Some questions related to Iserles' textbook

Joshua Lee Padgett

March 14, 2022

Contents

1 Euler's method and beyond

The following questions are meant to help ensure you have a solid *conceptual* understanding of the material from Chapter 1 of Iserles' textbook.

Setting 1.1. Let $T \in (0, \infty)$, $d \in \mathbb{N} = \{1, 2, 3, \dots\}$, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be a function which satisfies for all $u, v \in \mathbb{R}^d$, $s \in \mathbb{R}$ that $\|u+v\| \leq \|u\| + \|v\|$, $\|su\| = |s|\|u\|$, and $\|u\| = 0$ if and only if $u = 0$, let $\lfloor \cdot \rfloor_h \colon [0, T] \to [0, T]$, $h \in (0, \infty)$, be the functions which satisfy for all $h \in (0, \infty)$, $t \in [0, T]$ that $\lfloor t \rfloor_h = \max([0, t] \cap \{0, h, 2h, \dots\})$, let $f \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies that

$$\left[\sup_{v \in \mathbb{R}^d} \|f(v)\|\right] + \left[\sup_{v, w \in \mathbb{R}^d, v \neq w} \frac{\|f(v) - f(w)\|}{\|v - w\|}\right] < \infty, \tag{1.2}$$

let $y \colon [0, T] \to \mathbb{R}^d$ be a measurable function which satisfies for all $t \in [0, T]$ that

$$y(t) = y(0) + \int_0^t f\big(y(s)\big) \, \mathrm{d}s, \tag{1.3}$$

and for every $h \in (0, \infty)$ let $Y_{0,h}, Y_{1,h}, \dots, Y_{\lfloor T/h \rfloor, h} \in \mathbb{R}^d$ satisfy for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor - 1\}$ that $Y_{0,h} = y(0)$ and

$$Y_{n+1,h} = Y_{n,h} + hf\big(Y_{n,h}\big). \tag{1.4}$$

Problem 1.5. Do you understand Setting 1.1 above? Do you understand what each individual component means and do you see why each component is necessary to present a well-defined numerical method (i.e., the method in Eq. (1.4))?

Proof of Problem 1.5.

The proof of Problem 1.5 is thus complete. □

Definition 1.6. Assume Setting 1.1. We say that Eq. (1.4) is a convergent numerical method for Eq. (1.3) if and only if it holds that

$$\lim_{h \to 0^+} \left[\max_{n \in \{0,1,\dots,\lfloor T/h \rfloor\}} \left\| y(nh) - Y_{n,h} \right\| \right] = 0. \tag{1.7}$$

Problem 1.8. Do you understand conceptually what the notion of convergence is implying? Can you see how the topology of the problem would come into play if we were not considering a problem posed in a finite-dimensional space?

Proof of Problem 1.8.

The proof of Problem 1.8 is thus complete. □

Lemma 1.9. Let $\alpha \in [0, \infty)$ and let $a_0, a_1, a_2, \ldots \in [0, \infty)$ and $b_0, b_1, b_2, \ldots \in [0, \infty)$ satisfy for all $n \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ that

$$a_n \le \alpha + \sum_{k=0}^{n-1} b_k a_k. \tag{1.10}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$a_n \le \alpha \exp\left(\sum_{k=0}^{n-1} b_k \right). \tag{1.11}$$

Proof of Lemma 1.9. First, we claim that for all $n \in \mathbb{N}_0$ it holds that

$$a_n \le \alpha \left[\prod_{k=0}^{n-1} (1 + b_k) \right]. \tag{1.12}$$

We now prove Eq. (1.12) by mathematical induction on $n \in \mathbb{N}_0$. For the base case $n = 0$, note that Eq. (1.10) ensures that

$$a_0 \le \alpha + \sum_{k=0}^{-1} b_k a_k = \alpha + 0 = \alpha. \tag{1.13}$$

Combining this and the fact that $\prod_{k=0}^{-1}(1+b_k)=1$ establishes Eq. (1.12) in the base case $n=0$. For the induction step $\mathbb{N}_0 \ni (n-1) \dashrightarrow n \in \mathbb{N}$, let $n \in \mathbb{N}$ and assume that for every $m \in \{0,1,\dots,n-1\}$ it holds that

$$a_m \leq \alpha \left[\prod_{k=0}^{m-1}(1+b_k)\right]. \tag{1.14}$$

This and Eq. (1.10) assure that

$$a_n \leq \alpha + \sum_{k=0}^{n-1} b_k a_k \leq \alpha + \sum_{k=0}^{n-1} b_k \left(\alpha \left[\prod_{j=0}^{k-1}(1+b_j)\right]\right) = \alpha \left(1 + \sum_{k=0}^{n-1}\left[\prod_{j=0}^{k-1}(1+b_j)\right] b_k\right). \tag{1.15}$$

