18 Single-Step Methods

In this chapter, we begin the study of approximation methods for initial value problems (IVPs). The approach we will follow is motivated by (17.2), which is the defining relation that a mild solution must satisfy. Indeed, if $\boldsymbol{u}'(t) = \boldsymbol{f}(t, \boldsymbol{u}(t))$ and $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$, we can approximate $\boldsymbol{u}(t)$ via

$$\boldsymbol{u}(t) = \boldsymbol{u}_0 + \int_{t_0}^{t} \boldsymbol{f}(s, \boldsymbol{u}(s))\mathrm{d}s \approx u_0 + Q_{1,0}^{(t_0,t)}[\boldsymbol{f}(\,\cdot\,, \boldsymbol{u}(\,\cdot\,))],$$

where $Q_{1,0}^{(t_0,t)}$ is a quadrature formula. Notice that, in doing so, we now only require knowledge of $\boldsymbol{f}(\,\cdot\,, \boldsymbol{u}(\,\cdot\,))$ at the quadrature nodes. Thus, for instance, if we use the simple left-hand Riemann sum approximation:

$$Q_{1,0}^{(t_0,t)}[\boldsymbol{f}(\cdot, \boldsymbol{u}(\cdot))] = (t - t_0)\boldsymbol{f}(t_0, \boldsymbol{u}(t_0)).$$

Then the approximation becomes

$$\boldsymbol{u}(t) \approx \boldsymbol{u}_0 + (t - t_0)\boldsymbol{f}(t_0, \boldsymbol{u}(t_0))$$

or

$$\frac{1}{t - t_0}(\boldsymbol{u}(t) - \boldsymbol{u}_0) \approx \boldsymbol{f}(t_0, \boldsymbol{u}(t_0)).$$

This is Euler's famous approximation method. In this chapter, we will examine the convergence of methods of this type.

To simplify the discussion, in this and upcoming chapters, we consider the IVP over the "time" interval $I = [0, T]$ for $T > 0$. A simple linear transformation can be used to reduce the general case to this one. In addition, we will suppose that $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ is open, and $\boldsymbol{u}_0 \in \Omega$. We set $S = [0, T] \times \overline{\Omega}$ and assume that the slope function satisfies, at least, $\boldsymbol{f} \in C\big(S; \overline{\Omega}\big)$. We denote by $\boldsymbol{u} \in C^1([0, T]; \Omega)$ a classical solution on $[0, T]$ to the IVP

$$\boldsymbol{u}'(t) = \boldsymbol{f}(t, \boldsymbol{u}(t)), \quad \boldsymbol{u}(0) = \boldsymbol{u}_0. \tag{18.1}$$

The methods we will present here do not give us a function that approximates $\boldsymbol{u}$ but rather a sequence of vectors that approximate this function at a particular collection of points in time. More precisely, we let $K \in \mathbb{N}$, set $\tau = \frac{T}{K}$, which we call the *time step size*, and $t_k = k\tau$. We will then produce a finite sequence $\{\boldsymbol{w}^k\}_{k=0}^K \subset \mathbb{R}^d$ such that $\boldsymbol{w}^k \approx \boldsymbol{u}(t_k)$.

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

18.1 Single-Step Approximation Methods

We begin with a definition.

Definition 18.1 (single-step method)**.** The finite sequence $\left\{\boldsymbol{w}^k\right\}_{k=0}^K \subset \mathbb{R}^d$ is called a **single-step approximation** to $\boldsymbol{u}$ if and only if $\boldsymbol{w}^0 = \boldsymbol{u}_0$ and

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau \boldsymbol{G}\left(t_k, \tau, \boldsymbol{w}^k, \boldsymbol{w}^{k+1}\right), \quad k = 0, \dots, K-1, \tag{18.2}$$

where $\boldsymbol{G}$, called the **slope approximation**, satisfies $\boldsymbol{G}(t, 0, \boldsymbol{v}, \boldsymbol{v}) = \boldsymbol{f}(t, \boldsymbol{v})$ and

$$\boldsymbol{G} \in C([0, T] \times [0, T] \times \mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d).$$

The single-step approximation is called **explicit** if and only if $\boldsymbol{G}$ is independent of the last variable; otherwise, the approximation is called **implicit**. The **global error** of the single-step approximation is a finite sequence $\left\{\boldsymbol{e}^k\right\}_{k=0}^K$ defined via

$$\boldsymbol{e}^k = \boldsymbol{u}(t_k) - \boldsymbol{w}^k.$$

Let us present some examples of single-step approximation methods.

