

Then the proof is completed by Definitions 11.172, 11.175, and 11.176. □

11.6.10 Algebraic stability

Definition 11.279. An RK method is *algebraically stable* iff

(ABS-1) the RK weights $b_1,b_2,\ldots,b_s$ are nonnegative,

(ABS-2) the symmetric matrix $M\in\mathbb{R}^{s\times s}$ with

$$m_{i,j}=b_ia_{i,j}+b_ja_{j,i}-b_ib_j \tag{11.197}$$

is positive semidefinite.

The matrix M is called the *algebraic stability matrix*.

Theorem 11.280. An algebraically stable RK method is B-stable and A-stable.

Proof. Apply the s-stage RK method (11.122) to the ODE (11.1) for two initial conditions $\mathbf{U}(0)$, $\mathbf{V}(0)$, and we have

$$\begin{aligned} \boldsymbol{\xi}_i &= \mathbf{f}\left(\mathbf{U}^n+k\textstyle\sum_{j=1}^{s}a_{i,j}\boldsymbol{\xi}_j,t_n+c_ik\right),\\ \mathbf{U}^{n+1} &= \mathbf{U}^n+k\textstyle\sum_{j=1}^{s}b_j\boldsymbol{\xi}_j;\\ \boldsymbol{\eta}_i &= \mathbf{f}\left(\mathbf{V}^n+k\textstyle\sum_{j=1}^{s}a_{i,j}\boldsymbol{\eta}_j,t_n+c_ik\right),\\ \mathbf{V}^{n+1} &= \mathbf{V}^n+k\textstyle\sum_{j=1}^{s}b_j\boldsymbol{\eta}_j, \end{aligned}$$

where $i=1,2,\ldots,s$. Define

$$\mathbf{e}^n=\mathbf{U}^n-\mathbf{V}^n,\quad \mathbf{d}_i=\boldsymbol{\xi}_i-\boldsymbol{\eta}_i$$

and we have

$$\begin{aligned} & \|\mathbf{e}^{n+1}\|^2=\left\langle\mathbf{U}^{n+1}-\mathbf{V}^{n+1},\ \mathbf{U}^{n+1}-\mathbf{V}^{n+1}\right\rangle\\ &=\left\langle\mathbf{e}^n+k\textstyle\sum_{i=1}^{s}b_i\mathbf{d}_i,\mathbf{e}^n+k\sum_{j=1}^{s}b_j\mathbf{d}_j\right\rangle\\ &=\|\mathbf{e}^n\|^2+2k\left\langle\mathbf{e}^n,\textstyle\sum_{i=1}^{s}b_i\mathbf{d}_i\right\rangle+k^2\sum_{i=1}^{s}\sum_{j=1}^{s}b_ib_j\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle. \end{aligned}$$

For B-stability it remains to show

$$2\left\langle\mathbf{e}^n,\sum_{i=1}^{s}b_i\mathbf{d}_i\right\rangle+k\sum_{i=1}^{s}\sum_{j=1}^{s}b_ib_j\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle\le 0.$$

Define

$$\boldsymbol{\rho}_i=\mathbf{U}^n+k\sum_{j=1}^{s}a_{i,j}\boldsymbol{\xi}_j,\quad \boldsymbol{\sigma}_i=\mathbf{V}^n+k\sum_{j=1}^{s}a_{i,j}\boldsymbol{\eta}_j$$

and we have

$$\begin{aligned} & \left\langle\mathbf{e}^n,\textstyle\sum_{i=1}^{s}b_i\mathbf{d}_i\right\rangle=\sum_{i=1}^{s}b_i\left\langle\mathbf{e}^n,\mathbf{d}_i\right\rangle\\ = \;& \textstyle\sum_{i=1}^{s}b_i\left\langle\boldsymbol{\rho}_i-\boldsymbol{\sigma}_i-k\sum_{j=1}^{s}a_{i,j}\mathbf{d}_j,\mathbf{d}_i\right\rangle\\ = \;& \textstyle\sum_{i=1}^{s}b_i\left\langle\boldsymbol{\rho}_i-\boldsymbol{\sigma}_i,\mathbf{d}_i\right\rangle-k\sum_{i=1}^{s}\sum_{j=1}^{s}b_ia_{i,j}\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle. \end{aligned}$$