Next, observe that

$$\begin{aligned} 1 + \sum_{k=0}^{n-1}\left[\prod_{j=0}^{k-1}(1+b_j)\right] b_k &= 1 + \sum_{k=0}^{n-1}\left[\prod_{j=0}^{k-1}(1+b_j)\right]\big((1+b_k)-1\big) \\ &= 1 + \sum_{k=0}^{n-1}\left[\prod_{j=0}^{k}(1+b_j) - \prod_{j=0}^{k-1}(1+b_j)\right] \\ &= 1 + \prod_{j=0}^{n-1}(1+b_j) - \prod_{j=0}^{-1}(1+b_j) = \prod_{j=0}^{n-1}(1+b_j). \end{aligned} \tag{1.16}$$

Combining this, Eq. (1.16), and mathematical induction establishes Eq. (1.12). Moreover, note that the fact that for all $x \in [0,\infty)$ it holds that $1+x \leq \exp(x)$, the assumption that $b_0, b_1, b_2, \ldots \in [0,\infty)$, and Eq. (1.12) imply that for all $n \in \mathbb{N}_0$ it holds that

$$a_n \leq \alpha \left[\prod_{k=0}^{n-1}(1+b_k)\right] \leq \alpha \left[\prod_{k=0}^{n-1}\exp(b_k)\right] \leq \alpha \exp\left(\sum_{k=0}^{n-1} b_k\right). \tag{1.17}$$

This establishes Eq. (1.11). The proof of Lemma 1.9 is thus complete. □

Problem 1.18. Assume Setting 1.1. Using Lemma 1.9 above, prove that there exists $C \in [0,\infty)$ such that for all $h \in (0,\infty)$ it holds that

$$\max_{n \in \{0,1,\ldots,\lfloor T/h \rfloor\}} \big\| y(nh) - Y_{n,h} \big\| \leq Ch. \tag{1.19}$$

Explain how proving Eq. (1.19) holds would relate to the notion of convergence (cf. Definition 1.6).

Proof of Problem 1.18.

The proof of Problem 1.18 is thus complete. □

Problem 1.20. Can you present the theta method from the textbook in the rigorous format used in Setting 1.1 above?

Proof of Problem 1.20.

The proof of Problem 1.20 is thus complete. □

Setting 1.21. Let $T_{\text{new}}, p \in (0, \infty)$, $d \in \mathbb{N}$, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be a function which satisfies for all $u, v \in \mathbb{R}^d$, $s \in \mathbb{R}$ that $\|u + v\| \leq \|u\| + \|v\|$, $\|su\| = |s|\|u\|$, and $\|u\| = 0$ if and only if $u = 0$, let $\lfloor \cdot \rfloor_h \colon [0, T_{\text{new}}] \to [0, T_{\text{new}}]$, $h \in (0, \infty)$, be the functions which satisfy for all

$h \in (0,\infty)$, $t \in [0, T_{\text{new}}]$ that $\lfloor t \rfloor_h = \max([0,t] \cap \{0, h, 2h, \dots\})$, let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies that

$$\sup_{v,w \in \mathbb{R}^d, v \neq w} \frac{\|g(v) - g(w)\|}{\big(1 + \|v\|^p + \|w\|^p\big)\|v - w\|} < \infty, \tag{1.22}$$

let $z\colon [0, T_{\text{new}}] \to \mathbb{R}^d$ be a measurable function which satisfies for all $t \in [0, T_{\text{new}}]$ that

