20 Linear Multi-step Methods

The fundamental formula that defines mild solutions

$$\boldsymbol{u}(t_2) = \boldsymbol{u}(t_1) + \int_{t_1}^{t_2} \boldsymbol{f}(s, \boldsymbol{u}(s))\mathrm{d}s$$

was used in the previous chapter with $t_1 = t_k$ and $t_2 = t_{k+1}$. To approximate the integral, we used information about the slope function in the time interval $[t_k, t_{k+1}]$ only. However, we could use more information to build an approximation of the integral. For instance, for some $q \in \mathbb{N}_0$, we set $X_q = \{t_{k-q}, \ldots, t_k, t_{k+1}\}$ and replace the slope function by its interpolant on X_q

$$\mathcal{I}_{X_q}[\boldsymbol{f}(\cdot, \boldsymbol{u}(\cdot))](t) \approx \boldsymbol{f}(t, \boldsymbol{u}(t)).$$

Then

$$\boldsymbol{u}(t_2) \approx \boldsymbol{u}(t_1) + \int_{t_1}^{t_2} \mathcal{I}_{X_q}[\boldsymbol{f}(\cdot, \boldsymbol{u}(\cdot))](s)\mathrm{d}s.$$

This is the basis of Adams-type methods and other multi-step methods, which we examine in this chapter.

Before we begin, we recall that we are trying to approximate the solution to (18.1) by choosing $K \in \mathbb{N}$ and letting $\tau = T/K$ and $t_k = k\tau$. We will produce $\{\boldsymbol{w}^k\}_{k=0}^K \subset \mathbb{R}^d$ such that $\boldsymbol{w}^k \approx \boldsymbol{u}(t_k)$.

20.1 Consistency of Linear Multi-step Methods

In Chapters 18 and 19, we discussed single-step methods. With those, all that was needed to obtain the approximation at time level $k+1$ was the approximate solution at time level k. In this chapter, we will introduce multi-step methods, which use approximations at additional past time levels, $k-1$, $k-2$, etc.

Definition 20.1 (linear multi-step method)**.** Let $K, q \in \mathbb{N}$ with $q < K$. The finite sequence $\left\{\boldsymbol{w}^k\right\}_{k=0}^K \subset \mathbb{R}^d$ is called a **linear** q**-step approximation** (or just a **linear multi-step approximation**) to $\boldsymbol{u}$, solution of (18.1), with starting values $\boldsymbol{w}^0, \boldsymbol{w}^1, \ldots, \boldsymbol{w}^{q-1}$ if and only if, for $k = 0, \ldots, K-q$,

$$\sum_{j=0}^{q} a_j \boldsymbol{w}^{k+j} = \tau \sum_{j=0}^{q} b_j \boldsymbol{f}(t_{k+j}, \boldsymbol{w}^{k+j}), \tag{20.1}$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

where $\{a_j\}_{j=0}^q, \{b_j\}_{j=0}^q \subseteq \mathbb{R}$. The multi-step approximation is called **explicit** if $b_q = 0$; otherwise, it is called **implicit**. As before, the **global error** of the multi-step approximation is the finite sequence $\left\{\boldsymbol{e}^k\right\}_{k=0}^K \subseteq \mathbb{R}^d$ defined via

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

Remark 20.2 (convention)**.** Notice that, for the multi-step method (20.1) to make sense, we must have $a_q \neq 0$. In addition, observe that the coefficients $\{a_j\}_{j=0}^q, \{b_j\}_{j=0}^q$ are defined up to multiplication by a common constant. Because of these two considerations, here and in what follows we will assume that

$$a_q = 1.$$

Definition 20.3 (LTE and consistency)**.** Let $\boldsymbol{u} \in C^1([0,T];\Omega)$ be a classical solution on $[0,T]$ to (18.1). Let the sequence $\left\{\boldsymbol{w}^k\right\}_{k=0}^K$ be obtained with the q-step method (20.1) with starting values $\boldsymbol{w}^0, \boldsymbol{w}^1, \ldots, \boldsymbol{w}^{q-1}$. The **local truncation error** (LTE) or **consistency error** of the multi-step approximation is defined as

$$\begin{aligned}&\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,\tau)\\ &\quad= \frac{1}{\tau}\sum_{j=0}^{q}\left[a_j\boldsymbol{u}(t+(j-q)\tau) - \tau b_j \boldsymbol{f}(t+(j-q)\tau, \boldsymbol{u}(t+(j-q)\tau))\right] \qquad (20.2)\end{aligned}$$

for any $t \in [t_q, T]$. We make frequent use of the notation $\boldsymbol{\mathcal{E}}^k[\boldsymbol{u}] = \boldsymbol{\mathcal{E}}[\boldsymbol{u}](t_k,\tau)$ for $k = q, \ldots, K$. We say that the linear multi-step approximation is **consistent to at least order** $p \in \mathbb{N}$ if and only if, when

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

there is a constant $\tau_0 \in (0,T]$ and a constant $C > 0$ such that, for all $\tau \in (0,\tau_0]$ and all $t \in [t_q, T]$,

$$\|\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,\tau)\|_2 \leq C\tau^p. \qquad (20.3)$$

We say that the linear q-step approximation is **consistent to exactly order** p if and only if p is the largest positive integer for which (20.3) holds.

We say that the multi-step approximation **converges globally**, with at least order $p \in \mathbb{N}$, if and only if, when

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

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

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

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

The following result can be used to determine the order of a linear multi-step method.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Theorem 20.4 (method of C's)**.** *Let $\boldsymbol{f} \in \mathcal{F}^p(S)$ and $\boldsymbol{u} \in C^{p+1}([0,T];\mathbb{R}^d)$ be a classical solution on $[0,T]$ to* (18.1)*. Suppose that $\boldsymbol{u}$ is approximated by the linear q-step method* (20.1)*. Define*

$$C_m = \begin{cases} \sum_{j=0}^{q} a_j, & m = 0, \\ \sum_{j=0}^{q} \left(\frac{j^m}{m!} a_j - \frac{j^{m-1}}{(m-1)!} b_j \right), & m \in \{1,2,3,\dots\}, \end{cases} \tag{20.4}$$

with the convention that $0^0 = 1$. The method is consistent to exactly order p if and only if $C_0 = 0 = C_1 = \cdots = C_p$, but $C_{p+1} \neq 0$.

Proof. For simplicity of notation, let us suppose that $d = 1$. Consider $t \in [t_q, T]$ and we extend, for any $k \in \mathbb{Z}$, the definition $t_k = k\tau$. Using Taylor's Theorem, with the expansion point $t - t_q$, one finds, for each $j = 0, 1, \dots, q$,

$$u(t + t_{j-q}) = \sum_{m=0}^{p} u^{(m)}(t - t_q) \frac{(j\tau)^m}{m!} + u^{(p+1)}(\zeta_j) \frac{(j\tau)^{p+1}}{(p+1)!}$$

and

$$\begin{aligned} u'(t + t_{j-q}) &= \sum_{m=0}^{p-1} u^{(m+1)}(t - t_q) \frac{(j\tau)^m}{m!} + u^{(p+1)}(\xi_j) \frac{(j\tau)^p}{p!} \\ &= \sum_{m=1}^{p} u^{(m)}(t - t_q) \frac{(j\tau)^{m-1}}{(m-1)!} + u^{(p+1)}(\xi_j) \frac{(j\tau)^p}{p!}. \end{aligned}$$

Observe that the $j = 0$ case holds if we agree that $0^0 = 1$. Therefore,

$$\begin{aligned} \tau \mathcal{E}[u](t,\tau) &= \sum_{j=0}^{q} a_j u(t + t_{j-q}) - \tau \sum_{j=0}^{q} b_j u'(t + t_{j-q}) \\ &= \sum_{j=0}^{q} a_j \sum_{m=0}^{p} u^{(m)}(t - t_q) \frac{(j\tau)^m}{m!} - \tau \sum_{j=0}^{q} b_j \sum_{m=1}^{p} u^{(m)}(t - t_q) \frac{(j\tau)^{m-1}}{(m-1)!} \\ &\quad + \tau^{p+1} \sum_{j=0}^{q} \left[a_j u^{(p+1)}(\zeta_j) \frac{j^{p+1}}{(p+1)!} - b_j u^{(p+1)}(\xi_j) \frac{j^p}{p!} \right]. \end{aligned}$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Interchanging the summations, so that we sum by powers of τ, we have

$$\begin{aligned}\tau\mathcal{E}[u](t,\tau) &= u(t-t_q)\sum_{j=0}^{q} a_j + \sum_{m=1}^{p}\tau^m u^{(m)}(t-t_q)\sum_{j=0}^{q}\left[a_j\frac{j^m}{m!} - b_j\frac{j^{m-1}}{(m-1)!}\right] \\ &\quad + \tau^{p+1}\sum_{j=0}^{q}\left[a_j u^{(p+1)}(\zeta_j)\frac{j^{p+1}}{(p+1)!} - b_j u^{(p+1)}(\xi_j)\frac{j^p}{p!}\right] \\ &= C_0 u(t-t_q) + \sum_{m=1}^{p} C_m\tau^m u^{(m)}(t-t_q) \\ &\quad + \tau^{p+1}\sum_{j=0}^{q}\left[a_j u^{(p+1)}(\zeta_j)\frac{j^{p+1}}{(p+1)!} - b_j u^{(p+1)}(\xi_j)\frac{j^p}{p!}\right]\end{aligned} \tag{20.5}$$

for some constants $\zeta_j, \xi_j \in [t-t_q, t]$, $j = 0, \dots, q$.

($\Longrightarrow$) Suppose that the method is of exactly order p. Then, from (20.5), we must have $C_0 = C_1 = \cdots = C_p = 0$. If the true solution has higher regularity, say $u \in C^{p+2}([0,T])$, then we can extend the Taylor expansion by one term to obtain

$$\begin{aligned}\tau\mathcal{E}[u](t,\tau) &= C_{p+1}\tau^{p+1}u^{(p+1)}(t-t_q) \\ &\quad + \tau^{p+2}\sum_{j=0}^{q}\left[a_j u^{(p+2)}(\tilde{\zeta}_j)\frac{j^{p+2}}{(p+2)!} - b_j u^{(p+2)}(\tilde{\xi}_j)\frac{j^{p+1}}{(p+1)!}\right].\end{aligned}$$

Since the method does not exceed order p, it must be true that $C_{p+1} \neq 0$ generically, by the definition of the local truncation error.