The ODE system being contractive and (ABS-1) yield

$$\begin{aligned} & \textstyle\sum_{i=1}^{s}b_i\left\langle\boldsymbol{\rho}_i-\boldsymbol{\sigma}_i,\mathbf{d}_i\right\rangle\\ = \;& \textstyle\sum_{i=1}^{s}b_i\left\langle\boldsymbol{\rho}_i-\boldsymbol{\sigma}_i,\mathbf{f}(\boldsymbol{\rho}_i,t_n+c_ik)-\mathbf{f}(\boldsymbol{\sigma}_i,t_n+c_ik)\right\rangle\le 0, \end{aligned}$$

which further implies

$$\begin{aligned} \left\langle\mathbf{e}^n,\textstyle\sum_{i=1}^{s}b_i\mathbf{d}_i\right\rangle &\le -k\textstyle\sum_{i=1}^{s}\sum_{j=1}^{s}b_ia_{i,j}\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle,\\ \left\langle\mathbf{e}^n,\textstyle\sum_{i=1}^{s}b_i\mathbf{d}_i\right\rangle &\le -k\textstyle\sum_{j=1}^{s}\sum_{i=1}^{s}b_ja_{j,i}\left\langle\mathbf{d}_j,\mathbf{d}_i\right\rangle. \end{aligned}$$

It follows that

$$\begin{aligned} & 2\left\langle\mathbf{e}^n,\textstyle\sum_{i=1}^{s}b_i\mathbf{d}_i\right\rangle+k\sum_{i=1}^{s}\sum_{j=1}^{s}b_ib_j\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle\\ \le \;& -k\textstyle\sum_{i=1}^{s}\sum_{j=1}^{s}(b_ia_{i,j}+b_ja_{j,i}-b_ib_j)\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle\\ = \;& -k\textstyle\sum_{i=1}^{s}\sum_{j=1}^{s}m_{i,j}\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle. \end{aligned}$$

By (ABS-2), we can write $M=Q\Lambda Q^T$, where Q is orthogonal, and Λ is a diagonal matrix with $\lambda_k=\Lambda_{k,k}\ge 0$. Then the RK method is B-stable because

$$\begin{aligned} & \textstyle\sum_{i=1}^{s}\sum_{j=1}^{s}m_{i,j}\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle\\ = \;& \textstyle\sum_{i=1}^{s}\sum_{j=1}^{s}\left(\sum_{k=1}^{s}q_{i,k}\lambda_kq_{j,k}\right)\left\langle\mathbf{d}_i,\mathbf{d}_j\right\rangle\\ = \;& \textstyle\sum_{k=1}^{s}\lambda_k\left\langle\sum_{i=1}^{s}q_{i,k}\mathbf{d}_i,\sum_{j=1}^{s}q_{j,k}\mathbf{d}_j\right\rangle\\ = \;& \textstyle\sum_{k=1}^{s}\lambda_k\left\|\sum_{i=1}^{s}q_{i,k}\mathbf{d}_i\right\|^2\ge 0. \end{aligned}$$

The A-stability follows from Theorem 11.278. □

Example 11.281. A-stable RK methods may not be algebraically stable. For example, the three-stage method

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

is fourth-order accurate and A-stable. However, its algebraic stability matrix

$$M = \frac{1}{36}\begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

is not positive semidefinite.

Lemma 11.282. The algebraic stability matrix of an s-stage RK method is the zero matrix if

(i) the RK nodes $c_1, c_2, \ldots, c_s$ are pairwise distinct,

(ii) it is both $B(2s)$ and $C(s)$.

Proof. The Vandermonde matrix $V(c_1, c_2, \ldots, c_s)$ in Definition 2.3 is nonsingular since c_i's are pairwise distinct. Therefore $M = \mathbf{0}$ if and only if $\widetilde{M} := V^T M V = \mathbf{0}$. This actually holds because $\forall i, j = 1, 2, \ldots, s$, we have

$$\begin{aligned} &\widetilde{m}_{i,j} = \textstyle\sum_{k=1}^s \sum_{\ell=1}^s c_k^{i-1} m_{k,\ell} c_\ell^{j-1} \\ &= \textstyle\sum_{k=1}^s \sum_{\ell=1}^s c_k^{i-1}(b_k a_{k,\ell} + b_\ell a_{\ell,k} - b_k b_\ell) c_\ell^{j-1} \\ &= \textstyle\sum_{k=1}^s b_k c_k^{i-1} \sum_{\ell=1}^s a_{k,\ell} c_\ell^{j-1} + \sum_{\ell=1}^s b_\ell c_\ell^{j-1} \sum_{k=1}^s a_{\ell,k} c_k^{i-1} \\ &- \textstyle\sum_{k=1}^s b_k c_k^{i-1} \sum_{\ell=1}^s b_\ell c_\ell^{j-1} \\ &= \frac{1}{j}\sum_{k=1}^s b_k c_k^{i+j-1} + \frac{1}{i}\sum_{\ell=1}^s b_\ell c_\ell^{i+j-1} - \frac{1}{ij} \\ &= \left(\frac{1}{j} + \frac{1}{i}\right)\frac{1}{i+j} - \frac{1}{ij} = 0, \end{aligned}$$

where the second step follows from (ABS-2) in Definition (11.197) and the fourth and fifth steps from the conditions $B(2s)$, $C(s)$, Definition 11.194, and Definition 11.216. □

Theorem 11.283. The Gauss-Legendre RK methods are algebraically stable for $s \geq 1$, and hence are also B-stable and A-stable.