$$z(t) = z(0) + \int_0^t g\big(z(s)\big)\,\mathrm{d}s, \tag{1.23}$$

and for every $h \in (0,\infty)$ let $Z_{0,h}, Z_{1,h}, \dots, Z_{\lfloor T_{\text{new}}/h \rfloor, h} \in \mathbb{R}^d$ satisfy for all $n \in \{0, 1, \dots, \lfloor {}^{T_{\text{new}}}/_h \rfloor - 1\}$ that $Z_{0,h} = z(0)$ and

$$Z_{n+1,h} = Z_{n,h} + hg\big(Z_{n,h}\big). \tag{1.24}$$

Problem 1.25. Can we prove a result similar to that in Problem 1.18 under the assumptions outline in Setting 1.21 above? If not, can we prove a result that is "similar" to the result in Problem 1.18? What additional assumptions (if any) are needed to prove either of the above results?

Proof of Problem 1.25.

The proof of Problem 1.25 is thus complete. □

1.1 An exploration of the linear case

Definition 1.26. We denote by $\exp\colon (\cup_{d\in\mathbb{N}}\mathbb{C}^{d\times d}) \to (\cup_{d\in\mathbb{N}}\mathbb{C}^{d\times d})$ the function which satisfies for all $d \in \mathbb{N}$, $A \in \mathbb{C}^{d\times d}$ that $\exp(A) = \sum_{k=0}^{\infty}(1/k!)A^k$.

Definition 1.27. For every $d \in \mathbb{N}$ let $\mathfrak{N}_d = \{1, 2, \ldots, d\}$, for every $d \in \mathbb{N}$ let $S_d = \{(\sigma\colon \mathfrak{N}_d \to \mathfrak{N}_d)\colon \sigma \text{ is a bijection}\}$, let $\mathfrak{p}\colon (\cup_{d\in\mathbb{N}}S_d) \to \mathbb{N}_0$ be the function which satisfies for all $d \in \mathbb{N}$, $\sigma \in S_d$ that $\mathfrak{p}(\sigma) = \sum_{i=1}^{d}\sum_{j=i+1}^{d}\mathbb{1}_{(0,\infty)}(\sigma_i - \sigma_j)$, and let $\operatorname{sgn}\colon (\cup_{d\in\mathbb{N}}S_d) \to \{-1, 1\}$ be the function which satisfies for all $d \in \mathbb{N}$, $\sigma \in S_d$ that $\operatorname{sgn}(\sigma) = (-1)^{\mathfrak{p}(\sigma)}$. Then we denote by $\det\colon (\cup_{d\in\mathbb{N}}\mathbb{R}^{d\times d}) \to \mathbb{R}$ the function which satisfies for all $d \in \mathbb{N}$, $A = (a_{i,j})_{i,j\in\{1,2,\ldots,d\}} \in \mathbb{R}^{d\times d}$ that $\det(A) = \sum_{\sigma\in S}[\operatorname{sgn}(\sigma)\prod_{i=1}^{d} a_{i,\sigma_i}]$.

Definition 1.28. We denote by $\operatorname{tr}\colon (\cup_{d\in\mathbb{N}}\mathbb{R}^{d\times d}) \to \mathbb{R}$ the function which satisfies for all $d \in \mathbb{N}$, $A = (a_{i,j})_{i,j\in\{1,2,\ldots,d\}} \in \mathbb{R}^{d\times d}$ that $\operatorname{tr}(A) = \sum_{i=1}^{d} a_{i,i}$.

Lemma 1.29. Let $d \in \mathbb{N}$, $A, B \in \mathbb{R}^{d\times d}$ and let $\|\cdot\|\colon \mathbb{R}^d \to [0, \infty)$ be a function which satisfies for all $u, v \in \mathbb{R}^d$, $s \in \mathbb{R}$ that $\|u + v\| \leq \|u\| + \|v\|$, $\|su\| = |s|\|u\|$, and $\|u\| = 0$ if and only if $u = 0$. Then

(i) it holds that $\|\exp(A)\| \leq \exp(\|A\|) < \infty$,

(ii) it holds for all $s, t \in \mathbb{R}$ that $\exp(sA + tA) = \exp(sA)\exp(tA)$,

(iii) it holds that $\exp(A)\exp(-A) = \mathrm{id}_{\mathbb{R}^{d\times d}}$,

(iv) it holds that $\exp(A + B) = \exp(A)\exp(B)$ if and only if it holds that $AB = BA$, and

(v) it holds that $\det(\exp(A)) = \exp(\operatorname{tr}(A))$

(cf. Definitions 1.26, 1.27, and 1.28).