Example 18.1 The *forward (or explicit) Euler method*:[1]

$$\boldsymbol{G}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{G}_{FE}(t, s, \boldsymbol{v}_1) = \boldsymbol{f}(t, \boldsymbol{v}_1). \tag{18.3}$$

Example 18.2 The *backward (or implicit) Euler method*:

$$\boldsymbol{G}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{G}_{BE}(t, s, \boldsymbol{v}_2) = \boldsymbol{f}(t + s, \boldsymbol{v}_2). \tag{18.4}$$

Example 18.3 The *trapezoidal method*:

$$\boldsymbol{G}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{G}_{TR}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{2}\boldsymbol{f}(t, \boldsymbol{v}_1) + \frac{1}{2}\boldsymbol{f}(t + s, \boldsymbol{v}_2). \tag{18.5}$$

Example 18.4 *Taylor's method*:[2]

$$\begin{aligned} \boldsymbol{G}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) &= \boldsymbol{G}_{TM}(t, s, \boldsymbol{v}_1) \\ &= \boldsymbol{f}(t, \boldsymbol{v}_1) + \frac{s}{2}\left[\partial_t \boldsymbol{f}(t, \boldsymbol{v}_1) + D_{\boldsymbol{u}}\boldsymbol{f}(t, \boldsymbol{v}_1)\boldsymbol{f}(t, \boldsymbol{v}_1)\right], \end{aligned} \tag{18.6}$$

where $D_{\boldsymbol{u}}\boldsymbol{f} = \left[\partial_{u_j} f_i\right]_{i,j=1}^d$ is the $d \times d$ Jacobian matrix of partial derivatives of $\boldsymbol{f}$ with respect to $\boldsymbol{u}$.

Example 18.5 The *midpoint method*:

$$\boldsymbol{G}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{G}_{MR}(t, s, \boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{f}\left(t + \frac{s}{2}, \frac{1}{2}\boldsymbol{v}_1 + \frac{1}{2}\boldsymbol{v}_2\right). \tag{18.7}$$

[1] Named in honor of the Swiss mathematician, physicist, astronomer, geographer, logician, and engineer Leonhard Euler (1707–1783).

[2] Named in honor of the British mathematician Brook Taylor (1685–1731).

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

Definition 18.2 (LTE and convergence)**.** Let $\left\{\boldsymbol{w}^k\right\}_{k=0}^{K}$ be a single-step approximation to $\boldsymbol{u}$ generated by the slope approximation $\boldsymbol{G}$. The **local truncation error** (LTE) or **consistency error** of the single-step approximation is defined as

$$\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,s) = \frac{\boldsymbol{u}(t) - \boldsymbol{u}(t-s)}{s} - \boldsymbol{G}(t-s, s, \boldsymbol{u}(t-s), \boldsymbol{u}(t))$$

for any $t \in [s, T]$. We make frequent use of the notation $\boldsymbol{\mathcal{E}}^k[\boldsymbol{u}] = \boldsymbol{\mathcal{E}}[\boldsymbol{u}](t_k, \tau)$ for $k = 1, \dots, K$. We say that the approximation method is **consistent to at least order** $p \in \mathbb{N}$ if and only if, whenever

$$\boldsymbol{u} \in C^{p+1}\left([0, T]; \Omega\right),$$

there is a constant $\tau_0 \in (0, T]$ and a constant $C > 0$ independent of t and τ such that

$$\|\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,\tau)\|_2 \leq C\tau^p \tag{18.8}$$

for all $\tau \in (0, \tau_0]$ and $t \in [\tau, T]$. We say that the single-step approximation is **consistent to exactly order** p if and only if p is the largest positive integer for which (18.8) holds regardless of how smooth the exact solution $\boldsymbol{u}$ is.

We say that the single-step approximation method **converges globally** if

$$\lim_{K \to \infty} \max_{k=0,\dots,K} \|\boldsymbol{e}^k\|_2 = 0.$$

In addition, we say that it converges globally, with at least order $p \in \mathbb{N}$, if and only if, when

$$\boldsymbol{u} \in C^{p+1}\left([0, T]; \Omega\right),$$

there is some $\tau_1 \in (0, T]$ and a constant $C > 0$ independent of k and τ such that

$$\left\|\boldsymbol{e}^k\right\|_2 \leq C\tau^p$$

for all $k = 1, \dots, K$ and any $\tau \in (0, \tau_1]$.