Proof. We already know that the Gauss quadrature formula with s quadrature nodes has degree of exactness $2s-1$ and that the quadrature weights b_i's are positive. By Definition 11.279, Exercise 11.196, and Lemma 11.282, it suffices to show that the s-stage Gauss-Legendre RK method is $C(s)$:

$$\begin{aligned} \forall \ell = 1, \ldots, s, \quad &\textstyle\sum_{j=1}^s a_{i,j} c_j^{\ell-1} = \sum_{j=1}^s \left(\int_0^{c_i} \ell_j(\tau)\mathrm{d}\tau\right) c_j^{\ell-1} \\ &= \textstyle\int_0^{c_i} \left(\sum_{j=1}^s c_j^{\ell-1}\ell_j(\tau)\right)\mathrm{d}\tau = \int_0^{c_i} \tau^{\ell-1}\mathrm{d}\tau, \end{aligned}$$

where the first step follows from (11.150) and the third from the fact that the Lagrange interpolation polynomial of $\tau^{\ell-1}$ is itself. The proof is completed by Theorem 11.280. □

11.7 Programming assignments

11.7.1 The context

The restricted three-body system is a model for studying the motion of an earth-moon satellite or of an asteroid close enough to the earth to be strongly influenced by the earth and the sun.

In this system, the two heavy bodies are regarded as revolving in fixed orbits about their common centre of mass while the small body is attracted by the heavy bodies but never affecting their motion. To further simplify the problem, we approximate orbits of the heavy bodies as circles so that we can work in a frame of reference that rotates with the two heavy bodies. In this frame, the two heavy bodies are considered as fixed in space with their rotation translated into a modification of the equations of gravitational motion. For simplicity, we scale the units to reduce a number of constants to one. We also scale the masses of the two heavy bodies to $1-\mu$ and μ and their positions relative to the moving reference frame by the vectors μe_1 and $(\mu-1)e_1$ respectively so that their center of mass is at the origin of coordinates.

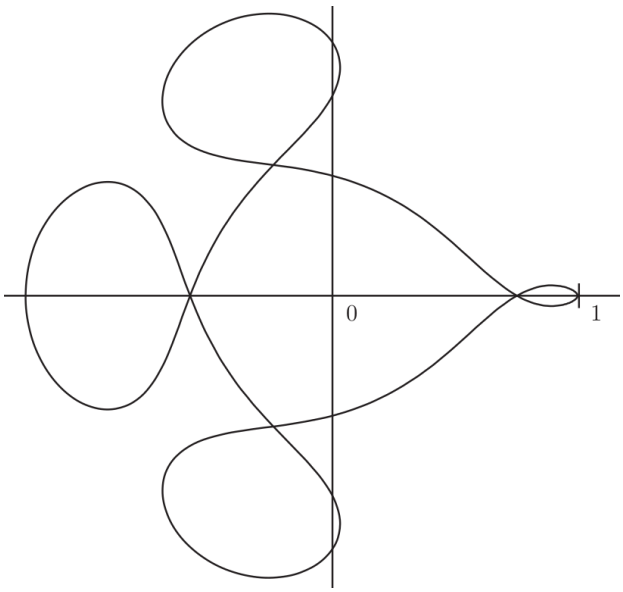
Denote by u_1, u_2 and u_3 the scalar variables representing the position coordinates of the small body and by u_4, u_5 and u_6 the corresponding velocity coordinates. The equations of motion are

$$\left\{\begin{array}{rl} u_1' = & u_4, \\ u_2' = & u_5, \\ u_3' = & u_6, \\ u_4' = & 2u_5 + u_1 - \frac{\mu(u_1+\mu-1)}{(u_2^2+u_3^2+(u_1+\mu-1)^2)^{3/2}} \\ & -\frac{(1-\mu)(u_1+\mu)}{(u_2^2+u_3^2+(u_1+\mu)^2)^{3/2}}, \\ u_5' = & -2u_4 + u_2 - \frac{\mu u_2}{(u_2^2+u_3^2+(u_1+\mu-1)^2)^{3/2}} \\ & -\frac{(1-\mu)u_2}{(u_2^2+u_3^2+(u_1+\mu)^2)^{3/2}}, \\ u_6' = & -\frac{\mu u_3}{(u_2^2+u_3^2+(u_1+\mu-1)^2)^{3/2}} \\ & -\frac{(1-\mu)u_3}{(u_2^2+u_3^2+(u_1+\mu)^2)^{3/2}}. \end{array}\right. \tag{11.198}$$