Proof of Lemma 1.29.

The proof of Lemma 1.29 is thus complete. $\square$

Problem 1.30. Let $A \in \mathbb{R}^{2\times 2}$ satisfy

$$A = \begin{pmatrix} -1 & 1 \\ -2 & -4 \end{pmatrix}. \tag{1.31}$$

(i) Show that there exist $D = (d_{i,j})_{i,j\in\{1,2\}} \in \mathbb{R}^{2\times 2}$, $P \in \mathbb{R}^{2\times 2}$ with $\det(P) \neq 0$, $d_{1,2} = d_{2,1} = 0$, and $A = PDP^{-1}$ (cf. Definition 1.27).

(ii) Use the results from item (i) to show that

$$\exp(A) = \begin{pmatrix} 2\exp(-2) - \exp(-3) & \exp(-2) - \exp(-3) \\ 2\exp(-3) - 2\exp(-2) & 2\exp(-3) - \exp(-2) \end{pmatrix} \tag{1.32}$$

(cf. Definition 1.26).

Proof of Problem 1.30.

The proof of Problem 1.30 is thus complete. □

Problem 1.33. Let $T \in (0,\infty)$, let $\|\cdot\| \colon \mathbb{R}^2 \to [0,\infty)$ be the function which satisfies for all $u = (u_1, u_2) \in \mathbb{R}^2$ that $\|u\| = [|u_1|^2 + |u_2|^2]^{1/2}$, let $\lfloor \cdot \rfloor_h \colon [0,T] \to [0,T]$, $h \in (0,\infty)$, be the functions which satisfy for all $h \in (0,\infty)$, $t \in [0,T]$ that $\lfloor t \rfloor_h = \max([0,t] \cap \{0, h, 2h, \dots\})$, let $A \in \mathbb{R}^{2\times 2}$, $y \in C([0,T], \mathbb{R}^2)$ satisfy for all $t \in [0,T]$ that

$$A = \begin{pmatrix} -1 & 1 \\ -2 & -4 \end{pmatrix} \qquad \text{and} \qquad y(t) = (1,1)^* + \int_0^t Ay(s)\,\mathrm{d}s, \tag{1.34}$$

and for every $h \in (0,\infty)$ let $Y_{0,h}, Y_{1,h}, \dots, Y_{\lfloor T/h \rfloor, h} \in \mathbb{R}^2$ satisfy for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor - 1\}$ that $Y_{0,h} = y(0)$ and

$$Y_{n+1,h} = Y_{n,h} + hAY_{n,h}. \tag{1.35}$$

(i) Prove that for all $t \in [0,T]$ it holds that $y(t) = \exp(tA)y(0)$ (cf. Definition 1.26).

(ii) Prove that for all $h \in (0,\infty)$, $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$ it holds that

$$Y_{n,h} = \big(\mathrm{id}_{\mathbb{R}^{2\times 2}} + hA\big)^n y(0). \tag{1.36}$$

(iii) Prove that for all $h \in (0,\infty)$ it holds that

$$\begin{split} \Big\|\exp(hA)y(0) - \big(\mathrm{id}_{\mathbb{R}^{2\times 2}} + hA\big)y(0)\Big\| &= \left\| \int_0^h (h-s)A^2 \exp\big(sA\big)y(0)\,\mathrm{d}s \right\| \\ &\le \frac{h^2}{2}\left[\sup_{\mathfrak{h}\in(0,h)}\left(\sup_{v\in\mathbb{R}^2\setminus\{0\}} \frac{\|\exp(\mathfrak{h}A)v\|}{\|v\|}\right)\right]\big\|A^2 y(0)\big\| \le \sqrt{17}\,h^2 \end{split} \tag{1.37}$$

(cf. Definition 1.26).

(iv) Prove that for all $h \in (0,\infty)$, $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$ it holds that

$$\begin{split} &y(nh) - Y_{n,h} \\ &= \sum_{k=0}^{n-1} \exp\big(khA\big)\Big[\exp\big(hA\big) - \big(\mathrm{id}_{\mathbb{R}^{2\times 2}} + hA\big)\Big]\big(\mathrm{id}_{\mathbb{R}^{2\times 2}} + hA\big)^{(n-k-1)} y(0) \end{split} \tag{1.38}$$

(cf. Definition 1.26).