18.2 Consistency and Convergence

Let us now study the consistency and convergence of some single-step approximation methods. We will present a few illustrative cases that highlight the type of techniques and ideas that are needed. The consistency and convergence of many other methods are found in the Problems at the end of the chapter.

A useful tool in this will be a discrete analogue of the Grönwall[3] inequality proved in Lemma 17.7.

Lemma 18.3 (discrete Grönwall)**.** *Let $K \in \mathbb{N}$. Suppose that the finite sequence $\{a_k\}_{k=0}^{K} \subset \mathbb{R}_+ \cup \{0\}$ satisfies $a_0 = 0$ and, for some $b > 1$, $c \geq 0$,*

$$a_{k+1} \leq b a_k + c, \quad k = 0, \dots, K-1.$$

[3] Named in honor of the Swedish–American mathematician Thomas Hakon Grönwall (1877–1932).

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

Then, for all $k = 0, \dots, K$,

$$a_k \leq \frac{c}{b-1}\left[b^k - 1\right].$$

Proof. The proof is by induction. The base case, $k = 0$, is trivial since $a_0 = 0$.

For the induction hypothesis, we assume that

$$a_k \leq \frac{c}{b-1}\left[b^k - 1\right]$$

holds for every $k = 0, \dots, n$.

For the induction step, we observe that, by assumption,

$$\begin{aligned} a_{n+1} &\leq b a_n + c \\ &\leq b\frac{c}{b-1}\left[b^n - 1\right] + c \\ &= \frac{c}{b-1}\left[b^{n+1} - b\right] + c \\ &= \frac{c}{b-1}b^{n+1} - \frac{bc}{b-1} + \frac{(b-1)c}{b-1} \\ &= \frac{c}{b-1}b^{n+1} - \frac{c}{b-1} \\ &= \frac{c}{b-1}\left[b^{n+1} - 1\right], \end{aligned}$$

which completes the induction argument. □

18.2.1 Forward Euler Method

Proposition 18.4 (consistency). *Suppose that $d = 1$ and $f \in \mathcal{F}^1(S)$. Then, for all $s \in (0, T]$ and $t \in [s, T]$,*

$$u(t) = u(t-s) + sf(t-s, u(t-s)) + s\mathcal{E}[u](t,s),$$

where $\mathcal{E}[u](t,s)$ satisfies

$$|\mathcal{E}[u](t,s)| \leq Cs$$

and $C > 0$ is a constant that is independent of t and s.

Proof. Since $f \in \mathcal{F}^1(S)$, $u \in C^2([0,T])$. Fix $s \in (0,T]$ and $t \in [s,T]$. By Taylor's Theorem B.31, for some $\eta \in (t-s, t)$,

$$u(t) = u(t-s) + u'(t-s)s + \frac{1}{2}u''(\eta)s^2.$$

Hence,

$$u(t) = u(t-s) + sf(t-s, u(t-s)) + s\mathcal{E}[u](t,s),$$

where

$$\mathcal{E}[u](t,s) = \frac{1}{2}u''(\eta)s.$$

The result is proved using that

$$|\mathcal{E}[u](t,s)| \leq \frac{s}{2}\max_{t-s\leq\eta\leq t}|u''(\eta)| \leq \frac{s}{2}\max_{0\leq\eta\leq T}|u''(\eta)|$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

and taking

$$C = \frac{1}{2} \max_{0 \le \eta \le T} |u''(\eta)|. \qquad \square$$

An analogous result, for $d > 1$, immediately implies consistency of the forward Euler method.

Corollary 18.5 (consistency)**.** *The forward Euler method* (18.3) *is of exactly order $p = 1$. In other words, if $\boldsymbol{f} \in \mathcal{F}^1(S)$ and $\boldsymbol{u} \in C^2([0,T];\mathbb{R}^d)$ is the unique solution to* (18.1)*, then, for all $\tau \in (0,T]$ and $t \in (\tau, T]$,*

$$\|\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,\tau)\|_2 \le C_{FE}\tau \tag{18.9}$$

for some $C_{FE} > 0$ that is independent of t and τ.