($\Longleftarrow$) Suppose that $C_0 = C_1 = \cdots = C_p = 0$, but $C_{p+1} \neq 0$. Then

$$\mathcal{E}[u](t,\tau) = \tau^p\sum_{j=0}^{q}\left[a_j u^{(p+1)}(\zeta_j)\frac{j^{p+1}}{(p+1)!} - b_j u^{(p+1)}(\xi_j)\frac{j^p}{p!}\right]$$

and the method is consistent to at least order p. Since $C_{p+1} \neq 0$, the order of accuracy cannot exceed p, even if the true solution has higher regularity, say $u \in C^{p+2}([0,T])$. □

The consistency criterion given in Theorem 20.4 is usually known as the *method of C's*. The algorithmic description of the computation of each one of the involved C's is given in Listing 20.1.

Definition 20.5 (characteristic polynomials)**.** For the linear q-step method (20.1), we define the **first** and **second characteristic polynomials**, respectively, as

$$\psi(z) = \sum_{j=0}^{q} a_j z^j \in \mathbb{P}_q, \qquad \chi(z) = \sum_{j=0}^{q} b_j z^j \in \mathbb{P}_q.$$

Corollary 20.6 (first-order consistency)**.** *Let* $\boldsymbol{f} \in \mathcal{F}^1(S)$*. Assume that the function* $\boldsymbol{u} \in C^2([0,T];\mathbb{R}^d)$ *is a classical solution to the initial value problem (IVP)* (18.1).

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Suppose that $\boldsymbol{u}$ *is approximated by the linear q-step method* (20.1). *The method is consistent to at least first order if and only if*

$$\psi(1) = 0, \qquad \psi'(1) - \chi(1) = 0.$$

Proof. This follows from Theorem 20.4; see Problem 20.1. □

The following result provides another way to verify the consistency of a linear multi-step method.

Theorem 20.7 (the log-method). *Let* $\boldsymbol{f} \in \mathcal{F}^p(S)$. *Assume that the function* $\boldsymbol{u} \in C^{p+1}([0,T];\mathbb{R}^d)$ *is a classical solution to the IVP* (18.1). *Suppose that* $\boldsymbol{u}$ *is approximated by the linear q-step method* (20.1). *The method is consistent to exactly order p if and only if the function*

$$\phi(\mu) = \frac{\psi(\mu)}{\ln(\mu)} - \chi(\mu),$$

which is complex analytic in a neighborhood of $\mu = 1$, *has the property that* $\mu = 1$ *is a p-fold zero or, equivalently, that*

$$\tilde{\phi}(\mu) = \psi(\mu) - \chi(\mu)\ln(\mu),$$

which is also complex analytic in a neighborhood of $\mu = 1$, *has the property that* $\mu = 1$ *is a* $p+1$*-fold zero.*

Proof. Once again, for simplicity of notation, we consider $d = 1$ and assume that u is real analytic. From Theorem 20.4, we observe that the method is consistent to order p if and only if $C_0 = C_1 = \cdots = C_p = 0$, but $C_{p+1} \neq 0$. Using the techniques developed in the proof of Theorem 20.4, we can expand to all orders to obtain

$$\tau\mathcal{E}[u](t,\tau) = \sum_{m=p+1}^{\infty} C_m u^{(m)}(t - t_q)\tau^m.$$

In particular, setting $u(t) = \exp(t)$, which is certainly real analytic, we find

$$\tau\mathcal{E}[\exp(\cdot)](t,\tau) = \exp(t - t_q)\sum_{m=p+1}^{\infty} C_m\tau^m.$$

On the other hand, by the definition of the local truncation error, we have

$$\begin{aligned}
\tau\mathcal{E}[\exp(\cdot)](t,\tau) &= \exp(t - t_q)\sum_{j=0}^{q}\{a_j\exp(t_j) - \tau b_j\exp(t_j)\} \\
&= \exp(t - t_q)\sum_{j=0}^{q}\{a_j\exp(j\tau) - \tau b_j\exp(j\tau)\} \\
&= \exp(t - t_q)\,[\psi(\exp(\tau)) - \tau\chi(\exp(\tau))].
\end{aligned}$$

Equating terms, we have

$$\tilde{\phi}(\exp(\tau)) = \tau\phi(\exp(\tau)) = \psi(\exp(\tau)) - \tau\chi(\exp(\tau)) = \sum_{m=p+1}^{\infty} C_m\tau^m.$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Thus, $\tau = 0$ is a p-fold zero of the function $\phi(\exp(\tau))$. Here, in fact, we can assume that τ is any complex number. Setting $\mu = \exp(\tau)$ and using the fact that, in a neighborhood of $\mu = 1$,

$$\begin{aligned}\ln(\mu) &= \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}(\mu-1)^m \\ &= (\mu-1) - \frac{1}{2}(\mu-1)^2 + \frac{1}{3}(\mu-1)^3 - \frac{1}{4}(\mu-1)^4 + \frac{1}{5}(\mu-1)^5 + \cdots,\end{aligned}$$

it follows that this is equivalent to the condition that $\phi(\mu)$ has a p-fold zero at $\mu = 1$, which is equivalent to the condition that $\tilde{\phi}(\mu)$ has a $p+1$-fold zero at $\mu = 1$. □

The consistency criterion given in Theorem 20.7 is usually known as the *log-method*.

Example 20.1 Consider the linear two-step method

$$\boldsymbol{w}^{k+2} - \boldsymbol{w}^k = \frac{\tau}{3}\left[\boldsymbol{f}(t_{k+2}, \boldsymbol{w}^{k+2}) + 4\boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1}) + \boldsymbol{f}(t_k, \boldsymbol{w}^k)\right].$$

This method is implicit and consistent to exactly order $p = 4$. To prove this, assuming that the slope function is sufficiently regular, $\boldsymbol{f} \in \mathcal{F}^4(S)$, we need only to show that $C_0 = C_1 = \cdots = C_4 = 0$, but $C_5 \neq 0$, where these constants are defined in (20.4). Clearly, $C_0 = 0$. Now

$$\begin{aligned}C_1 &= \sum_{j=0}^{2}(ja_j - b_j) = 2\cdot 1 + 1\cdot 0 + 0\cdot(-1) - \left(\frac{1}{3} + \frac{4}{3} + \frac{1}{3}\right) = 2 - 2 = 0, \\ C_2 &= \sum_{j=0}^{2}\left(\frac{j^2}{2!}a_j - jb_j\right) = \frac{2^2}{2}\cdot 1 - \left(2\cdot\frac{1}{3} + 1\cdot\frac{4}{3}\right) = 2 - 2 = 0, \\ C_3 &= \sum_{j=0}^{2}\left(\frac{j^3}{6}a_j - \frac{j^2}{2}b_j\right) = \frac{2^3}{6}\cdot 1 - \left(\frac{2^2}{2}\cdot\frac{1}{3} + \frac{1^2}{2}\cdot\frac{4}{3}\right) = \frac{8}{6} - \frac{8}{6} = 0, \\ C_4 &= \sum_{j=0}^{2}\left(\frac{j^4}{24}a_j - \frac{j^3}{6}b_j\right) = \frac{2^4}{24}\cdot 1 - \left(\frac{2^3}{6}\cdot\frac{1}{3} + \frac{1^3}{6}\cdot\frac{4}{3}\right) = \frac{2}{3} - \frac{2}{3} = 0.\end{aligned}$$

But

$$C_5 = \sum_{j=0}^{2}\left(\frac{j^5}{120}a_j - \frac{j^4}{24}b_j\right) = \frac{2^5}{120}\cdot 1 - \left(\frac{2^4}{24}\cdot\frac{1}{3} + \frac{1^4}{24}\cdot\frac{4}{3}\right) = \frac{4}{15} - \frac{5}{18} \neq 0.$$

Example 20.2 In this example, we use Theorem 20.7 to show that the two-step implicit method

$$\boldsymbol{w}^{k+2} - \boldsymbol{w}^{k+1} = \tau\left[\frac{5}{12}\boldsymbol{f}(t_{k+2}, \boldsymbol{w}^{k+2}) + \frac{8}{12}\boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1}) - \frac{1}{12}\boldsymbol{f}(t_k, \boldsymbol{w}^k)\right]$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

is consistent to exactly order $p = 3$. To do so, it is convenient to make the change of variables $z = \mu - 1$. Then

$$\begin{aligned}
\psi(\mu) &= \mu^2 - \mu \\
&= (z+1)^2 - (z+1) \\
&= z^2 + z, \\
\chi(\mu) &= \frac{5}{12}\mu^2 + \frac{8}{12}\mu - \frac{1}{12} \\
&= \frac{5}{12}(z+1)^2 + \frac{8}{12}(z+1) - \frac{1}{12} \\
&= \frac{5}{12}z^2 + \frac{3}{2}z + 1,
\end{aligned}$$

and

$$\begin{aligned}
\ln(\mu) &= (\mu-1) - \frac{1}{2}(\mu-1)^2 + \frac{1}{3}(\mu-1)^3 - \frac{1}{4}(\mu-1)^4 + \frac{1}{5}(\mu-1)^5 + \cdots \\
&= z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4 + \frac{1}{5}z^5 + \cdots .
\end{aligned}$$

We need to consider the difference

$$\frac{\psi(\mu)}{\ln(\mu)} - \chi(\mu).$$

To do so, we first find an expansion, in terms of z, for $\frac{\psi(\mu)}{\ln(\mu)}$:

$$\frac{\psi(\mu)}{\ln(\mu)} = c_0 + c_1 z + c_2 z^2 + c_3 z^3 + c_4 z^4 + \cdots .$$

Then

$$\left(c_0 + c_1 z + c_2 z^2 + c_3 z^3 + \cdots\right)\left(z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4 + \cdots\right) = z + z^2,$$

which implies that

$$\begin{aligned}
c_0 &= 1, & & \\
c_1 - \frac{1}{2}c_0 &= 1 & \implies\ & c_1 = \frac{3}{2}, \\
c_2 - \frac{1}{2}c_1 + \frac{1}{3}c_0 &= 0 & \implies\ & c_2 = \frac{5}{12}, \\
c_3 - \frac{1}{2}c_2 + \frac{1}{3}c_1 - \frac{1}{4}c_0 &= 0 & \implies\ & c_3 = -\frac{1}{24}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{\psi(\mu)}{\ln(\mu)} - \chi(\mu) &= 1 + \frac{3}{2}z + \frac{5}{12}z^2 - \frac{1}{24}z^3 + c_4 z^4 + \cdots - \left(1 + \frac{3}{2}z + \frac{5}{12}z^2\right) \\
&= -\frac{1}{24}z^3 + c_4 z^4 + \cdots ,
\end{aligned}$$

which proves that the method is of exactly third order.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