(v) Prove that

$$\sup_{h\in(0,\infty)}\left[\max_{n\in\{0,1,\dots,\lfloor T/h\rfloor\}}\big\|y(nh)-Y_{n,h}\big\|\right]\leq T\exp\big(9T/2\big)\sqrt{34}\,h. \tag{1.39}$$

Proof of Problem 1.33.

The proof of Problem 1.33 is thus complete. □

2 Multistep methods

Setting 2.1. Let $T\in(0,\infty)$, $d,s\in\mathbb{N}$, $a_0,a_1,\dots,a_s\in\mathbb{R}$, $b_0,b_1,\dots,b_s\in\mathbb{R}$, let $\|\cdot\|\colon\mathbb{R}^d\to[0,\infty)$ be a function which satisfies for all $u,v\in\mathbb{R}^d$, $s\in\mathbb{R}$ that $\|u+v\|\leq\|u\|+\|v\|$, $\|su\|=|s|\|u\|$, and $\|u\|=0$ if and only if $u=0$, let $\lfloor\cdot\rfloor_h\colon[0,T]\to[0,T]$, $h\in(0,\infty)$, be the functions which satisfy for all $h\in(0,\infty)$, $t\in[0,T]$ that $\lfloor t\rfloor_h=\max([0,t]\cap\{0,h,2h,\dots\})$, let $\mathcal{A}=\{g\colon[0,T]\to\mathbb{R}^d\colon f$ is analytic in $[0,T]\}$, for every $h\in(0,\infty)$, $n\in\{0,1,\dots,\lfloor T/h\rfloor-1\}$, $g\in\mathcal{A}$ let $\mathcal{D}\colon\mathcal{A}\to\mathcal{A}$ and $\mathcal{E}_h\colon\mathcal{A}\to\mathcal{A}$ satisfy

$$\big(\mathcal{D}g\big)(nh)=\big(\tfrac{\mathrm{d}}{\mathrm{d}t}g\big)(nh)\qquad\text{and}\qquad\big(\mathcal{E}_hg\big)(nh)=g\big((n+1)h\big), \tag{2.2}$$

let $f\colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies that

$$\left[\sup_{v\in\mathbb{R}^d}\|f(v)\|\right] + \left[\sup_{v,w\in\mathbb{R}^d, v\neq w}\frac{\|f(v)-f(w)\|}{\|v-w\|}\right] < \infty, \tag{2.3}$$

let $y\colon [0,T] \to \mathbb{R}^d$ be a measurable function which satisfies for all $t \in [0,T]$ that

$$y(t) = y(0) + \int_0^t f\big(y(s)\big)\,\mathrm{d}s, \tag{2.4}$$

and for every $h \in (0,\infty)$ let $Y_{0,h}, Y_{1,h}, \ldots, Y_{\lfloor T/h\rfloor,h} \in \mathbb{R}^d$ satisfy for all $n \in \{0,1,\ldots,\lfloor T/h\rfloor - s\}$ that $Y_{0,h} = y(0)$ and

$$\sum_{m=0}^{s} a_m Y_{n+m,h} = h\sum_{m=0}^{s} b_m f\big(Y_{n+m,h}\big). \tag{2.5}$$

Definition 2.6. Assume Setting 2.1. We say that Eq. (2.5) is a numerical method of order $p \in \mathbb{N}_0$ if and only if there exists $C \in (0,\infty)$ such that for all $h \in (0,\infty)$, $n \in \{0,1,\ldots,\lfloor T/h\rfloor\}$ with h sufficiently close to zero it holds that

$$\left\|\textstyle\sum_{m=0}^{s} a_m y\big((n+m)h\big) - h\sum_{m=0}^{s} b_m f\big(y\big((n+m)h\big)\big)\right\| \leq Ch^{p+1}. \tag{2.7}$$

Lemma 2.8. Assume Setting 2.1 and let $p \in \mathbb{N}$. Then Eq. (2.5) is of order p if and only if there exists $C \in (0,\infty)$ such that for all $z \in \mathbb{R}$ with z sufficiently close to one it holds that

$$\left|\textstyle\sum_{m=0}^{s} a_m z^m - \ln(z)\sum_{m=0}^{s} b_m z^m\right| \leq C|z-1|^{p+1} \tag{2.9}$$

(cf. Definition 2.6).