Proof. See Problem 18.3. □

Theorem 18.6 (convergence)**.** *Suppose that $\boldsymbol{f} \in \mathcal{F}^1(S)$. Let $L > 0$ be its $\boldsymbol{u}$-Lipschitz constant on S. Suppose that the forward Euler method* (18.3) *is used to approximate $\boldsymbol{u}$, the unique solution to* (18.1)*. Then, for all $k = 0, \dots, K$,*

$$\left\|\boldsymbol{e}^k\right\|_2 \le \frac{C_{FE}}{L}\left[e^{TL} - 1\right]\tau,$$

where $C_{FE} > 0$ is the LTE constant from (18.9)*. Consequently,*

$$\max_{k=1,\dots,K} \left\|\boldsymbol{e}^k\right\|_2 \le \frac{C_{FE}}{L}\left[e^{TL} - 1\right]\tau.$$

Proof. Recall that $\boldsymbol{e}^k = \boldsymbol{u}(t_k) - \boldsymbol{w}^k$ and notice that we have the error equation

$$\boldsymbol{e}^{k+1} = \boldsymbol{e}^k + \tau \boldsymbol{f}(t_k, \boldsymbol{u}(t_k)) - \tau \boldsymbol{f}(t_k, \boldsymbol{w}^k) + \tau \boldsymbol{\mathcal{E}}^{k+1}[\boldsymbol{u}]$$

for $k = 0, \dots, K-1$ with $\boldsymbol{e}^0 = \boldsymbol{0}$. Using the triangle inequality, the Lipschitz condition, and the LTE bound (18.9), we have, for $k = 0, \dots, K-1$,

$$\left\|\boldsymbol{e}^{k+1}\right\|_2 \le \left\|\boldsymbol{e}^k\right\|_2 + \tau L \left\|\boldsymbol{e}^k\right\|_2 + \tau \left\|\boldsymbol{\mathcal{E}}^{k+1}[u]\right\|_2 \le (1 + \tau L)\left\|\boldsymbol{e}^k\right\|_2 + C_{FE}\tau^2.$$

Using Lemma 18.3, we find, for $k = 0, \dots, K$,

$$\left\|\boldsymbol{e}^k\right\|_2 \le \frac{C_{FE}}{L}\left[(1 + \tau L)^k - 1\right]\tau.$$

Now, since $\tau L > 0$, $1 + \tau L < e^{\tau L}$. Hence,

$$\left\|\boldsymbol{e}^k\right\|_2 \le \frac{C_{FE}}{L}\left[e^{\tau k L} - 1\right]\tau \le \frac{C_{FE}}{L}\left[e^{TL} - 1\right]\tau$$

for all $k = 0, \dots, K$. □

18.2.2 Trapezoidal Method

Proposition 18.7 (consistency)**.** *Suppose that $d = 1$ and $f \in \mathcal{F}^2(S)$. Then, for all $s \in (0,T]$ and $t \in [s,T]$, we have*

$$u(t) = u(t-s) + \frac{s}{2}\left[f(t-s, u(t-s)) + f(t, u(t))\right] + s\mathcal{E}[u](t,s),$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

where $\mathcal{E}[u](t,s)$ satisfies

$$|\mathcal{E}[u](t,s)| \leq Cs^2,$$

where $C > 0$ is a constant that is independent of s and t.

Proof. Since $f \in \mathcal{F}^2(S)$, we have that $u \in C^3([0,T])$. By Taylor's Theorem B.31, for some $\eta \in (t-s, t-s/2)$,

$$u(t-s/2) = u(t-s) + u'(t-s)\frac{s}{2} + \frac{1}{2}u''(\eta)\frac{s^2}{4}.$$

Likewise, for some $\zeta \in (t-s/2, t)$,

$$u(t-s/2) = u(t) + u'(t)\frac{(-s)}{2} + \frac{1}{2}u''(\zeta)\frac{(-s)^2}{4}.$$

Subtracting, we have

$$u(t) = u(t-s) + \frac{s}{2}\left(u'(t-s) + u'(t)\right) + \frac{s^2}{8}\left(u''(\eta) - u''(\zeta)\right).$$

Using the Mean Value Theorem B.30, for some $\chi \in (\eta, \zeta)$,

$$u''(\eta) - u''(\zeta) = u'''(\chi)(\eta - \zeta).$$

Hence,

$$u(t) = u(t-s) + \frac{s}{2}\left[f(t-s, u(t-s)) + f(t, u(t))\right] + s\mathcal{E}[u](t,s),$$

where

$$\mathcal{E}[u](t,s) = s\frac{\eta - \zeta}{8}u'''(\chi).$$

Since $|\eta - \zeta| \leq s$ and $u \in C^3([0,T])$, the result is proved via the following estimate:

$$|\mathcal{E}[u](t,s)| \leq \frac{s^2}{8}\max_{t-s\leq\chi\leq t}|u'''(\chi)| \leq \frac{s^2}{8}\max_{t\in[0,T]}|u'''(t)|. \qquad \square$$

An extension, to $d > 1$, of this result implies consistency of the trapezoidal method.

