Consistency of numerical methods for ODEs

In the previous course note we provided an overview of numerical methods to solve an initial value problem for the system of ODEs

$$\begin{cases} \dfrac{d\boldsymbol{y}}{dt} = \boldsymbol{f}(\boldsymbol{y}, t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases} \tag{1}$$

In particular, we discussed linear multistep methods (LMM), Runge-Kutta methods (RK), and backward differentiation formulas (BFD) methods. All these methods can be written in the general form (see [1, p. 24])

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = \Delta t \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_{k+q}, \dots, \boldsymbol{u}_k, t_k, \Delta t), \tag{2}$$

where $\Phi_{\boldsymbol{f}}$ is an *iteration function* that depends on $\boldsymbol{f}$ (right hand side of Eq. (1)), t_k, Δt and the approximate solution $\boldsymbol{u}_k$ at different time steps. Equation (2) has only *one unknown*, i.e., $\boldsymbol{u}_{k+q}$ (numerical solution at time t_{k+q}, which can be computed from (2) either implicitly or explicitly given $\{\boldsymbol{u}_{k+q-1}, \dots, \boldsymbol{u}_k\}$ (numerical solution at previous time steps).

In (2) we set $\alpha_q = 1$ to avoid non-uniqueness of the set of coefficients $\{\alpha_j\}$, e.g., when we multiply (2) by a nonzero constant. The iteration function $\Phi_{\boldsymbol{f}}$ satisfies the following conditions

$$\boldsymbol{\Phi}_{\boldsymbol{0}}(\boldsymbol{u}_{k+q}, \dots, \boldsymbol{u}_k, t_k, \Delta t) = 0, \tag{3}$$

$$\left\|\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{z}_{k+q}, \dots, \boldsymbol{z}_k, t_k, \Delta t) - \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_{k+q}, \dots, \boldsymbol{u}_k, t_k, \Delta t)\right\| \leq M \sum_{j=0}^{q} \left\|\boldsymbol{u}_{k+j} - \boldsymbol{z}_{k+j}\right\|. \tag{4}$$

The second condition follows from the assumption that $\boldsymbol{f}$ is Lipshitz continuous. Let us show how to write a few well-known schemes in the form (2).

- **Crank-Nicolson (AM1 or implicit RK2):**

  The Crank-Nicolson scheme corresponds to $q = 1$ (one step), with $\alpha_1 = 1$, $\alpha_0 = -1$ and iteration function given by

$$\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_k, \boldsymbol{u}_{k+1}, t_k, \Delta t) = \frac{1}{2}\left[\boldsymbol{f}(\boldsymbol{u}_k, t_k) + \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_k + \Delta t)\right]. \tag{5}$$

- **Heun (explicit RK2):**

  Here we have again $q = 1$ (one step), $\alpha_1 = 1$, $\alpha_0 = -1$ and iteration function given by

$$\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_k, t_k, \Delta t) = \frac{1}{2}\left[\boldsymbol{f}(\boldsymbol{u}_k, t_k) + \boldsymbol{f}(\boldsymbol{u}_k + \Delta t \boldsymbol{f}(\boldsymbol{u}_k, t_k), t_k + \Delta t)\right]. \tag{6}$$

- **Adams-Bashforth 2:**

  Here we have $q = 2$ (two-steps), $\alpha_2 = 1$, $\alpha_1 = -1$, $\alpha_0 = 0$

$$\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_{k+1}, \boldsymbol{u}_k, t_k, \Delta t) = \frac{1}{2}\left[3\boldsymbol{f}(\boldsymbol{u}_{n+1}, t_k + \Delta t) - \boldsymbol{f}(\boldsymbol{u}_k, t_k)\right]. \tag{7}$$

Clearly, if $\boldsymbol{f}$ is Lipschitz continuous then the iteration functions (5)-(7) satisfy condition (4) for some $M \geq 0$. Moreover, as easily seen, they all satisfy condition (3).

Local truncation error and consistency

The local truncation error of a numerical scheme is the error arising from the scheme when we perform one step forward from an exact initial condition, i.e., an initial condition defined by the analytical solution of the ODE system (1). In other words, it is the error that we make when we substitute the exact solution into the scheme.

Let $\boldsymbol{y}(t)$ be the analytical solution of (1), and denote by $\boldsymbol{y}_k = \boldsymbol{y}(t_k)$. The *local truncation error* of the scheme (2), denoted as $\boldsymbol{\tau}_{k+q}$, is defined as

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{y}_{k+j} = \Delta t \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{y}_{k+q}, \ldots, \boldsymbol{y}_k, t_k, \Delta t) + \Delta t \boldsymbol{\tau}_{k+q}, \tag{8}$$

i.e.,

$$\boldsymbol{\tau}_{k+q} = \frac{1}{\Delta t} \sum_{j=0}^{q} \alpha_j \boldsymbol{y}_{k+j} - \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{y}_{k+q}, \ldots, \boldsymbol{y}_k, t_k, \Delta t), \qquad \text{(local truncation error)} \tag{9}$$

The *global truncation error* of a numerical scheme that undergoes multiple iterations, say N within a certain time interval $[0, T]$, is defined as

$$\|\boldsymbol{\tau}\| = \max_{k=1,\ldots,N} \|\boldsymbol{\tau}_{k+q}\|. \tag{10}$$

Definition 1 (**Consistency**)**.** Let $\boldsymbol{\tau}_{k+q}$ be the local truncation error of the scheme (2). If

$$\lim_{\Delta t \to 0} \|\boldsymbol{\tau}_{k+q}\| = 0 \tag{11}$$

then we say that the numerical scheme (2) is *consistent*. If $\|\boldsymbol{\tau}_{k+q}\|$ goes to zero as Δt^p then we say that the numerical scheme is *consistent with order p*.