20.2 Adams–Bashforth and Adams–Moulton Methods

In this section, we derive some examples of the so-called Adams–Moulton and Adams–Bashforth multi-step methods. We make extensive use of the Lagrange interpolation techniques of Chapter 9. Let $q < K$. Assume that we have computed $k+q-1 < K$ approximations $\{\boldsymbol{w}^j\}_{j=0}^{k+q-1}$. Now, from the definition of mild solution (17.2), we have that

$$\boldsymbol{u}(t_{k+q}) - \boldsymbol{u}(t_{q+k-1}) = \int_{t_{k+q-1}}^{t_{k+q}} \boldsymbol{f}(s, \boldsymbol{u}(s))\mathrm{d}s.$$

We could approximate this integral using the values $\boldsymbol{w}^k$ and $\boldsymbol{w}^{k+1}$, and this is the idea behind the single-step methods studied in Chapter 18. However, in doing so, we are not making use of all the information that we had computed before, i.e., $\{\boldsymbol{w}^j\}_{j=0}^{k+q-1}$. The idea of the *Adams methods* is to use a subset of $\{\boldsymbol{f}(t_j, \boldsymbol{w}^j)\}_{j=0}^{k+q-1}$ to construct an interpolating polynomial and use this polynomial to approximate the integral in the previous identity. Two important classes of methods here are:

1. *Adams–Bashforth* methods,[1] which use $\{\boldsymbol{f}(t_j, \boldsymbol{w}^j)\}_{j=k}^{k+q-1}$ and thus are *explicit* methods.
2. *Adams–Moulton* methods,[2] which use $\{\boldsymbol{f}(t_j, \boldsymbol{w}^j)\}_{j=k}^{k+q}$ and thus are *implicit*.

The general strategy is clear, so we confine ourselves to presenting a few examples.

Example 20.3 The *Adams–Bashforth four-step method* (AB4) is defined as follows: for $k = 0, \dots, K-4$,

$$\boldsymbol{w}^{k+4} - \boldsymbol{w}^{k+3} = \tau \left[\frac{55}{24}\boldsymbol{f}^{k+3} - \frac{59}{24}\boldsymbol{f}^{k+2} + \frac{37}{24}\boldsymbol{f}^{k+1} - \frac{9}{24}\boldsymbol{f}^{k} \right], \tag{20.6}$$

where $\boldsymbol{f}^j = \boldsymbol{f}(t_j, \boldsymbol{w}^j)$. This requires the starting values $\boldsymbol{w}^0$, $\boldsymbol{w}^1$, $\boldsymbol{w}^2$, $\boldsymbol{w}^3$. The coefficients are $a_4 = 1$, $a_3 = -1$, $a_2 = a_1 = a_0 = 0$, and $b_4 = 0$, $b_3 = \frac{55}{24}$, $b_2 = -\frac{59}{24}$, $b_1 = \frac{37}{24}$, $b_0 = -\frac{9}{24}$.

Theorem 20.8 (LTE of AB4)**.** *Suppose that* $\boldsymbol{f} \in \mathcal{F}^4(S)$ *and* $\boldsymbol{u} \in C^5([0,T];\mathbb{R}^d)$ *is the classical solution to* (18.1)*. Then the local truncation error for the AB4 method* (20.6) *may be expressed as*

$$\boldsymbol{\mathcal{E}}^{k+4}[\boldsymbol{u}] \le \frac{251 d^{1/2}}{720} \max_{\eta \in [t_k, t_{k+4}]} \|\boldsymbol{u}^{(5)}(\eta)\|_2 \tau^4$$

for every $k = 0, 1, \dots, K-4$.

[1] Named in honor of the British mathematicians John Couch Adams (1819–1892) and Francis Bashforth (1819–1912).

[2] Named in honor of the British mathematician John Couch Adams (1819–1892) and American astronomer Forest Ray Moulton (1872–1952).

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Proof. Suppose that $0 \le k \le K-4$. Let $\boldsymbol{p}_k \in [\mathbb{P}_3]^d$ be the vector-valued Lagrange interpolating polynomial with respect to the four interpolation points

$$\{(t_{k+j}, \boldsymbol{f}(t_{k+j}, \boldsymbol{u}(t_{k+j})))\}_{j=0}^{3}.$$

Then, for all $t \in [t_k, t_{k+4}]$,

$$\boldsymbol{f}(t, \boldsymbol{u}(t)) = \boldsymbol{p}_k(t) + \boldsymbol{E}_k(t),$$

where $\boldsymbol{E}_k$ is an error function. Thus,

$$\boldsymbol{u}(t_{k+4}) - \boldsymbol{u}(t_{k+3}) = \int_{t_{k+3}}^{t_{k+4}} \boldsymbol{f}(t, \boldsymbol{u}(t))\mathrm{d}t = \int_{t_{k+3}}^{t_{k+4}} \boldsymbol{p}_k(t)\mathrm{d}t + \int_{t_{k+3}}^{t_{k+4}} \boldsymbol{E}_k(t)\mathrm{d}t.$$

According to the error theory for Lagrange interpolation (Theorem 9.16) we have, for all $i = 1, \dots, d$,

$$[\boldsymbol{E}_k]_i(t) = \frac{1}{4!}\frac{\mathrm{d}^4}{\mathrm{d}t^4}[\boldsymbol{f}(t, \boldsymbol{u}(t))]_{i|t=\xi_i} \prod_{j=0}^{3}(t - t_{k+j}) = \frac{1}{24}[\boldsymbol{u}^{(5)}]_i(\xi_i(t)) \prod_{j=0}^{3}(t - t_{k+j})$$

for all $t \in [t_k, t_{k+4}]$ and some $\xi_i = \xi_i(t) \in (t_k, t_{k+4})$. We find then that, after the change of variables $t = r\tau + t_{k+3}$,

$$\int_{t_{k+3}}^{t_{k+4}} [\boldsymbol{E}_k]_i(t)\mathrm{d}t = \frac{\tau^5}{24}\int_0^1 r(r+1)(r+2)(r+3)[\boldsymbol{u}^{(5)}]_i(\xi_i(t))\mathrm{d}r.$$

Evidently, $\boldsymbol{E}_k$ is a continuous function on $[0, T]$ since it is the difference of continuous functions. Now set

$$g(r) = r(r+1)(r+2)(r+3)$$

and observe that $g \ge 0$ on $[0, 1]$. Then, taking norms,

$$\begin{aligned}
\left\| \int_{t_{k+3}}^{t_{k+4}} \boldsymbol{E}_k(t)\mathrm{d}t \right\|_2 &\le \frac{\tau^5}{24}\int_0^1 g(r)\left(\sum_{i=1}^{d}\left|[\boldsymbol{u}^{(5)}]_i(\xi_i(t))\right|^2\right)^{1/2}\mathrm{d}r \\
&\le d^{1/2}\max_{\xi\in[t_k,t_{k+4}]}\left\|\boldsymbol{u}^{(5)}(\xi)\right\|_2 \frac{\tau^5}{24}\int_0^1 g(r)\mathrm{d}r \\
&= \max_{\xi\in[t_k,t_{k+4}]}\left\|\boldsymbol{u}^{(5)}(\xi)\right\|_2 \frac{\tau^5}{24}\frac{251d^{1/2}}{30} \\
&= \frac{251d^{1/2}}{720}\tau^5 \max_{\xi\in[t_k,t_{k+4}]}\left\|\boldsymbol{u}^{(5)}(\xi)\right\|_2 .
\end{aligned}$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Let us use the notation $\boldsymbol{f}_e^{k+j} = \boldsymbol{f}(t_{k+j}, \boldsymbol{u}(t_{k+j}))$, $j = 0, 1, 2, 3$. Then

$$\begin{aligned}\boldsymbol{p}_k(t) = \boldsymbol{f}_e^k &\frac{(t - t_{k+1})(t - t_{k+2})(t - t_{k+3})}{(t_k - t_{k+1})(t_k - t_{k+2})(t_k - t_{k+3})} \\ &+ \boldsymbol{f}_e^{k+1}\frac{(t - t_k)(t - t_{k+2})(t - t_{k+3})}{(t_{k+1} - t_k)(t_{k+1} - t_{k+2})(t_{k+1} - t_{k+3})} \\ &+ \boldsymbol{f}_e^{k+2}\frac{(t - t_k)(t - t_{k+1})(t - t_{k+3})}{(t_{k+2} - t_k)(t_{k+2} - t_{k+1})(t_{k+2} - t_{k+3})} \\ &+ \boldsymbol{f}_e^{k+3}\frac{(t - t_k)(t - t_{k+1})(t - t_{k+2})}{(t_{k+3} - t_k)(t_{k+3} - t_{k+1})(t_{k+3} - t_{k+2})}.\end{aligned}$$

Observe that

$$\boldsymbol{p}_k(t_{k+j}) = \boldsymbol{f}_e^{k+j} = \boldsymbol{f}(t_{k+j}, \boldsymbol{u}(t_{k+j})), \quad j = 0, 1, 2, 3.$$

Integrating $\boldsymbol{p}_k$ on the interval $[t_{k+3}, t_{k+4}]$, the reader can confirm that

$$\int_{t_{k+3}}^{t_{k+4}} \boldsymbol{p}_k(t)\mathrm{d}t = \tau\left[\frac{55}{24}\boldsymbol{f}_e^{k+3} - \frac{59}{24}\boldsymbol{f}_e^{k+2} + \frac{37}{24}\boldsymbol{f}_e^{k+1} - \frac{9}{24}f_e^k\right].$$

Using the definition of the local truncation error (20.2), the result is proven. □

Remark 20.9 (consistency)**.** With the same hypotheses as in Theorem 20.8, we can use the result of Theorem 20.4 to come to the same conclusion as above. In particular, for the AB4 method, we find

$$C_0 = C_1 = C_2 = C_3 = C_4 = 0, \qquad C_5 = \frac{251}{720}.$$

Example 20.4 The *Adams–Moulton four-step method* (AM4) is defined as follows: for $k = 0, \ldots, K - 4$,

$$\begin{aligned}\boldsymbol{w}^{k+4} - \boldsymbol{w}^{k+3} = \tau\Bigg[&\frac{251}{720}\boldsymbol{f}^{k+4} + \frac{646}{720}\boldsymbol{f}^{k+3} - \frac{264}{720}\boldsymbol{f}^{k+2} \\ &+\frac{106}{720}\boldsymbol{f}^{k+1} - \frac{19}{720}\boldsymbol{f}^k\Bigg],\end{aligned} \tag{20.7}$$

where $\boldsymbol{f}^j = \boldsymbol{f}(t_j, \boldsymbol{w}^j)$. This requires the starting values $\boldsymbol{w}^0$, $\boldsymbol{w}^1$, $\boldsymbol{w}^2$, $\boldsymbol{w}^3$. The coefficients are $a_4 = 1$, $a_3 = -1$, $a_2 = a_1 = a_0 = 0$, and $b_4 = \frac{251}{720}$, $b_3 = \frac{646}{720}$, $b_2 = -\frac{264}{720}$, $b_1 = \frac{106}{720}$, $b_0 = -\frac{19}{720}$.