Corollary 18.8 (consistency)**.** *The trapezoidal method* (18.5) *is of order exactly $p = 2$. Precisely, if $\boldsymbol{f} \in \mathcal{F}^2(S)$, then, for all $\tau \in (0,T]$ and $t \in [\tau, T]$,*

$$\|\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,\tau)\|_2 \leq C_{TR}\tau^2 \tag{18.10}$$

for some $C_{TR} > 0$ that is independent of t and s.

Proof. See Problem 18.6. □

We can now obtain global convergence of the trapezoidal method.

Theorem 18.9 (convergence)**.** *Let $\boldsymbol{f} \in \mathcal{F}^2(S)$ and $L > 0$ be its $\boldsymbol{u}$-Lipschitz constant on S. Suppose that the trapezoidal method* (18.5) *is used to approximate the solution to* (18.1)*. Then, for all $k = 0, \ldots, K$, we have*

$$\left\|\boldsymbol{e}^k\right\|_2 \leq \frac{C_{TR}}{L}\left[\exp(2TL) - 1\right]\tau^2,$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

provided that $0 < \tau L < 1$*, where* $C_{TR} > 0$ *is the LTE constant from* (18.10).

Proof. As in the case of the forward Euler method, we begin by identifying an equation for the error $\boldsymbol{e}^k = \boldsymbol{u}(t_k) - \boldsymbol{w}^k$. In this case, we have

$$\begin{aligned}\boldsymbol{e}^{k+1} = \boldsymbol{e}^k &+ \frac{\tau}{2}\left[\boldsymbol{f}(t_k, \boldsymbol{u}(t_k)) - \boldsymbol{f}(t_k, \boldsymbol{w}^k)\right] + \frac{\tau}{2}\left[\boldsymbol{f}(t_{k+1}, \boldsymbol{u}(t_{k+1})) - \boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1})\right] \\ &+ \tau \boldsymbol{\mathcal{E}}^{k+1}[\boldsymbol{u}].\end{aligned}$$

We take norms and apply the triangle inequality, the $\boldsymbol{u}$-Lipschitz condition on $\boldsymbol{f}$ and (18.10), to obtain

$$\left(1 - \frac{\tau L}{2}\right)\|\boldsymbol{e}^{k+1}\|_2 \le \left(1 + \frac{\tau L}{2}\right)\|\boldsymbol{e}^k\|_2 + C_{TR}\tau^3.$$

Since, by assumption, $\frac{1}{2} < 1 - \frac{\tau L}{2} < 1$, Lemma 18.3 then implies that

$$\begin{aligned}\|\boldsymbol{e}^{k+1}\|_2 &\le \frac{C_{TR}\tau^3}{1 - \frac{\tau L}{2}}\left[\left(\frac{1 + \frac{\tau L}{2}}{1 - \frac{\tau L}{2}}\right) - 1\right]^{-1}\left[\left(\frac{1 + \frac{\tau L}{2}}{1 - \frac{\tau L}{2}}\right)^k - 1\right] \\ &= \frac{C_{TR}}{L}\left[\left(\frac{1 + \frac{\tau L}{2}}{1 - \frac{\tau L}{2}}\right)^k - 1\right]\tau^2.\end{aligned}$$

Finally, notice that

$$\frac{1 + \frac{\tau L}{2}}{1 - \frac{\tau L}{2}} = 1 + \frac{\tau L}{1 - \frac{\tau L}{2}} \le 1 + 2\tau L \le e^{2\tau L}$$

to, in conclusion, obtain that

$$\|\boldsymbol{e}^{k+1}\|_2 \le \frac{C_{TR}}{L}[e^{2k\tau L} - 1]\tau^2 \le \frac{C_{TR}}{L}[e^{2TL} - 1]\tau^2,$$

as claimed. □

18.2.3 Taylor's Method

As a last example, we consider the consistency and convergence of Taylor's method.