Proof of Lemma 2.8. Throughout this proof let $h \in (0,\infty)$ be sufficiently small, let $\rho\colon \mathbb{R} \to \mathbb{R}$ and $\sigma\colon \mathbb{R} \to \mathbb{R}$ be the functions which satisfy for all $z \in \mathbb{R}$ that

$$\rho(z) = \sum_{m=0}^{s} a_m z^m \qquad \text{and} \qquad \sigma(z) = \sum_{m=0}^{s} b_m z^m, \tag{2.10}$$

and without loss of generality assume that $y \in \mathcal{A}$. Note that Taylor's theorem guarantees that for all $n \in \{0,1,\ldots,\lfloor T/h\rfloor\}$, $k \in \mathbb{N}_0$ it holds that

$$\begin{aligned}\Big(\mathcal{E}_h\big(\tfrac{\mathrm{d}^k}{\mathrm{d}t^k}y\big)\Big)(nh) = \big(\tfrac{\mathrm{d}^k}{\mathrm{d}t^k}y\big)\big((n+1)h\big) &= \sum_{j=0}^{\infty}\frac{h^j}{j!}\big(\tfrac{\mathrm{d}^{k+j}}{\mathrm{d}t^{k+j}}y\big)(nh)\\ &= \sum_{j=0}^{\infty}\frac{h^j}{j!}\Big(\tfrac{\mathrm{d}^j}{\mathrm{d}t^j}\big(\tfrac{\mathrm{d}^k}{\mathrm{d}t^k}y\big)\Big)(nh)\\ &= \sum_{j=0}^{\infty}\frac{h^j}{j!}\Big(\mathcal{D}^j\big(\tfrac{\mathrm{d}^k}{\mathrm{d}t^k}y\big)\Big)(nh).\end{aligned} \tag{2.11}$$

Combining this and the fact that $\mathcal{D}$ is a bounded linear operator (something we have not shown, but which can be shown) ensures that

$$\mathcal{E}_h = \exp\big(h\mathcal{D}\big). \tag{2.12}$$

Next, observe that Eq. (2.4) assures that for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor - s\}$ it holds that

$$\begin{aligned} &\sum_{m=0}^{s} a_m y\big((n+m)h\big) - h \sum_{m=0}^{s} b_m f\big(y((n+m)h)\big) \\ &\quad = \sum_{m=0}^{s} a_m y\big((n+m)h\big) - h \sum_{m=0}^{s} b_m \big(\tfrac{\mathrm{d}}{\mathrm{d}t} y\big)\big((n+m)h\big) \\ &\quad = \sum_{m=0}^{s} a_m \big(\mathcal{E}_h^m y\big)(nh) - h \sum_{m=0}^{s} b_m \Big(\mathcal{E}_h^m \big(\mathcal{D}y\big)\Big)(nh). \end{aligned} \tag{2.13}$$

This, the fact that Eq. (2.12) implies that for all $g \in \mathcal{A}$ it holds that $(\mathcal{D}(\mathcal{E}_h g)) = (\mathcal{E}_h(\mathcal{D}g))$, the fact that $\mathcal{D}$ is a linear operator, and the so-called Borel functional calculus guarantee that for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor - s\}$ it holds that

$$\begin{aligned} &\sum_{m=0}^{s} a_m y\big((n+m)h\big) - h \sum_{m=0}^{s} b_m f\big(y((n+m)h)\big) \\ &\quad = \sum_{m=0}^{s} a_m \big(\mathcal{E}_h^m y\big)(nh) - h \left(\mathcal{D} \sum_{m=0}^{s} b_m \big(\mathcal{E}_h^m y\big)\right)(nh) \\ &\quad = \left(\left(\sum_{m=0}^{s} a_m \mathcal{E}_h^m - h\mathcal{D} \sum_{m=0}^{s} b_m \mathcal{E}_h^m\right) y\right)(nh) = \Big(\big(\rho(\mathcal{E}_h) - h\mathcal{D}\sigma(\mathcal{E}_h)\big) y\Big)(nh). \end{aligned} \tag{2.14}$$