As we will see in the proof of the following result, the difference between the AB4 and AM4 methods is that the interpolating polynomial in the AM4 method uses the additional implicit time level point $\boldsymbol{f}(t_{k+4}, \boldsymbol{u}(t_{k+4}))$. This results in the AM4 method being more accurate, but comes at the cost of producing an implicit method. Implicit methods are usually always more complicated in practice than explicit methods.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Theorem 20.10 (LTE for AM4)**.** *Assume that* $\boldsymbol{f} \in \mathcal{F}^5(S)$*. Let the function* $\boldsymbol{u} \in C^6([0,T];\mathbb{R}^d)$ *be the classical solution to* (18.1)*. Then the local truncation error for the AM4 method* (20.7) *may be expressed as*

$$\|\boldsymbol{\mathcal{E}}^{k+4}[\boldsymbol{u}]\|_2 \le \frac{3d^{1/2}}{160} \max_{\eta\in[t_k,t_{k+4}]} \left\|\boldsymbol{u}^{(6)}(\eta)\right\|_2 \tau^5$$

for every $k = 0, 1, \dots, K-4$.

Proof. The proof is similar to that of Theorem 20.8. Suppose that $0 \le k \le K-4$. Let $\boldsymbol{p}_k \in [\mathbb{P}_4]^d$ be the vector-valued Lagrange interpolating polynomial uniquely determined by the five interpolation points

$$\{(t_{k+j}, \boldsymbol{f}(t_{k+j}, \boldsymbol{u}(t_{k+j})))\}_{j=0}^4 .$$

Then, for all $t \in [t_k, t_{k+4}]$,

$$\boldsymbol{f}(t, \boldsymbol{u}(t)) = \boldsymbol{p}_k(t) + \boldsymbol{E}_k(t)$$

and, for all $i = 1, \dots, d$,

$$[\boldsymbol{E}_k]_i(t) = \frac{1}{5!}\frac{\mathrm{d}^5}{\mathrm{d}t^5}[\boldsymbol{f}(t,\boldsymbol{u}(t))]_{i|t=\xi}\prod_{j=0}^4(t-t_{k+j}) = \frac{1}{120}[\boldsymbol{u}^{(6)}]_i(\xi_i(t))\prod_{j=0}^4(t-t_{k+j})$$

for all $t \in [t_k, t_{k+4}]$ and some $\xi_i = \xi_i(t) \in (t_k, t_{k+4})$. After the change of variables $t = r\tau + t_{k+3}$,

$$\int_{t_{k+3}}^{t_{k+4}} [\boldsymbol{E}_k]_i(t)\mathrm{d}t = \frac{\tau^6}{120}\int_0^1 (r-1)r(r+1)(r+2)(r+3)[\boldsymbol{u}^{(6)}]_i(\xi_i(t))\mathrm{d}r.$$

Setting

$$g(r) = -(r-1)r(r+1)(r+2)(r+3),$$

we observe that $g \ge 0$ on $[0,1]$ and, as before,

$$\left\|\int_{t_{k+3}}^{t_{k+4}} \boldsymbol{E}_k(t)\mathrm{d}t\right\|_2 \le \frac{3d^{1/2}}{160}\tau^6 \max_{\eta\in[t_k,t_{k+4}]}\left\|\boldsymbol{u}^{(6)}(\eta)\right\|_2 .$$

For the Lagrange interpolating polynomial, the reader can confirm that

$$\int_{t_{k+3}}^{t_{k+4}} \boldsymbol{p}_k(t)\mathrm{d}t = \tau\left[\frac{251}{720}\boldsymbol{f}_e^{k+4} + \frac{646}{720}\boldsymbol{f}_e^{k+3} - \frac{264}{720}\boldsymbol{f}_e^{k+2} + \frac{106}{720}\boldsymbol{f}_e^{k+1} - \frac{19}{720}\boldsymbol{f}_e^{k}\right],$$

where $\boldsymbol{f}_e^{k+j} = \boldsymbol{f}(t_{k+j}, \boldsymbol{u}(t_{k+j}))$, $j = 0, \dots, 4$. The result follows from the definition of the local truncation error (20.2). □

20.3 Backward Differentiation Formula Methods

Another important class of multi-step methods is the Backward Differentiation Formula (BDF) methods. These differ from the Adams–Bashforth (AB) and Adams–Moulton (AM) methods in a fundamental way. Whereas the AB and AM

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

methods are derived via an integration procedure, the BDF methods are derived via differentiation. We will demonstrate this point shortly. But before that, we give a general definition for these methods and derive some of their properties.

Definition 20.11 (BDF)**.** A linear q-step method (20.1) is called a **BDF method** (or a **BDFq method**) if and only if it is of order q, exactly, and

$$b_q \neq 0, \quad b_{q-1} = b_{q-2} = \cdots = b_1 = b_0 = 0.$$

The reader should recall that, herein, we always assume that $a_q = 1$.

Theorem 20.12 (construction of BDF)**.** *Let $q \in \mathbb{N}$ and set $\beta = \left[\sum_{j=1}^{q} \frac{1}{j}\right]^{-1}$. Suppose that the linear q-step method* (20.1) *is a BDF method. Then $b_q = \beta$ and*

$$\psi(z) = \sum_{j=0}^{q} a_j z^j = \beta \sum_{j=1}^{q} \frac{1}{j} z^{q-j}(z-1)^j$$

or, equivalently,

$$a_q = 1, \qquad a_{q-m} = \beta \sum_{j=m}^{q} \frac{(-1)^m}{j} \binom{j}{m}, \qquad m = 1, \dots, q.$$

Proof. Appealing to Theorem 20.7, we see that, in a neighborhood of $\mu = 1$,

$$\psi(\mu) - b_q \mu^q \ln(\mu) = \sum_{m=q+1}^{\infty} \tilde{C}_m (\mu - 1)^m$$

with $\tilde{C}_{q+1} \neq 0$. In other words, $\mu = 1$ is a $(q+1)$-fold zero. Now we make the substitution $\mu = \nu^{-1}$. In a neighborhood of $\nu = 1$,

$$\nu^q \psi(\nu^{-1}) + b_q \ln(\nu) = \sum_{m=q+1}^{\infty} \hat{C}_m (\nu - 1)^m$$

with $\hat{C}_{q+1} \neq 0$. Thus, in a neighborhood of $\nu = 1$,

$$\nu^q \psi(\nu^{-1}) = b_q \sum_{m=1}^{\infty} \frac{(-1)^m (\nu - 1)^m}{m} + \sum_{m=q+1}^{\infty} \hat{C}_m (\nu - 1)^m \in \mathbb{P}_q.$$

This implies that the tail of the series vanishes:

$$\nu^q \psi(\nu^{-1}) = b_q \sum_{m=1}^{q} \frac{(-1)^m (\nu - 1)^m}{m}.$$

Therefore,

$$\psi(\mu) = b_q \sum_{m=1}^{q} \frac{(-1)^m \mu^q (\mu^{-1} - 1)^m}{m} = b_q \sum_{m=1}^{q} \frac{\mu^{q-m} (\mu - 1)^m}{m} = \sum_{j=0}^{q} a_j \mu^j.$$

It is clear at this point that $a_q = b_q \beta^{-1}$. Thus, to achieve our standard normalization, we require $a_q = 1$, $b_q = \beta$.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Now we establish the equivalence. Using the binomial theorem,

$$\begin{aligned}\sum_{j=0}^{q} a_j\mu^j &= b_q\sum_{j=1}^{q}\frac{1}{j}\mu^{q-j}(\mu-1)^j\\ &= b_q\sum_{j=1}^{q}\frac{1}{j}\mu^{q-j}\sum_{m=0}^{j}\binom{j}{m}\mu^{j-m}(-1)^m\\ &= b_q\sum_{j=1}^{q}\sum_{m=0}^{j}\frac{1}{j}\binom{j}{m}(-1)^m\mu^{q-m}\\ &= b_q\sum_{m=0}^{q}\sum_{j=\max\{1,m\}}^{q}\frac{(-1)^m}{j}\binom{j}{m}\mu^{q-m}.\end{aligned}$$

Thus, we have $a_q = 1$ and

$$a_{q-m} = \beta\sum_{j=m}^{q}\frac{(-1)^m}{j}\binom{j}{m}, \qquad m = 1, \ldots, q. \qquad \square$$

Owing to the previous result, we have the following BDF coefficients.

Example 20.5 For $q = 1$, we find

$$b_1 = 1, \quad a_1 = 1, \quad a_0 = -1.$$

Of course, this corresponds to the single-step backward Euler method.