Theorem 18.10 (consistency)**.** *Suppose that* $\boldsymbol{f} \in \mathcal{F}^2(S)$*. For any* $s \in (0, T]$ *and* $t \in (s, T]$*, we have that the LTE of Taylor's method satisfies*

$$\|\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t, s)\|_2 \le C_{TM}s^2 \tag{18.11}$$

for some $C_{TM} > 0$ *that is independent of* t *and* s*. In other words, Taylor's method is consistent to exactly order* $p = 2$*.*

Proof. For simplicity, we will only consider the case $d = 1$. Since $f \in \mathcal{F}^2(S)$, we know that $u \in C^3([0, T])$. Thus, using Taylor's Theorem, we get that

$$u(t) = u(t - s) + su'(t - s) + \frac{s^2}{2}u''(t - s) + \frac{s^3}{6}u'''(\xi)$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

for some $\xi \in (t-s,t)$. Now we note that $u'(t-s) = f(t-s,u(t-s))$ and also

$$u''(t-s) = \frac{\partial f}{\partial t}(t-s,u(t-s)) + \frac{\partial f}{\partial u}(t-s,u(t-s))f(t-s,u(t-s)).$$

Hence,

$$\begin{aligned} s\mathcal{E}[u](t,s) &= u(t) - u(t-s) - sf(t-s,u(t-s)) - \frac{s^2}{2}\frac{\partial f}{\partial t}(t-s,u(t-s)) \\ &\quad - \frac{s^2}{2}\frac{\partial f}{\partial u}(t-s,u(t-s))f(t-s,u(t-s)) \\ &= \frac{s^3}{6}u'''(\xi). \end{aligned}$$

The result follows. □

Theorem 18.11 (convergence)**.** *Suppose that* $\boldsymbol{f} \in \mathcal{F}^2(S)$ *and there is some constant* $B > 0$ *such that*

$$|D^{\boldsymbol{\alpha}} f_i(t,\boldsymbol{v})| \leq B$$

for all multi-indices $\boldsymbol{\alpha} \in \mathbb{N}^{d+1}$ *with* $|\boldsymbol{\alpha}| = 2$*, for all* $i = 1,\dots,d$*, and for all* $(t,\boldsymbol{v}) \in S$*. (In other words, all second derivatives are bounded on* S*.) Then Taylor's method* (18.6) *is convergent and the global rate of convergence is* $p = 2$*. In particular,*

$$\left\|\boldsymbol{e}^k\right\|_2 \leq \frac{C_{TM}}{L'}\left[e^{TL'} - 1\right]s^2,$$

where $C_{TM} > 0$ *is the LTE constant from* (18.11) *and* $L' > 0$ *is a Lipschitz constant given below.*

Proof. The key step in this proof is to establish a global $\boldsymbol{u}$-Lipschitz continuity for the slope approximation, i.e., an estimate of the form

$$\|\boldsymbol{G}_{TM}(t,s,\boldsymbol{v}_1) - \boldsymbol{G}_{TM}(t,s,\boldsymbol{v}_2)\|_2 \leq L'\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2,$$

for any $s \in (0,T]$, for any $t \in [s,T]$, and for all $\boldsymbol{v}_1,\boldsymbol{v}_2 \in \mathbb{R}^d$. This requires the extra assumptions on the slope function that are included in the hypotheses. The details are left to the reader as an exercise; see Problem 18.9. □

18.3 Linear Slope Functions

In the case that the slope function is linear, we can provide more direct proofs of convergence. The following example will be of interest in Chapter 28, where we examine numerical methods for the heat equation.

Theorem 18.12 (convergence)**.** *Let* $\mathsf{A} \in \mathbb{R}^{d\times d}$ *be symmetric. Suppose that* $\boldsymbol{u}\colon [0,T] \to \mathbb{R}^d$ *is the solution to*

$$\boldsymbol{u}'(t) = \mathsf{A}\boldsymbol{u}(t), \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0 \in \mathbb{R}^d.$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

Let $K \in \mathbb{N}$. Suppose that the sequence $\left\{\boldsymbol{w}^k\right\}_{k=0}^K$ is generated using the forward Euler method (18.3)*. Then, for all $k = 0, \dots, K$, we have*

$$\left\|\boldsymbol{e}^k\right\|_2 \le \|\boldsymbol{u}_0\|_2 \max_{\lambda\in\sigma(\mathsf{A})} \left|e^{\lambda k\tau} - (1+\lambda\tau)^k\right|. \tag{18.12}$$

