Classical Numerical Analysis, Chapter 18

Abner J. Salgado and Steven M. Wise

asalgad1@utk.edu swise1@.utk.edu
University of Tennessee

Chapter 18
Single Step Methods

Euler's Method

In this chapter we begin the study of approximation methods for initial value problems. The approach we will follow is motivated by definition of the mild solution. Indeed, if $\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t))$ and $\mathbf{u}(t_0) = \mathbf{u}_0$, we can approximate $\mathbf{u}(t)$ via

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

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

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

Then, the approximation becomes,

$$\mathbf{u}(t) \approx \mathbf{u}_0 + (t - t_0)\mathbf{f}(t_0, \mathbf{u}(t_0)), \quad \Longrightarrow \quad \frac{1}{t - t_0}\left(\mathbf{u}(t) - \mathbf{u}_0\right) \approx \mathbf{f}(t_0, \mathbf{u}(t_0)).$$

This is Euler's famous approximation method.

Notation and Assumptions

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 $d \in \mathbb{N}$ and $\mathbf{u}_0 \in \mathbb{R}^d$.

We will assume throughout this and the next few chapters that the slope function satisfies

,

$$\mathbf{f} \in F^1(S), \quad \text{where} \quad S = [0, T] \times \mathbb{R}^d,$$

which implies that $\mathbf{f}$ is globally $\mathbf{u}$-Lipschitz continuous. This simplification guarantees the existence of a classical solution. More importantly, it allows us to apply the Lipschitz estimate, with a single constant L, with either the solution values or with the approximate solution values, without worrying about whether those values are in some bounded open set Ω. This simplifying assumption can often be lifted by proving that the true approximate solutions lie in a certain bounded set.

Notation and Assumptions

We denote by $\mathbf{u} \in C^1([0, T]; \Omega)$ a classical solution on $[0, T]$ to the initial value problem

$$\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t)), \quad \mathbf{u}(0) = \mathbf{u}_0. \tag{1}$$

That such a solution exists and is unique follows from our assumptions on the slope function $\mathbf{f}$.

Approximations at Discrete Points

The methods we will present here do not give us a function that approximates $\mathbf{u}$ but rather a sequence of vectors that approximates this function at a particular collection of points in time. More precisely, we let $K \in \mathbb{N}$, and we define

$$\tau = \frac{T}{K},$$

which we call the *time step size*, and define

$$t_k = k\tau.$$

We will then produce a finite sequence $\{\mathbf{w}^k\}_{k=0}^K \subset \mathbb{R}^d$ such that

$$\mathbf{w}^k \approx \mathbf{u}(t_k).$$

Single-Step Approximation Methods

Definition (single–step method)

The finite sequence $\left\{\mathbf{w}^k\right\}_{k=0}^{K} \subset \mathbb{R}^d$, *is called a* **single–step approximation** *to* $\mathbf{u}$ *iff* $\mathbf{w}^0 = \mathbf{u}_0$ *and*

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \tau \mathbf{G}\left(t_k, \tau, \mathbf{w}^k, \mathbf{w}^{k+1}\right), \quad k = 0, \ldots, K-1, \tag{2}$$

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

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

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

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

Example

The *forward (or explicit) Euler method*:

$$\mathbf{G}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) = \mathbf{G}_{FE}\left(t, \mathbf{v}_1\right) = \mathbf{f}(t, \mathbf{v}_1). \tag{3}$$

Example

The *backward (or implicit) Euler method*:

$$\mathbf{G}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) = \mathbf{G}_{BE}\left(t, s, \mathbf{v}_2\right) = \mathbf{f}(t + s, \mathbf{v}_2). \tag{4}$$

Example

The *trapezoidal method*:

$$\mathbf{G}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) = \mathbf{G}_{TR}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) = \frac{1}{2}\mathbf{f}(t, \mathbf{v}_1) + \frac{1}{2}\mathbf{f}(t + s, \mathbf{v}_2). \tag{5}$$

Example

Taylor's method:

$$\begin{aligned} \mathbf{G}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) &= \mathbf{G}_{TM}\left(t, s, \mathbf{v}_1\right) \\ &= \mathbf{f}(t, \mathbf{v}_1) + \frac{s}{2}\left[\partial_t \mathbf{f}(t, \mathbf{v}_1) + D_{\mathrm{u}}\mathbf{f}(t, \mathbf{v}_1)\mathbf{f}(t, \mathbf{v}_1)\right], \end{aligned} \tag{6}$$

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

Example

The *midpoint method*:

$$\mathbf{G}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) = \mathbf{G}_{MR}\left(t, s, \mathbf{v}_1, \mathbf{v}_2\right) = \mathbf{f}\left(t + \frac{s}{2}, \frac{1}{2}\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2\right). \tag{7}$$

Definition (LTE and convergence)

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

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

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

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

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

$$\|\boldsymbol{\mathcal{E}}[\mathbf{u}](t,\tau)\|_2 \leq C\tau^p, \tag{8}$$

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 *iff* p *is the largest positive integer for which* (8) *holds regardless of how smooth the exact solution* $\mathbf{u}$ *is.*

Definition (LTE and convergence Cont.)

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

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

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

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

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

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

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

Consistency and Convergence

Lemma (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 $b > 1$, $c \geq 0$,

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

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

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

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 $k = 0, \ldots, n$.

Proof Cont.

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

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

which completes the induction argument. □

Proposition (Consistency of Forward Euler ($d = 1$))

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, for some $\eta \in (t - s, t)$,

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

Hence,

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

Proof Cont.