Example 20.6 For $q = 2$, we find

$$b_2 = \frac{2}{3}, \quad a_2 = 1, \quad a_1 = -\frac{4}{3}, \quad a_0 = \frac{1}{3}.$$

Example 20.7 For $q = 3$, we find

$$b_3 = \frac{6}{11}, \quad a_3 = 1, \quad a_2 = -\frac{18}{11}, \quad a_1 = \frac{9}{11}, \quad a_0 = -\frac{2}{11}.$$

We conclude by commenting that the traditional way to develop BDF methods is via differentiation of the Lagrange interpolating polynomials studied in Chapter 9. The following example illustrates how to obtain the BDFq method using this approach.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Example 20.8 Let $d = 1$ and $u \in C^1([0, T])$ be a classical solution to (18.1). For $0 \leq k \leq K - 2$, and any $t \in [t_k, t_{k+2}]$,

$$
\begin{aligned}
u(t) = u(t_k)\frac{(t - t_{k+1})(t - t_{k+2})}{(t_k - t_{k+1})(t_k - t_{k+2})} + u(t_{k+1})\frac{(t - t_k)(t - t_{k+2})}{(t_{k+1} - t_k)(t_{k+1} - t_{k+2})} \\
+ u(t_{k+2})\frac{(t - t_k)(t - t_{k+1})}{(t_{k+2} - t_k)(t_{k+2} - t_{k+1})} + \mathcal{E}(t),
\end{aligned}
$$

where $\mathcal{E}$ is an error term. Differentiating and evaluating at $t = t_{k+2}$, we get

$$
\begin{aligned}
u'(t_{k+2}) &= u(t_k)\frac{(t_{k+2} - t_{k+1}) + (t_{k+2} - t_{k+2})}{(t_k - t_{k+1})(t_k - t_{k+2})} \\
&\quad + u(t_{k+1})\frac{(t_{k+2} - t_k) + (t_{k+2} - t_{k+2})}{(t_{k+1} - t_k)(t_{k+1} - t_{k+2})} \\
&\quad + u(t_{k+2})\frac{(t_{k+2} - t_k) + (t_{k+2} - t_{k+1})}{(t_{k+2} - t_k)(t_{k+2} - t_{k+1})} + \mathcal{E}'(t_{k+2}) \\
&= \frac{1}{2\tau}u(t_k) - \frac{2}{\tau}u(t_{k+1}) + \frac{3}{2\tau}u(t_{k+2}) + \mathcal{E}'(t_{k+2}) \\
&= f(t_{k+2}, u(t_{k+2})).
\end{aligned}
$$

Equivalently,

$$
u(t_{k+2}) - \frac{4}{3}u(t_{k+1}) + \frac{1}{3}u(t_k) + \frac{2\tau}{3}\mathcal{E}'(t_{k+2}) = \frac{2\tau}{3}f(t_{k+2}, u(t_{k+2})).
$$

This yields the BDF2 method, as claimed.

20.4 Zero Stability

We now examine the issue of zero stability of multi-step methods. It turns out that not all consistent multi-step approximation methods are zero stable. This was not an issue for single-step methods; they are generally always stable. For linear multi-step methods, we need to take great care: if a consistent method is not also stable, it will not be a convergent method.

Definition 20.13 (zero stability)**.** Suppose that $\boldsymbol{f} \in \mathcal{F}^1(S)$ and $\boldsymbol{u} \in C^2([0, T]; \mathbb{R}^d)$ is a classical solution to (18.1). Let, for $i = 1, 2$, $\left\{\boldsymbol{w}_i^k\right\}_{k=0}^K$ be approximations generated by the linear q-step method (20.1) with the starting values $\left\{\boldsymbol{w}_i^k\right\}_{k=0}^{q-1}$, $i = 1, 2$, respectively. The method is called **zero stable** if and only if there is a $C > 0$ independent of $\tau > 0$ and the starting values such that, for any $k = q, \ldots, K$,

$$
\left\|\boldsymbol{w}_1^k - \boldsymbol{w}_2^k\right\|_2 \leq C \max_{m=0,\ldots,q-1} \|\boldsymbol{w}_1^m - \boldsymbol{w}_2^m\|_2 .
$$

Definition 20.14 (root condition)**.** The linear q-step method (20.1) satisfies the **root condition** if and only if:

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

1. All of the roots of the first characteristic polynomial $\psi(z) = \sum_{j=0}^{q} a_j z^j$ are inside the unit disk
$$\{z \in \mathbb{C} \mid |z| \leq 1\} \subset \mathbb{C}.$$
2. If $\psi(\xi) = 0$ and $|\xi| = 1$, then ξ is a simple root, i.e., its multiplicity is exactly one, i.e., $\psi'(\xi) \neq 0$.

Definition 20.15 (homogeneous zero stability)**.** Suppose that $\boldsymbol{f} \equiv \mathbf{0}$ and $\boldsymbol{u}_0 = \mathbf{0}$, so that the unique solution to (18.1) is $\boldsymbol{u}(t) = \mathbf{0}$ for all $t \geq 0$. Let $\left\{\boldsymbol{w}^k\right\}_{k=0}^{K}$ be the approximation generated by the linear q-step method (20.1) with the starting values $\left\{\boldsymbol{w}^k\right\}_{k=0}^{q-1}$. The method is called **homogeneous zero stable** if and only if there is a $C > 0$ independent of $\tau > 0$ and the starting values such that, for any $k = q, \dots, K$,
$$\left\|\boldsymbol{w}^k\right\|_2 \leq C \max_{m=0,\dots,q-1} \|\boldsymbol{w}^m\|_2 .$$

Definition 20.16 (stable solutions)**.** Suppose that $\{a_j\}_{j=0}^{q-1} \subset \mathbb{C}$ are given. An equation of the form
$$\zeta_{k+q} + \sum_{j=0}^{q-1} a_j \zeta_{k+j} = 0, \quad k = 0, 1, 2, \dots \tag{20.8}$$
is called a **homogeneous difference equation**. We say that solutions to (20.8) are **stable** if and only if, given any starting values $\{\zeta_k\}_{k=0}^{q-1} \subset \mathbb{R}$, the sequence $\{\zeta_k\}_{k=0}^{\infty} \subset \mathbb{R}$ is bounded by a constant $C > 0$ that only depends upon the starting values.

We will see that the concepts of stability, homogeneous zero stability, and the root condition are all actually equivalent.

Example 20.9 In this example, we exhibit a method that does not satisfy the root condition and is *not* homogeneously zero stable. Consider the method $q = 2$, $a_2 = 1$, $a_1 = -3$, $a_0 = 2$ and $b_2 = 0$, $b_1 = 0$, $b_0 = -1$. In other words,
$$\boldsymbol{w}^{k+2} - 3\boldsymbol{w}^{k+1} + 2\boldsymbol{w}^k = -\tau \boldsymbol{f}(t_k, \boldsymbol{w}^k)$$
with the starting values $\boldsymbol{w}^0$, $\boldsymbol{w}^1$. The method is consistent. We find
$$C_0 = 0 = C_1, \quad C_2 = \frac{1}{2},$$
which implies that the method is consistent to order $p = 1$.

The first characteristic polynomial is $\psi(z) = z^2 - 3z + 2 = (z - 1)(z - 2)$. Thus, the method fails to satisfy the root condition. Since we are considering homogeneous zero stability, we take $\boldsymbol{f} \equiv \mathbf{0}$ and $u_0 = \mathbf{0}$. The solution of the homogeneous linear constant coefficient difference equation,
$$\zeta_{k+2} - 3\zeta_{k+1} + 2\zeta_k = 0, \quad k = 0, 1, \dots,$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

is precisely

$$\zeta_k = 2\zeta_0 - \zeta_1 + 2^k(\zeta_1 - \zeta_0).$$

This can be verified by a simple induction argument. For starting values, let us take $\zeta_0 = 0$, $\zeta_1 = \tau$. Then $\zeta_k = \tau(2^k - 1)$, $k = 0, 1, 2, \ldots$. Let us examine the approximation at time $T = 1$. In this case, $\tau = 1/K$ and we have, as $K \to \infty$,

$$w^K = \frac{2^K - 1}{K} \to \infty.$$

Thus, the method is not homogeneously zero stable.

We need the following technical lemma in order to construct general solutions of linear homogeneous difference equations. Here, we follow the exposition in the book by Kress [52].

Lemma 20.17 (operator Q)**.** *Suppose that $q \in \mathbb{N}$ and $\nu(t) = \sum_{j=0}^{q} \beta_j t^j \in \mathbb{P}_q$ with coefficients $\beta_j \in \mathbb{C}$, $j = 0, \ldots, q$. Assume that $\beta_q \neq 0$, $\beta_0 \neq 0$. Define the operator $Q \colon \mathbb{P}_q \to \mathbb{P}_q$ via $Q[u](t) = tu'(t)$. The number $\lambda \in \mathbb{C}$ is a root of the polynomial ν of multiplicity m, $m = 1, \ldots, q$ if and only if, for every $p = 0, \ldots, m-1$,*

$$Q^p[\nu](\lambda) = 0, \tag{20.9}$$

where Q^0 is the identity operator but $Q^m[\nu](\lambda) \neq 0$.

Proof. We start with a couple of observations. First, since $\beta_0 \neq 0$, $\lambda = 0$ is not a root of ν. Next, as the case $m = 1$ is trivial, we will assume that $m > 1$. Finally, the repeated application of Q results in

$$Q^p[\nu](t) = \sum_{j=0}^{q} \beta_j j^p t^j \in \mathbb{P}_q,$$

which holds for all $p = 0, 1, 2, \ldots$, provided we interpret $0^0 = 1$.

($\Longrightarrow$) Suppose that $\lambda \neq 0$ is a root of ν of multiplicity $m > 1$. Then $\nu(t) = (t-\lambda)^m \phi_0(t)$, where $\phi_0 \in \mathbb{P}_{q-m}$ and $\phi_0(\lambda) \neq 0$. Then

$$Q[\nu](t) = t\phi_0'(t)(t-\lambda)^m + t\phi_0(t)m(t-\lambda)^{m-1} = (t-\lambda)^{m-1}\phi_1(t),$$

where $\phi_1 \in \mathbb{P}_{q-m+1}$ and $\phi_1(\lambda) \neq 0$. Clearly, $Q[\nu](\lambda) = 0$. Continuing recursively, we observe that

$$Q^p[\nu](t) = (t-\lambda)^{m-p}\phi_p(t),$$

where $\phi_p \in \mathbb{P}_{q-m+p}$ and $\phi_p(\lambda) \neq 0$, for all $p = 1, \ldots, m-1$, with $Q^p[\nu](\lambda) = 0$. However, it is clear that $Q^m[\nu](\lambda) \neq 0$.

($\Longleftarrow$) To obtain a contradiction, suppose that $\lambda \neq 0$ is *not* a root of multiplicity $m > 1$, but property (20.9) holds. Then we proceed in three steps.

1. If λ is not a root of ν at all, then we arrive at a contradiction, namely $Q^0[\nu](\lambda) = \nu(\lambda) \neq 0$.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

2. If λ is a root of ν of multiplicity $n < m$, then we again get a contradiction, since $Q^n[\nu](\lambda) \neq 0$.
3. Lastly, if λ is a root of ν of multiplicity $n > m$, we get a contradiction, since $Q^m[\nu](\lambda) = 0$. It must be that property (20.9) implies that $\lambda \neq 0$ is a root of multiplicity exactly m.