Planar motion is possible; that is, solutions which satisfy $u_3 = u_6 = 0$ at all times. One of these is shown below, with the values of (u_1, u_2) plotted as the orbit evolves. The heavier mass is at the point $(-\mu, 0)$ and the lighter mass is at $(1-\mu, 0)$, where $(0,0)$ is marked 0 and $(1,0)$ is marked 1. For this calculation the value of $\mu = 1/81.45$ was selected, corresponding to the earth-moon system. The initial values for this computation were

$$\begin{array}{rl} & (u_1, u_2, u_3, u_4, u_5, u_6) \\ = & (0.994, 0, 0, 0, -2.0015851063790825224, 0) \end{array} \tag{11.199}$$

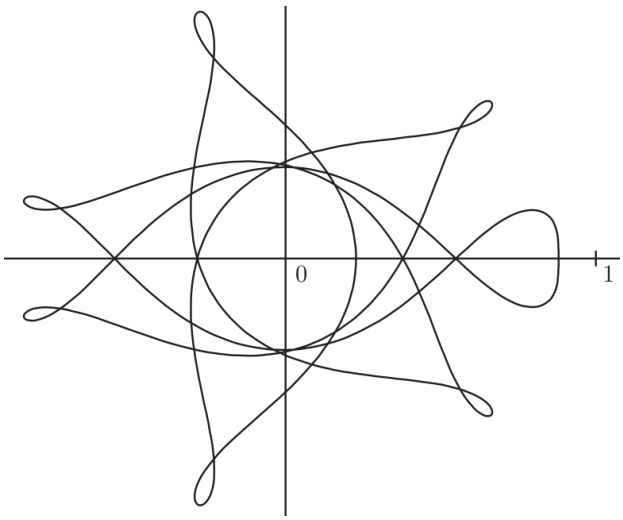
and the period was T_1 =17.06521656015796.



The plot below illustrates another solution of (11.198) with a different initial condition

$$\begin{array}{rl} & (u_1, u_2, u_3, u_4, u_5, u_6) \\ = & (0.87978, 0, 0, 0, -0.3797, 0) \end{array} \tag{11.200}$$

and a different period T_2 =19.14045706162071.



11.7.2 The assignments

Write a `C++` package to implement

- Adams-Bashforth methods ($p = 1, 2, 3, 4$),
- Adams-Moulton methods ($p = 2, 3, 4, 5$),
- BDFs ($p = 1, 2, 3, 4$),
- the classical RK method,
- the ESDIRK method in Example 11.205,
- Gauss-Legendre RK methods ($s = 1, 2, 3$),
- Fehlberg 4(5) embedded RK method,
- Dormand-Prince 5(4) embedded RK method.

You must test each method against the restricted three-body system with (11.199) and (11.200) in Section 11.7.1 as the initial conditions. Each solution should be computed to T_i for $i = 1, 2$ with a stable time-step size. For (11.199), the solution is periodic and you should use the initial condition as the exact solution at T_1 for calculating the solution errors. For (11.200), you should use Richardson extrapolation to compute the errors and convergence rates. For each method and for each test case, you are supposed to

(a) itemize all details of the test in an input file where the name and order of the method are parsed in the `main` program to generate an object of (a derived class of) the class `TimeIntegrator` by the singleton object `TimeIntegratorFactory`,

(b) perform a sequence of grid refinement test where solution errors, convergence rates, and CPU time are reported to demonstrate convergence with a correct order of accuracy (this is the criterion of whether your test passes or fails!),

(c) determine a time step size, as large as possible, with which the plot of the solutions is visually indistinguishable from the corresponding plot in Section 11.7.1.

Write a report to organize your results in a coherent story that relates your numerical results to the theory in this chapter. Your story should contain three plots of your solution to (11.199): Euler's method with 24000 steps of fixed size, the classical RK methods with 6000 steps of fixed size, and Dormand-Prince 5(4) embedded RK method with 100 steps of adaptive size. In addition, to achieve an error of 10^{-3} based on the max-norm of the solution error, which method is the winner in terms of total CPU time?