Suppose that the maximum on the right-hand side of (18.12) *is achieved at $\lambda_{\max} \in \sigma(\mathsf{A})$ and, furthermore, that $\lambda_{\max} < 0$. Then there is a constant $\tau_0 \in (0, T]$ such that, for all $\tau \in (0, \tau_0]$ and all $k = 1, 2, \dots, K$,*

$$\left\|\boldsymbol{e}^k\right\|_2 \le \frac{T}{2}\lambda_{\max}^2 \|\boldsymbol{u}_0\|_2 \tau.$$

Proof. Since $\mathsf{A} \in \mathbb{R}^{d\times d}$ is symmetric, it is orthogonally diagonalizable. In other words, there exists a diagonal matrix D (whose diagonal entries $[\mathsf{D}]_{i,i} = \lambda_i$ are the eigenvalues of A) and an orthogonal matrix Q such that $\mathsf{A} = \mathsf{Q}\mathsf{D}\mathsf{Q}^\intercal$. The exact solution of the equation is then given by

$$\boldsymbol{u}(t) = \mathsf{Q}e^{t\mathsf{D}}\mathsf{Q}^\intercal \boldsymbol{u}_0,$$

where $e^{t\mathsf{D}}$ is a diagonal matrix whose diagonal entries are precisely $\left[e^{t\mathsf{D}}\right]_{i,i} = e^{t\lambda_i}$. Now using the forward Euler method, it is easy to see that

$$\boldsymbol{w}^k = (\mathsf{I} + \tau\mathsf{A})^k \boldsymbol{u}_0 = \mathsf{Q}(\mathsf{I} + \tau\mathsf{D})^k\, \mathsf{Q}^\intercal \boldsymbol{u}_0.$$

Thus,

$$\boldsymbol{e}^k = \mathsf{Q}e^{k\tau\mathsf{D}}\mathsf{Q}^\intercal\boldsymbol{u}_0 - \mathsf{Q}(\mathsf{I}+\tau\mathsf{D})^k\,\mathsf{Q}^\intercal\boldsymbol{u}_0 = \mathsf{Q}\left(e^{k\tau\mathsf{D}} - (\mathsf{I}+\tau\mathsf{D})^k\right)\mathsf{Q}^\intercal\boldsymbol{u}_0.$$

Taking norms, we get

$$\left\|\boldsymbol{e}^k\right\|_2 \le \left\|\mathsf{Q}\left(e^{k\tau\mathsf{D}} - (\mathsf{I}+\tau\mathsf{D})^k\right)\mathsf{Q}^\intercal\right\|_2 \|\boldsymbol{u}_0\|_2 \le \left\|e^{k\tau\mathsf{D}} - (\mathsf{I}+\tau\mathsf{D})^k\right\|_2 \|\boldsymbol{u}_0\|_2.$$

Notice that $(\mathsf{I}+\tau\mathsf{D})^k$ is a diagonal matrix with the entries $(1+\tau\lambda_i)^k$. To conclude, we use the fact that the 2-norm of a diagonal matrix is simply the largest diagonal element in absolute value. Hence,

$$\left\|\boldsymbol{e}^k\right\|_2 \le \|\boldsymbol{u}_0\|_2 \max_{\lambda\in\sigma(\mathsf{A})} \left|e^{k\tau\lambda} - (1+\tau\lambda)^k\right|.$$

Now, using Taylor expansions, we see that, for any $x \le 0$,

$$1 + x \le e^x \le 1 + x + \frac{1}{2}x^2.$$

Equivalently,

$$1 + x - \frac{1}{2}x^2 \le e^x - \frac{1}{2}x^2 \le 1 + x \le e^x.$$

Using the binomial expansion, it follows that, if $n \ge 2$,

$$(1-\alpha)^n = 1 - n\alpha + \binom{n}{2}\alpha^2 + \sum_{j=3}^n \binom{n}{j}(-1)^j\alpha^j.$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

There is an $\alpha_0 \in (0,1)$ such that if $\alpha \in (0,\alpha_0)$, then

$$\binom{n}{2}\alpha^2 + \sum_{j=3}^{n}\binom{n}{j}(-1)^j\alpha^j \geq 0.$$

Essentially, the first term, which is positive, dominates the others. Thus, if $n \geq 2$ and $\alpha \in (0,\alpha_0)$,

$$(1-\alpha)^n \geq 1 - n\alpha.$$

We apply the last estimate with

$$\alpha = \frac{x^2}{2e^x}.$$

Thus,

$$1 - n\frac{x^2}{2e^x} \leq \left(1 - \frac{x^2}{2e^x}\right)^n.$$

Multiplying by e^{nx}, we get

$$e^{nx} - n\frac{x^2}{2}e^{(n-1)x} \leq \left(e^x - \frac{x^2}{2}\right)^n \leq (1+x)^n \leq e^{nx},$$

provided that $x \in (-1,0]$ and x is sufficiently small in absolute value.