This concludes the proof. □

Theorem 20.18 (solution of difference equations)**.** *Suppose that $q \in \mathbb{N}$, $a_q = 1$, $\{a_j\}_{j=0}^{q-1} \subset \mathbb{R}$, with $a_0 \neq 0$, and $\{\zeta_j\}_{j=0}^{q-1} \subset \mathbb{R}$ are given. The linear homogeneous difference* equation (20.8) *has a unique solution $\{\zeta_k\}_{k=0}^{\infty}$. Assume that $\lambda_j \in \mathbb{C}$, $j = 1, \ldots, r$, where $r \leq q$, are the distinct roots of the characteristic polynomial ψ, with multiplicities m_j, respectively. Hence,*

$$\psi(z) = \prod_{j=1}^{r}(z - \lambda_j)^{m_j}, \qquad \sum_{j=1}^{r} m_j = q.$$

Then the solution to (20.8) *is given by*

$$\zeta_k = \sum_{j=1}^{r} p_j(k)\lambda_j^k,$$

where $p_j \in \mathbb{P}_{m_j-1}$. In particular,

$$p_j(z) = \sum_{m=0}^{m_j-1} \alpha_{j,m} z^m$$

for some coefficients $\alpha_{j,m} \in \mathbb{C}$, $\#\{\alpha_{j,m}\} = q$, that are uniquely determined by the q starting values $\zeta_0, \ldots, \zeta_{q-1}$.

Proof. By linearity, uniqueness follows from the fact that the only possible solution to (20.8) with zero starting values: $\zeta_j = 0$ for $j = 0, \ldots, q-1$, is a sequence of zeros $\zeta_k = 0$, $k = 0, 1, \ldots$.

To show existence, suppose that λ is the root of ψ of multiplicity $m \geq 1$. Consider the sequence

$$\zeta_k = k^n \lambda^k \tag{20.10}$$

for some $n = 0, \ldots, m-1$. Then

$$\begin{aligned}\zeta_{k+q} + \sum_{j=0}^{q-1} a_j \zeta_{k+j} &= \sum_{j=0}^{q} a_j (k+j)^n \lambda^{k+j} \\ &= \sum_{j=0}^{q} a_j \sum_{i=0}^{n} \binom{n}{i} k^i j^{n-i} \lambda^{k+j} \\ &= \lambda^k \sum_{i=0}^{n} \binom{n}{i} k^i \sum_{j=0}^{q} a_j j^{n-i} \lambda^j \\ &= \lambda^k \sum_{i=0}^{n} \binom{n}{i} k^i Q^{n-i}[\psi](\lambda),\end{aligned}$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

where Q is the operator defined in Lemma 20.17. Observe that, since $\psi(z) = (z-\lambda)^m\phi(z)$, where $\phi \in \mathbb{P}_{q-m}$,

$$Q^{n-i}[\psi](\lambda) = 0$$

for all $n = 0, \dots, m-1$ and $i = 0, \dots, n$, appealing to Lemma 20.17. This proves that (20.10) is a solution to (20.8).

Next, we aim to prove that the general solution is just a linear combination of the solutions from above. To establish that the general solution has the form

$$\zeta_k = \sum_{j=1}^{r}\sum_{m=0}^{m_j-1} \alpha_{j,m} k^m \lambda_j^k, \quad k = 0, 1, \dots, \tag{20.11}$$

we need to determine the q free parameters $\alpha_{1,0}, \alpha_{1,1}, \dots, \alpha_{r,m_r-1} \in \mathbb{C}$ such that (20.11) holds for the q starting values, i.e., (20.11) holds for $k = 0, \dots, q-1$. This forms a square $q \times q$ system of linear equations that is uniquely solvable if and only if the corresponding homogeneous system has only the trivial solution. In other words, we want to show that

$$\sum_{j=1}^{r}\sum_{m=0}^{m_j-1} \alpha_{j,m} k^m \lambda_j^k = 0, \quad k = 0, 1, \dots, q-1 \tag{20.12}$$

implies that $\alpha_{j,m} = 0$ for $j = 1, \dots, r$ and $m = 0, \dots, m_j - 1$.

Note that we can represent the homogeneous system above as $\mathsf{B}\boldsymbol{\alpha} = \mathbf{0}$, where $\boldsymbol{\alpha}^\intercal = [\alpha_{1,0}, \alpha_{1,2}, \dots, \alpha_{r,m_r-1}]^\intercal$. We recall that B is nonsingular if and only if $\mathsf{B}^\intercal$ is nonsingular if and only if B^H is nonsingular. In order to prove that $\mathsf{B}^\intercal$ is nonsingular, we consider the homogeneous adjoint equation $\mathsf{B}^\intercal\boldsymbol{\beta} = \mathbf{0}$. In particular, suppose that $\beta_k \in \mathbb{C}$, $k = 0, \dots, q-1$ satisfy the homogeneous adjoint equations

$$\sum_{k=0}^{q-1} \beta_k k^m \lambda_j^k = 0, \quad j = 1, \dots, r, \quad m = 0, \dots, m_j - 1. \tag{20.13}$$

Now define

$$\eta(t) = \sum_{k=0}^{q-1} \beta_k t^k.$$

Then, for any $m \in \mathbb{N}$,

$$Q^m[\eta](t) = \sum_{k=0}^{q-1} \beta_k k^m t^k,$$

and it is clear from (20.13) that

$$Q^m[\eta](\lambda_j) = 0, \quad j = 1, \dots, r, \quad m = 0, \dots, m_j - 1.$$

Appealing to Lemma 20.17, this proves that λ_j is a root η of multiplicity m_j for $j = 1, \dots, r$. All told, $\eta \in \mathbb{P}_{q-1}$ has q roots, counting multiplicities. The only possibility, therefore, is that $\eta \equiv 0$. In other words, $\beta_k = 0$, $k = 0, \dots, q-1$. □

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Because of the form of the general solution, there is a clear connection between the root condition and the stability of solutions to the homogeneous difference equation (20.8).

Theorem 20.19 (root condition and stability)**.** *Suppose that $q \in \mathbb{N}$. Consider a linear q-step method* (20.1) *with coefficients $a_j, b_j \in \mathbb{R}$, $j = 0, \dots, q$, with $a_q = 1$ and $a_0 \neq 0$. The solutions to the corresponding homogeneous difference equation* (20.8) *are bounded, i.e., stable, if and only if the root condition is satisfied.*

Proof. Given the starting values ζ_j, $j = 0, \dots, q-1$, the general solution to (20.8) is

$$\zeta_k = \sum_{j=1}^{r} \sum_{m=0}^{m_j - 1} \alpha_{j,m} k^m \lambda_j^k, \qquad k = 0, 1, \dots,$$

where the q coefficients $\alpha_{1,1}, \alpha_{1,2}, \dots, \alpha_{r,m_r} \in \mathbb{C}$ are determined uniquely by the q starting values.

($\Longrightarrow$) From the form of the general solution, it is clear that, if the root condition is not satisfied, the approximations can grow unboundedly, as in Example 20.9.

($\Longleftarrow$) Conversely, if the approximations remain bounded, the only possibility is that the root condition is satisfied. The details are left to the reader as an exercise; see Problem 20.8. □

Theorem 20.20 (nonhomogeneous difference equations)**.** *Let $q \in \mathbb{N}$ and, for $m = 0, \dots, q-1$, the sequence $\left\{ g_k^{(m)} \right\}_{k=0}^{\infty}$ be the unique solution to the homogeneous linear difference equation* (20.8) *with $a_j \in \mathbb{R}$, $j = 0, \dots, q$, $a_q = 1$, and $a_0 \neq 0$ and starting values*

$$g_k^{(m)} = \delta_{k,m}, \qquad k, m = 0, 1, 2, \dots, q-1. \tag{20.14}$$

Let $\{c_k\}_{k=q}^{\infty} \subset \mathbb{C}$ be a given sequence. Then there is a unique solution to the linear difference equation

$$\zeta_{k+q} + \sum_{j=0}^{q-1} a_j \zeta_{k+j} = c_{k+q}, \qquad k = 0, 1, \dots, \tag{20.15}$$

with starting values $\{\zeta_k\}_{k=0}^{q-1}$. This solution is given by

$$\zeta_{k+q} = \sum_{j=0}^{q-1} \zeta_j g_{k+q}^{(j)} + \sum_{j=0}^{k} c_{j+q} g_{k+q-j-1}^{(q-1)}, \qquad k = 0, 1, 2, \dots. \tag{20.16}$$

Proof. See Kress [52, Chapter 10] or Gautschi [32, Chapter 6]. □

Theorem 20.21 (zero stability)**.** *A linear q-step method* (20.1) *is homogeneous zero stable if and only if solutions to the corresponding homogeneous equations* (20.8) *are stable.*

Proof. It suffices to prove the result for the scalar case, i.e., $d = 1$.

($\Longrightarrow$) Suppose that the linear q-step method (20.1) with the first characteristic polynomial $\psi(z) = \sum_{j=0}^{q} a_j z^j$ is homogeneous zero stable, i.e., if $\{w^k\}_{k=q}^{K}$ solves (20.1) with starting values $\{w^m\}_{m=0}^{q-1}$, then

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

$$|w^k| \leq C \max_{m=0,\dots,q-1} |w^m|.$$

Let now the sequence $\{\zeta_k\}_{k=q}^{\infty}$ solve (20.8) with starting values $\zeta_k = w^k$ for $k = 0, \dots, q-1$. Notice that we must necessarily have $\zeta_k = w^k$ for $k = q, \dots, K$, where $\tau K = T$. In other words, the product τK is always the same fixed constant. The homogeneous zero stability of (20.1) then shows that there exists a constant $C > 0$ independent of $\tau > 0$ such that, for any $k = q, \dots, K$,

$$|\zeta_k| \leq C \max_{m=0,\dots,q-1} |\zeta_m| . \tag{20.17}$$

Since C is independent of τ, it must also be independent of K. In other words, $K \in \mathbb{N}$ may be arbitrarily large. It follows that (20.17) holds for any $k \in \mathbb{N}$. It must be that $\{\zeta_k\}_{k=q}^{\infty}$ is bounded for any given set of starting values $\{\zeta_k\}_{k=0}^{q-1}$.