The result is proved using that

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

and taking

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

□

Corollary (Consistency of Forward Euler)

The forward Euler method (3) *is of exactly order* $p = 1$. *In other words, if* $\mathbf{f} \in \mathcal{F}^1(S)$ *and* $\mathbf{u} \in C^2([0, T]; \mathbb{R}^d)$ *is the unique solution to* (1), *then, for all* $\tau \in (0, T]$ *and* $t \in (\tau, T]$,

$$\|\boldsymbol{\mathcal{E}}[\mathbf{u}](t, \tau)\|_2 \leq C_{FE}\tau \tag{9}$$

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

Theorem (Convergence of Forward Euler)

Suppose that $\mathbf{f} \in \mathcal{F}^1(S)$. *Let* $L > 0$ *be its* $\mathbf{u}$*-Lipschitz constant on* S. *Suppose that the forward Euler method* (3) *is used to approximate* $\mathbf{u}$, *the unique solution to* (1). *Then, for all* $k = 0, \ldots, K$,

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

where $C_{FE} > 0$ *is the LTE constant from* (9). *Consequently,*

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

Proof.

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

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

for $k = 0, \ldots, K-1$ with $\mathbf{e}^0 = \mathbf{0}$.

Proof Cont.

Using the triangle inequality, the Lipschitz condition, and the LTE bound (9), we have, for $k = 0, \ldots, K-1$,

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

Using the discrete Grönwall inequality, we find, for $k = 0, \ldots, K$,

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

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

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

for all $k = 0, \ldots, K$. □

Proposition (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

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

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

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

Proof Cont.

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, 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

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

Proof Cont.

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

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

Corollary (Consistency of the Trapezoidal Method)

The trapezoidal method (5) *is of order exactly* $p = 2$. *Precisely, if* $\mathbf{f} \in \mathcal{F}^2(S)$, *then, for all* $\tau \in (0, T]$ *and* $t \in [\tau, T]$,

$$\|\boldsymbol{\mathcal{E}}[\mathbf{u}](t, \tau)\|_2 \leq C_{TR}\tau^2 \tag{10}$$

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

Theorem (Convergence of the Trapezoidal Method)

Let $\mathbf{f} \in \mathcal{F}^2(S)$ *and* $L > 0$ *be its* $\mathbf{u}$*-Lipschitz constant on* S*. Suppose that the trapezoidal method* (5) *is used to approximate the solution to* (1)*. Then, for all* $k = 0, \ldots, K$*, we have*

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

provided that $0 < \tau L < 1$*, where* $C_{TR} > 0$ *is the LTE constant from* (10)*.*

Proof.

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

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

Proof Cont.

We take norms and apply the triangle inequality, the **u**-Lipschitz condition on **f** and (10), to obtain

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

Since, by assumption, $\frac{1}{2} < 1-\frac{\tau L}{2} < 1$, the discrete Grönwall inequality then implies that

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

Proof Cont.

Finally, notice that

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

to, in conclusion, obtain that

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

as claimed. □

Theorem (Consistency of Taylor's Method)

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

$$\|\boldsymbol{\mathcal{E}}[\mathbf{u}](t, s)\|_2 \leq C_{TM} s^2 \tag{11}$$

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

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

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

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

Proof Cont.

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 (Convergence of Taylor's Method)

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

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

for all multi-indices $\boldsymbol{\alpha} \in \mathbb{N}^{d+1}$ *with* $|\boldsymbol{\alpha}| = 2$*, for all* $i = 1, \ldots, d$*, and for all* $(t, \mathbf{v}) \in S$*. (In other words, all second derivatives are bounded on* S*.) Then Taylor's method* (6) *is convergent and the global rate of convergence is* $p = 2$*. In particular,*

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

where $C_{TM} > 0$ *is the LTE constant from* (11) *and* $L' > 0$ *is a Lipschitz constant given below.*

Proof.

We need to establish an estimate of the form

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

for any $s \in (0, T]$, for any $t \in [s, T]$, and for all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$. The details are an exercise. □

Linear Slope Functions

Theorem (Convergence of Forward Euler for a Linear ODE System)

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

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

Let $K \in \mathbb{N}$. *Suppose that the sequence* $\left\{\mathbf{w}^k\right\}_{k=0}^K$ *is generated using the forward Euler method* (3). *Then, for all* $k = 0, \ldots, K$, *we have*

$$\left\|\mathbf{e}^k\right\|_2 \leq \|\mathbf{u}_0\|_2 \max_{\lambda\in\sigma(\mathsf{A})} \left|e^{\lambda k\tau} - (1+\lambda\tau)^k\right|. \tag{12}$$

Suppose that the maximum on the right-hand side of (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* $k = 1, 2, \ldots, K$,

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

Proof.

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

$$\mathbf{u}(t) = Qe^{tD}Q^\intercal \mathbf{u}_0,$$

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

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

Thus,

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

Proof Cont.

Taking norms, we get

$$\left\|\mathbf{e}^k\right\|_2 \leq \left\|\mathrm{Q}\left(e^{k\tau\mathrm{D}} - (\mathrm{I}+\tau\mathrm{D})^k\right)\mathrm{Q}^\intercal\right\|_2 \|\mathbf{u}_0\|_2 \leq \left\|e^{k\tau\mathrm{D}} - (\mathrm{I}+\tau\mathrm{D})^k\right\|_2 \|\mathbf{u}_0\|_2 .$$

Notice that $(\mathrm{I}+\tau\mathrm{D})^k$ 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,

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

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

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

Equivalently,

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

Proof Cont.

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

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

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

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

Proof Cont.

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. □