Now, if $\tau\lambda_{\max} \in [-1,0)$ is sufficiently small in absolute value,

$$-k\frac{(\tau\lambda_{\max})^2}{2}e^{(k-1)\tau\lambda_{\max}} \leq (1+\tau\lambda_{\max})^k - e^{k\tau\lambda_{\max}} \leq 0,$$

or, equivalently,

$$0 \leq e^{k\tau\lambda_{\max}} - (1+\tau\lambda_{\max})^k \leq k\frac{(\tau\lambda_{\max})^2}{2}e^{(k-1)\tau\lambda_{\max}} \leq \frac{T}{2}\lambda_{\max}^2\tau.$$

The result follows. □

Problems

18.1 Alternate discrete Grönwall inequality: Let $K \in \mathbb{N}$. Suppose that the sequence $\{a_k\}_{k=0}^{K} \subset \mathbb{R}_+$ is such that there are $b > 0$ and $c \geq 0$ for which

$$a_{k+1} \leq ba_k + c, \quad k = 0,\ldots,K-1.$$

Prove that, for all $k = 0,\ldots,K$,

$$a_k \leq b^k a_0 + \left(\sum_{j=0}^{k-1} b^j\right)c.$$

For $b \neq 1$, show that

$$a_k \leq b^k a_0 + \frac{b^k - 1}{b-1}c.$$

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press

18.2 Yet another discrete Grönwall inequality: Let $\gamma, \beta \in \mathbb{R}$ with $\beta > 0$ and $\gamma > -1$. Let $\{a_n\}_{n\in\mathbb{N}}$ and $\{f_n\}_{n\in\mathbb{N}}$ be two sequences of nonnegative real numbers satisfying

$$(1+\gamma)a_{n+1} \leq a_n + \beta f_n.$$

Prove that

$$a_{n+1} \leq \frac{a_0}{(1+\gamma)^{n+1}} + \beta \sum_{k=0}^{n} \frac{f_k}{(1+\gamma)^{n-k+1}}.$$

18.3 Prove Corollary 18.5.

18.4 Suppose that $\boldsymbol{f} \in \mathcal{F}^1(S)$. Prove that the backward Euler method is:

a) Consistent to exactly order $p = 1$.

b) Globally convergent with order $p = 1$.

18.5 Show that a sharper LTE estimate than the Proposition 18.7 estimate can be obtained. Specifically, under the same assumptions, prove that

$$|\mathcal{E}[u](t,s)| \leq \frac{s^2}{12} \max_{t\in[0,T]} |u'''(t)|.$$

18.6 Prove Corollary 18.8.

18.7 Suppose that $\boldsymbol{f} \in \mathcal{F}^2(S)$. Prove that the midpoint method is:

a) Consistent to exactly order $p = 2$.

b) Globally convergent with order $p = 2$.

18.8 Let $\theta \in [0,1]$. The θ-method is defined as

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau \boldsymbol{f}(t_k + (1-\theta)\tau, \theta \boldsymbol{w}^k + (1-\theta)\boldsymbol{w}^{k+1}).$$

Assuming that $\boldsymbol{f} \in \mathcal{F}^2(S)$, find the consistency order of this method and show that it is convergent.

Hint: The order of consistency depends on the value of θ.

18.9 Complete the proof of Theorem 18.11.

18.10 Consider the IVP

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

For $K \in \mathbb{N}$, set $\tau = \frac{1}{K}$. Apply the following methods to obtain approximations of $u(1) = e$.

a) Forward Euler method.

b) Taylor's method.

c) Heun's[4] method:

$$w^{k+1} = w^k + \frac{\tau}{2}\left[f(t_k, w^k) + f\big(t_{k+1}, w^k + \tau f(t_k, w^k)\big)\right].$$

d) Modified Euler's method:

$$w^{k+1} = w^k + \tau f\left[t_k + \frac{\tau}{2}, w^k + \frac{\tau}{2} f(t_k, w^k)\right].$$

For these approximations, show directly (without appealing to any convergence theorems) that $w^K \to u(1) = e$ as $K \to \infty$.

[4] Named in honor of the German mathematician Karl Heun (1859–1929).

https://doi.org/10.1017/9781108942607.024 Published online by Cambridge University Press