($\Longleftarrow$) Let, for $m = 0, \dots, q-1$, $\left\{g_k^{(m)}\right\}_{k=0}^{\infty}$ be the unique solution to the homogeneous linear difference equation (20.8) with the "impulse" starting values (20.14). Then, for all $m = 0, \dots, q-1$,

$$\left|g_k^{(m)}\right| \leq C_m$$

for all $k \in \mathbb{N}$, where $C_m > 0$ is independent of k. We can then define

$$\widehat{C} = \max_{m=0,\dots,q-1} C_m$$

to obtain a constant that is independent of m. Suppose that, with the starting values $\left\{w^k\right\}_{k=0}^{q-1}$, the sequence $\left\{w^k\right\}_{k=q}^{\infty}$ satisfies (20.1) with $f \equiv 0$. Then, using (20.16), we have

$$w^{k+q} = \sum_{j=0}^{q-1} w^j g_{k+q}^{(j)}, \qquad k = 0, 1, 2, \dots .$$

Taking absolute values and using the triangle inequality, we get

$$\left|w^{k+q}\right| \leq \sum_{j=0}^{q-1} \left|w^j\right| \left|g_{k+q}^{(j)}\right| \leq \widehat{C} \sum_{j=0}^{q-1} \left|w^j\right| \leq q\widehat{C} \max_{m=0,\dots,q-1} |w^m| .$$

The proof is complete. □

Remark 20.22 ($a_0 = 0$)**.** We have only discussed the case for which $a_0 \neq 0$. What if $a_0 = 0$?

20.5 Convergence of Linear Multi-step Methods

Having discussed the notions of consistency and stability for multi-step methods, we are now ready to present the theory regarding their convergence. We begin with yet another discrete incarnation of Grönwall's[3] Lemma.

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

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Lemma 20.23 (discrete Grönwall)**.** *Let $\{a_n\}_{n=0}^{\infty} \subset \mathbb{R}_+ \cup \{0\}$ be a sequence with the property that, for $n = 1, 2, \ldots,$*

$$a_n \leq b \sum_{m=0}^{n-1} a_m + c$$

for some constants $b > 0$ and $c \geq 0$. Then, for all $n = 1, 2, \ldots,$

$$a_n \leq (ba_0 + c)\, e^{(n-1)b}.$$

Proof. The result follows by an induction argument, which is left to the reader as an exercise; see Problem 20.11. □

Theorem 20.24 (convergence)**.** *Let $p, K, q \in \mathbb{N}$ with $q < K$. Suppose that $\boldsymbol{f} \in \mathcal{F}^p(S)$, so that $\boldsymbol{u} \in C^{p+1}([0,T]; \mathbb{R}^d)$ satisfies the IVP (18.1). Let $\left\{\boldsymbol{w}^k\right\}_{k=0}^{K} \subset \mathbb{R}^d$ be an approximation generated by the linear q-step method (20.1) with the starting values $\left\{\boldsymbol{w}^k\right\}_{k=0}^{q-1} \subset \mathbb{R}^d$. Assume that the starting values are such that, for some constant $\tau_0 \in (0,T]$ and some $C_0 > 0$, we have*

$$\max_{k=0,\ldots,q-1} \left\|\boldsymbol{e}^k\right\|_2 \leq C_0 \tau^p, \quad \forall \tau \in (0, \tau_0].$$

Assume, in addition, that the multi-step method is consistent to order p and satisfies the root condition. In this setting, there are constants $\tau_1 \in (0,T]$ and $C_1 > 0$ such that

$$\max_{k=0,\ldots,K} \left\|\boldsymbol{e}^k\right\|_2 \leq C_1 \tau^p, \quad \forall \tau \in (0, \tau_1].$$

Proof. Let us assume that the method is implicit. The explicit case is simpler. Then we have the following error equation: for $k = 0, \ldots, K - q$,

$$\begin{aligned} \boldsymbol{e}^{k+q} + \sum_{j=0}^{q-1} a_j \boldsymbol{e}^{k+j} &= \tau \sum_{j=0}^{q} b_j \left(\boldsymbol{f}(t_{k+j}, \boldsymbol{u}(t_{k+j})) - \boldsymbol{f}(t_{k+j}, \boldsymbol{w}^{k+j})\right) + \tau \boldsymbol{\mathcal{E}}^{k+q}[\boldsymbol{u}] \\ &= \tau \boldsymbol{c}^{k+q}. \end{aligned}$$

Since $\boldsymbol{f}$ satisfies a global $\boldsymbol{u}$-Lipschitz condition, we can estimate $\boldsymbol{c}^{k+q}$, as follows: by the triangle inequality and the Lipschitz continuity,

$$\left\|\boldsymbol{c}^{k+q}\right\|_2 \leq LB \sum_{j=0}^{q} \left\|\boldsymbol{e}^{k+j}\right\|_2 + \left\|\boldsymbol{\mathcal{E}}^{k+q}[\boldsymbol{u}]\right\|_2, \tag{20.18}$$

where $B = \max_{j=0,\ldots,q} |b_j|$ and $L > 0$ is the standard Lipschitz constant.

By Theorem 20.20, the solution of the error equation can be represented as

$$\boldsymbol{e}^{k+q} = \sum_{j=0}^{q-1} g_{k+q}^{(j)} \boldsymbol{e}^j + \tau \sum_{j=0}^{k} g_{k+q-j-1}^{(q-1)} \boldsymbol{c}^{j+q}$$

for all $k = 0, \ldots, K - q$. Because the q-step method satisfies the root condition and is, therefore, homogeneous zero stable, the solutions $g_k^{(j)}$ are bounded for every

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

$j = 0, \dots, q-1$ and every $k = 0, 1, 2, \dots$. In other words, there is a constant $C > 0$ such that, for all $j = 0, \dots, q-1$ and every $k = 0, 1, 2, \dots$,

$$\left| g_k^{(j)} \right| \le C.$$

So, we can estimate the error as

$$\left\| \boldsymbol{e}^{k+q} \right\|_2 \le \sum_{j=0}^{q-1} C\tau^p + \tau \sum_{j=0}^{k} C \left\| \boldsymbol{c}^{j+q} \right\|_2 \le C \left(\tau^p + \tau \sum_{j=0}^{k} \left\| \boldsymbol{c}^{j+q} \right\|_2 \right),$$

provided that $0 < \tau \le \tau_0$. Here and in what follows, the constant C may change value from line to line. The important point is that it is a constant, and it is independent of all the involved quantities. Now we can use our estimate (20.18) above to obtain

$$\begin{aligned} \left\| \boldsymbol{e}^{k+q} \right\|_2 &\le C \left\{ \tau^p + \tau BL \sum_{j=0}^{k} \sum_{m=0}^{q} \left\| \boldsymbol{e}^{j+m} \right\|_2 + \tau \sum_{j=0}^{k} \left\| \boldsymbol{\mathcal{E}}^{j+q}[\boldsymbol{u}] \right\|_2 \right\} \\ &\le C \left\{ \tau^p + BL\tau \sum_{j=0}^{k} \sum_{m=0}^{q} \left\| \boldsymbol{e}^{j+m} \right\|_2 + \tau(k+1)C_1\tau^p \right\} \\ &\le C \left\{ \tau^p + BL\tau \sum_{j=0}^{k} \sum_{m=0}^{q} \left\| \boldsymbol{e}^{j+m} \right\|_2 + TC_1\tau^p \right\} \end{aligned} \tag{20.19}$$

for all $k = 0, \dots, K-q$ and for all $0 < \tau \le \tau_1$. Now observe that

$$\begin{aligned} \sum_{j=0}^{k} \sum_{m=0}^{q} \left\| \boldsymbol{e}^{j+m} \right\|_2 &= \sum_{m=0}^{q} \sum_{j=0}^{k} \left\| \boldsymbol{e}^{j+m} \right\|_2 \\ &\le (q+1) \sum_{j=0}^{k+q} \left\| \boldsymbol{e}^{j} \right\|_2 \\ &\le (q+1) \sum_{j=0}^{q-1} \left\| \boldsymbol{e}^{j} \right\|_2 + (q+1) \sum_{j=q}^{k+q} \left\| \boldsymbol{e}^{j} \right\|_2 \\ &\le q(q+1)C\tau^p + (q+1) \sum_{j=q}^{k+q} \left\| \boldsymbol{e}^{j} \right\|_2 . \end{aligned} \tag{20.20}$$

Combining estimates (20.19) and (20.20), we get

$$\begin{aligned} \left\| \boldsymbol{e}^{k+q} \right\|_2 &\le C \left\{ \tau^p + \bar{C}\tau^{p+1} + BL(q+1)\tau \sum_{j=q}^{k+q} \left\| \boldsymbol{e}^{j} \right\|_2 + T\bar{\bar{C}}\tau^p \right\} \\ &\le C\tau^p + \tau\tilde{C} \sum_{j=q}^{k+q} \left\| \boldsymbol{e}^{j} \right\|_2 . \end{aligned}$$

Provided that τ is sufficiently small, in particular

$$0 < \tau\tilde{C} < 1,$$

we have, for all $k = 0, 1, \dots, K - q$,

$$\left\| \boldsymbol{e}^{k+q} \right\|_2 \leq \frac{C}{1 - \tau\tilde{C}}\tau^p + \tau\frac{\tilde{C}}{1 - \tau\tilde{C}}\sum_{j=q}^{k+q-1} \left\| \boldsymbol{e}^j \right\|_2 .$$

Notice that we have made the sum on the right-hand side explicit with respect to the left-hand side. Reindexing the summation, we have, for every $m = q, q+2, \dots, K$,

$$\left\| \boldsymbol{e}^m \right\|_2 \leq \frac{C}{1 - \tau\tilde{C}}\tau^p + \tau\frac{\tilde{C}}{1 - \tau\tilde{C}}\sum_{j=q}^{m-1} \left\| \boldsymbol{e}^j \right\|_2 .$$

If we further restrict the time step so that

$$0 < \tau\tilde{C} \leq \frac{1}{2},$$

then it follows that

$$\frac{1}{1 - \tau\tilde{C}} \leq 2,$$

and, for every $m = q, q+2, \dots, K$,

$$\left\| \boldsymbol{e}^m \right\|_2 \leq 2\tilde{C}\tau^p + 2\tau\tilde{C}\sum_{j=q}^{m-1} \left\| \boldsymbol{e}^j \right\|_2 .$$

Applying Lemma 20.23, we have

$$\left\| \boldsymbol{e}^m \right\|_2 \leq \left(2\tau\tilde{C} \left\| \boldsymbol{e}^q \right\|_2 + 2C\tau^p\right) \exp\left[2\tau(m - q - 1)\tilde{C}\right] \leq C\tau^p e^{2T\tilde{C}}.$$

The theorem is proven with $C_1 = Ce^{2T\tilde{C}}$. □

The reader will notice that, in the proof of Theorem 20.24, we only used the homogeneous zero stability of the method. In fact, we now have the tools to prove that, if $\boldsymbol{f}$ satisfies the usual global Lipschitz condition, the notions of zero stability and homogeneous zero stability are equivalent.