This shows that for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor - s\}$ it holds that

$$\begin{aligned} &\left|\sum_{m=0}^{s} a_m y\big((n+m)h\big) - h \sum_{m=0}^{s} b_m f\big(y((n+m)h)\big)\right| \\ &\quad = \left|\Big(\big(\rho(\mathcal{E}_h) - h\mathcal{D}\sigma(\mathcal{E}_h)\big) y\Big)(nh)\right| \leq \left[\sup_{g \in \mathcal{A}\setminus\{0\}} \frac{\big|\big((\rho(\mathcal{E}_h) - h\mathcal{D}\sigma(\mathcal{E}_h))g\big)(nh)\big|}{|g(nh)|}\right] \big|y(nh)\big|. \end{aligned} \tag{2.15}$$

In addition, note that Eq. (2.12), the fact that for all $g \in \mathcal{A}$, $t \in [0, T]$ it holds that $\lim_{z \to 0^+} (\mathcal{E}_z g)(t) = g(t)$ (can you see that this is true?), and the implicit function theorem demonstrate that for all $g \in \mathcal{A}$, $t \in [0, T]$ it holds that

$$\big(h\mathcal{D}g\big)(t) = \big(\ln(\mathcal{E}_h) g\big)(t) = \left(\sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} \big(\mathcal{E}_h - \mathrm{id}\big)^{k+1} g\right)(t). \tag{2.16}$$

This and the Borel functional calculus yield that there exists $\gamma_h \subseteq \mathbb{C}$ (with the spectrum of $\mathcal{E}_h$ contained inside of γ_h—we can discuss this, if desired) such that for all $g \in \mathcal{A}$, $t \in [0, T]$ it holds that

$$\Big(\big(\rho(\mathcal{E}_h) - \ln(\mathcal{E}_h)\sigma(\mathcal{E}_h)\big) g\Big)(t) = \frac{1}{2\pi i} \int_{\gamma_h} \big[\rho(z) - \ln(z)\sigma(z)\big] \big((z\,\mathrm{id} - \mathcal{E}_h)^{-1} g\big)(t)\, \mathrm{d}z. \tag{2.17}$$

Combining Eqs. (2.15) and (2.17) hence proves Eq. (2.9). The proof of Lemma 2.8 is thus complete. □

3 Runge-Kutta methods

4 Stiff equations

Definition 4.1. Let $y_\lambda\colon [0,\infty) \to \mathbb{C}$, $\lambda \in \mathbb{C}$, be measurable functions which satisfy for all $\lambda \in \mathbb{C}$, $t \in [0,\infty)$ that

$$y_\lambda(t) = 1 + \lambda \int_0^t y(s)\,\mathrm{d}s, \tag{4.2}$$

let $h \in (0,\infty)$, for every $\lambda \in \mathbb{C}$ let $Y_{0,\lambda}, Y_{1,\lambda}, Y_{2,\lambda}, \ldots \in \mathbb{R}$ satisfy $Y_{0,\lambda} = 1$, and assume there exists $p, C \in (0,\infty)$ such that for all $\lambda \in \mathbb{C}$ with $\lambda + \bar{\lambda} \in (-\infty, 0)$ it holds that

$$\sup_{n\in\mathbb{N}_0} \left| y_\lambda(nh) - Y_{n,\lambda} \right| \leq Ch^p. \tag{4.3}$$

Then the set

$$\mathcal{D} = \left\{ h\lambda \in \mathbb{C}\colon \, \lim_{n\to\infty} Y_{n,\lambda} = 0 \right\} \subseteq \mathbb{C} \tag{4.4}$$

is the *linear stability domain* of the numerical method $\{Y_{n,\lambda}\}_{(n,\lambda)\in\mathbb{N}_0\times\mathbb{C}}$. Moreover, we say that the numerical method $\{Y_{n,\lambda}\}_{(n,\lambda)\in\mathbb{N}_0\times\mathbb{C}}$ is A-stable if it holds that

$$\left\{ z \in \mathbb{C}\colon \, z + \bar{z} \in (-\infty, 0) \right\} \subseteq \mathcal{D}. \tag{4.5}$$