Corollary 20.25 (equivalence)**.** *Let $K, q \in \mathbb{N}$ with $q < K$. Suppose that $\boldsymbol{f} \in \mathcal{F}^1(S)$, so that $\boldsymbol{u} \in C^2([0, T]; \mathbb{R}^d)$ satisfies the IVP* (18.1)*. Let $\left\{\boldsymbol{w}^k\right\}_{k=0}^{K} \subset \mathbb{R}^d$ be an approximation generated by the linear q-step method* (20.1) *with the starting values $\left\{\boldsymbol{w}^k\right\}_{k=0}^{q-1} \subset \mathbb{R}^d$. The multi-step method is zero stable if and only if it is homogeneously zero stable.*

Proof. See Problem 20.12. □

20.6 Dahlquist Theorems

As a last topic in our discussion of the linear multi-step method, we present a series of results due to Dahlquist. The first one is known as the *Dahlquist Equivalence Theorem*. This essentially gives a converse to Theorem 20.24. The proof can be found in [32].

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Theorem 20.26 (Dahlquist Equivalence Theorem [4])**.** *Let $p, K, q \in \mathbb{N}$ with $q < K$. Suppose that $\boldsymbol{f} \in \mathcal{F}^p(S)$, so that $\boldsymbol{u} \in C^{p+1}([0,T];\mathbb{R}^d)$ satisfies the IVP* (18.1)*. Let $\left\{\boldsymbol{w}^k\right\}_{k=0}^K \subset \mathbb{R}^d$ be an approximation generated by the linear q-step method* (20.1) *with the starting values $\left\{\boldsymbol{w}^k\right\}_{k=0}^{q-1} \subset \mathbb{R}^d$. Suppose that the method satisfies the root condition. Then the multi-step method is consistent to order p if and only if it is globally convergent with order p.*

The next theorem is known as the *Dahlquist First Barrier Theorem*; see, for example, [12] or [32]. The result gives us a firm upper limit on the order of a *zero stable* multi-step method.

Theorem 20.27 (Dahlquist First Barrier Theorem)**.** *The order of accuracy (consistency) of a zero stable linear q-step method* (20.1) *cannot exceed $q+1$ if q is odd or $q+2$ if q is even.*

Problems

20.1 Complete the proof of Corollary 20.6.

20.2 Show that the two-step (implicit) Adams–Moulton method,

$$\boldsymbol{w}^{k+2} - \boldsymbol{w}^{k+1} = \tau\left[\frac{5}{12}\boldsymbol{f}(t_{k+2},\boldsymbol{w}^{k+2}) + \frac{8}{12}\boldsymbol{f}(t_{k+1},\boldsymbol{w}^{k+1}) \quad - \frac{1}{12}\boldsymbol{f}(t_k,\boldsymbol{w}^k)\right],$$

is at least order one using the conditions $\psi(1) = 0$ and $\psi'(1) - \chi(1) = 0$. Prove that, in fact, the method is exactly third order.

20.3 Show that the three-step (implicit) Adams–Moulton method,

$$\boldsymbol{w}^{k+3} - \boldsymbol{w}^{k+2} = \tau\left[\frac{9}{24}\boldsymbol{f}(t_{k+3},\boldsymbol{w}^{k+3}) + \frac{19}{24}\boldsymbol{f}(t_{k+2},\boldsymbol{w}^{k+2}) - \frac{5}{24}\boldsymbol{f}(t_{k+1},\boldsymbol{w}^{k+1}) + \frac{1}{24}\boldsymbol{f}(t_k,\boldsymbol{w}^k)\right],$$

is at least order one using the conditions $\psi(1) = 0$ and $\psi'(1) - \chi(1) = 0$. Prove that, in fact, the method is exactly fourth order using both the method of C's and the log-method.

20.4 Find all of the values of α and β, so that the three-step method,

$$\boldsymbol{w}^{k+3} + \alpha(\boldsymbol{w}^{k+2} - \boldsymbol{w}^{k+1}) - \boldsymbol{w}^k = \tau\beta\left[\boldsymbol{f}(t_{k+2},\boldsymbol{w}^{k+2}) + \boldsymbol{f}(t_{k+1},\boldsymbol{w}^{k+1})\right],$$

is of order four.

20.5 Show that the BDF3 method,

$$\boldsymbol{w}^{k+3} - \frac{18}{11}\boldsymbol{w}^{k+2} + \frac{9}{11}\boldsymbol{w}^{k+1} - \frac{2}{11}\boldsymbol{w}^k = \frac{6}{11}\tau\boldsymbol{f}(t_{k+3},\boldsymbol{w}^{k+3}),$$

is consistent to order $p = 3$ using both the method of C's and the log-method.

20.6 The Adams–Bashforth two-step method is given by

$$\boldsymbol{w}^{k+2} = \boldsymbol{w}^{k+1} + \tau\left[\frac{3}{2}\boldsymbol{f}(t_{k+1},\boldsymbol{w}^{k+1}) - \frac{1}{2}\boldsymbol{f}(t_k,\boldsymbol{w}^k)\right].$$

[4] Named in honor of the Swedish mathematician Germund Dahlquist (1925–2005).

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

Derive this method by an integration procedure and give an exact expression for the local truncation error.

20.7 Derive the general form of a q-step BDF method using the method of C's.

20.8 Complete the proof of Theorem 20.19.

20.9 Show that the BDF3 method,

$$\boldsymbol{w}^{k+3} - \frac{18}{11}\boldsymbol{w}^{k+2} + \frac{9}{11}\boldsymbol{w}^{k+1} - \frac{2}{11}\boldsymbol{w}^{k} = \frac{6}{11}\tau \boldsymbol{f}(t_{k+3}, \boldsymbol{w}^{k+3}),$$

satisfies the root condition.

Hint: One of the roots is $w = 1$.

20.10 Consider the method

$$\boldsymbol{w}^{k+2} - \boldsymbol{w}^{k} = \frac{\tau}{3}\left[\boldsymbol{f}(t_{k+2}, \boldsymbol{w}^{k+2}) + 4\boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1}) + \boldsymbol{f}(t_k, \boldsymbol{w}^{k})\right].$$

Show that it is of fourth order and it obeys the root condition.

20.11 Prove Lemma 20.23.

20.12 Prove Corollary 20.25.

20.13 Show that, for all the values of α and β that make the three-step method

$$\boldsymbol{w}^{k+3} + \alpha(\boldsymbol{w}^{k+2} - \boldsymbol{w}^{k+1}) - \boldsymbol{w}^{k} = \tau\beta\left[\boldsymbol{f}(t_{k+2}, \boldsymbol{w}^{k+2}) + \boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1})\right]$$

of order four, the resulting method does not satisfy the root condition and is, therefore, not convergent.

20.14 Show that a linear multi-step method is of order $p \geq 1$ if and only if it yields the exact solution to an ordinary differential equation problem whose solution is a polynomial of degree no greater than p.

20.15 Recall Simpson's quadrature rule:

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] + E[f](a,b),$$

where $E[f](a,b)$ is an error term that satisfies

$$|E[f](a,b)| \leq C(b-a)^4$$

and $C > 0$ is a constant that depends on f. Starting from the identity

$$\boldsymbol{u}(t_{k+1}) = \boldsymbol{u}(t_{k-1}) + \int_{t_{k-1}}^{t_{k+1}} \boldsymbol{f}(s, \boldsymbol{u}(s))\mathrm{d}s,$$

use Simpson's rule to derive a two-step method. Determine its order and whether it is convergent.

20.16 Show that the method

$$\boldsymbol{w}^{k+2} - 3\boldsymbol{w}^{k+1} + 2\boldsymbol{w}^{k} = \tau\left[\frac{13}{12}\boldsymbol{f}(t_{k+2}, \boldsymbol{w}^{k+2}) - \frac{5}{3}\boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1}) - \frac{5}{12}\boldsymbol{f}(t_k, \boldsymbol{w}^{k})\right]$$

is of order two. However, this method does not converge. Why?

20.17 Show that the explicit multi-step method,

$$\begin{aligned}&\boldsymbol{w}^{k+3} + \alpha_2\boldsymbol{w}^{k+2} + \alpha_1\boldsymbol{w}^{k+1} + \alpha_0\boldsymbol{w}^{k}\\ &\qquad = \tau\left[\beta_2\boldsymbol{f}(t_{k+2}, \boldsymbol{w}^{k+2}) + \beta_1\boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1}) + \beta_0\boldsymbol{f}(t_k, \boldsymbol{w}^{k})\right],\end{aligned}$$

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press

is fourth order only if $\alpha_0 + \alpha_2 = 8$ and $\alpha_1 = -9$. Prove that this method cannot be both fourth order and convergent.

20.18 Consider the method

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

What is the order of the method? Is it a convergent method?

20.19 Study the order and convergence of the method

$$\boldsymbol{w}^{k+1} - \boldsymbol{w}^k = \frac{\tau}{12}\left[5\boldsymbol{f}(t_{k+1}, \boldsymbol{w}^{k+1}) + 8\boldsymbol{f}(t_k, \boldsymbol{w}^k) - \boldsymbol{f}(t_{k-1}, \boldsymbol{w}^{k-1})\right].$$

Listings

```
function res = MethodCs( a, b, m )
% The method of Cs to determine the order of consistency of a
% linear multistep method.
%
% Input
% a : The coefficients of the first characteristic polynomial
% b : The coefficients of the second characteristic polynomial
% m : The number of C that one wants to compute
%
% Output
% res: The number C_m. A method is consistent to order exactly
%        p if C_0 = ... = C_p = 0, but C_{p+1} != 0
  if m == 0
    res = sum( a );
  else
    res = 0.;
    q = length(a);
    factmminusone = factorial(m-1);
    factm = m*factmminusone;
    for j=1:q
      res = res + a(j)*(j-1)^m/factm - b(j)*(j-1)^(m-1) ...
        /factmminusone;
    end
  end
end
```

Listing 20.1 Algorithmic description of the method of C's.

https://doi.org/10.1017/9781108942607.026 Published online by Cambridge University Press