Problem 4.6. Let $T \in (0,\infty)$, $d \in \mathbb{N}$, let $\|\cdot\|\colon \mathbb{R}^d \to [0,\infty)$ be a function which satisfies for all $u, v \in \mathbb{R}^d$, $s \in \mathbb{R}$ that $\|u + v\| \leq \|u\| + \|v\|$, $\|su\| = |s|\|u\|$, and $\|u\| = 0$ if and only if $u = 0$, let $\lfloor \cdot \rfloor_h\colon [0,T] \to [0,T]$, $h \in (0,\infty)$, be the functions which satisfy for all $h \in (0,\infty)$, $t \in [0,T]$ that $\lfloor t \rfloor_h = \max([0,t] \cap \{0, h, 2h, \ldots\})$, let $f \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ satisfy

$$\left[\sup_{v\in\mathbb{R}^d} \|f(v)\| \right] + \left[\sup_{v,w\in\mathbb{R}^d, v\neq w} \frac{\|f(v) - f(w)\|}{\|v - w\|} \right] < \infty, \tag{4.7}$$

let $y\colon [0,T] \to \mathbb{R}^d$ be a measurable function which satisfies for all $t \in [0,T]$ that

$$y(t) = y(0) + \int_0^t f\big(y(s)\big)\,\mathrm{d}s, \tag{4.8}$$

and for every $h \in (0,\infty)$ let $Y_{0,h}, Y_{1,h}, \ldots, Y_{\lfloor T/h \rfloor, h} \in \mathbb{R}^d$ satisfy for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor - 1\}$ that $Y_{0,h} = y(0)$ and

$$Y_{n+1,h} = Y_{n,h} + \tfrac{h}{4}\Big[f\big(Y_{n,h}\big) + 3f\big(Y_{n+1,h}\big) \Big]. \tag{4.9}$$

a. Determine whether or not Eq. (4.9) is consistent (cf. Definition 2.6). If Eq. (4.9) is consistent, determine its order.

b. Determine whether or not Eq. (4.9) is convergent (cf. Definition 1.6).

c. Determine whether or not Eq. (4.9) is A-stable (cf. Definition 4.1).

Proof of Problem 4.6.

The proof of Problem 4.6 is thus complete. □

5 Geometric numerical integration

6 Error control

7 Nonlinear algebraic systems

8 Finite difference schemes

Problem 8.1. Let $N \in \mathbb{N}_0$, $\alpha, \beta \in \mathbb{R}$, let $f \in C(\mathbb{R}, \mathbb{R})$ and $u \in C^4([0,1], \mathbb{R})$ satisfy for all $x \in [0,1]$ that $u(0) = \alpha$, $u(1) = \beta$, and

$$\big(\tfrac{\mathrm{d}^2}{\mathrm{d}x^2} u\big)(x) = f(x), \tag{8.2}$$

and let $h_0, h_1, \ldots, h_N, x_0, x_1, \ldots, x_{N+1} \in [0,1]$ satisfy for all $n \in \{0, 1, \ldots, N\}$ that

$$0 = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1 \qquad \text{and} \qquad h_n = x_{n+1} - x_n. \tag{8.3}$$

a. Construct a three-point finite difference scheme for approximating the solution to Eq. (8.2) on the non-uniform grid $\{x_n\}_{n \in \{0,1,\ldots,N+1\}} \subseteq [0,1]$ given by Eq. (8.3).

b. Determine the order of the method constructed in item a. above. Determine what additional assumptions are necessary (if any) for guaranteeing this order. Compare these results with the case from Section 8.2 of the textbook (i.e., the case when $h_0 = h_1 = \ldots = h_N$).

c. Write the finite difference scheme constructed in item a. above in the form of a linear system (i.e., as a matrix-vector equation).

d. Determine whether the linear system obtained in item c. is always nonsingular. If the linear system is not always nonsingular, provide sufficient conditions to guarantee that the linear system is nonsingular.

e. Implement your finite difference scheme (i.e., the difference equations from item a. above or the linear system from item c. above) in Python. Numerically compare the approximate solution with the true solution for some "test case."

Proof of Problem 8.1.

The proof of Problem 8.1 is thus complete. $\square$