Stated in simple terms, a consistent numerical scheme is a scheme that converges to the ODE (1) (to some order in Δt) when we send Δt to zero. As we will see, this is not alone sufficient to guarantee that the numerical solution we obtain by iterating the scheme (i.e., $\boldsymbol{u}_k$), converges to the solution to (1), i.e.,

$$\lim_{\Delta t \to 0} \|\boldsymbol{u}_k - \boldsymbol{y}_k\| = 0 \qquad \text{for all} \qquad k = 0, \ldots, N \qquad \text{(convergence)}. \tag{12}$$

In order for $\boldsymbol{u}_k$ to converge to $\boldsymbol{y}_k = \boldsymbol{y}(t_k)$ as $\Delta t \to 0$ the scheme must to be *consistent* and *zero-stable.*

Remark: In equations (11)-(12) $\|\cdot\|$ denotes any norm in $\mathbb{R}^n$. Since all norms are equivalent in $\mathbb{R}^n$ (see Appendix A, in course note 2) we have that consistency in one norm implies consistency in any other norm. Also, the order of consistency does not depend on the norm that is used. Let us consider a few examples in which we show how to calculate the local truncation error and the order of consistency.

- **Consistency of Euler-Forward:** The Euler forward scheme

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \Delta t \boldsymbol{f}(\boldsymbol{u}_k, t_k) \tag{13}$$

is consistent with order one. To this end, suppose we have available the analytical solution $\boldsymbol{y}(t)$ to (1), and denote by

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}(t_{k+1}). \tag{14}$$

A substitution of this expression into (13) yields

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \Delta t \boldsymbol{f}(\boldsymbol{y}_k, t_k) + \Delta t \boldsymbol{\tau}_{k+1}, \tag{15}$$

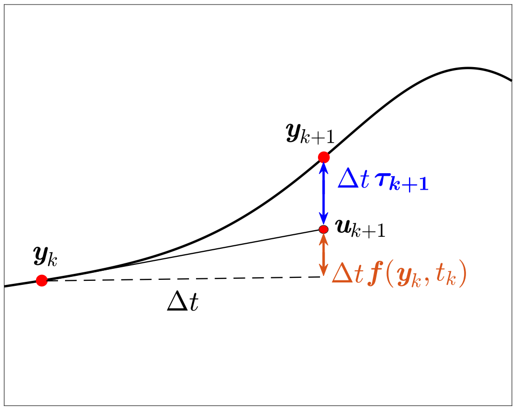
$\boldsymbol{y}_{k+1}$
$\Delta t\, \boldsymbol{\tau}_{k+1}$
$\boldsymbol{u}_{k+1}$
$\Delta t\, \boldsymbol{f}(\boldsymbol{y}_k, t_k)$
$\boldsymbol{y}_k$
Δt

Figure 1: Geometric meaning of the local truncation error (LTE) for the Euler forward scheme. The LTE represents is the difference (divided by Δt) between the analytical solution at time t_{k+1}, i.e., $\boldsymbol{y}_{t_k+1}$, and the numerical solution $\boldsymbol{u}_{k+1}$ obtained by applying the Euler scheme to the "exact" initial condition $\boldsymbol{y}_k$.

i.e.,

$$\boldsymbol{\tau}_{k+1} = \frac{\boldsymbol{y}_{k+1} - \boldsymbol{y}_k}{\Delta t} - \boldsymbol{f}(\boldsymbol{y}_k, t_k). \tag{16}$$

By expanding $\boldsymbol{y}_{k+1} = \boldsymbol{y}(t_{k+1})$ in a Taylor series (in time) we obtain[1]

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \Delta t \frac{d\boldsymbol{y}(t_k)}{dt} + \underbrace{\frac{\Delta t^2}{2} \frac{d^2 \boldsymbol{y}(\boldsymbol{\xi}_k)}{dt^2}}_{\text{remainder}}, \qquad \boldsymbol{\xi}_k \in [t_k, t_{k+1}]^n. \tag{17}$$

In this equation, $\boldsymbol{y}(\boldsymbol{\xi}_k)$ denotes a vector with components

$$\boldsymbol{y}(\boldsymbol{\xi}_k) = \begin{bmatrix} y_1(\xi_{1k}) \\ y_2(\xi_{2k}) \\ \vdots \\ y_n(\xi_{nk}) \end{bmatrix}. \tag{18}$$

Substituting the series (17) into the expression (16) and computing the norm yields

$$\|\boldsymbol{\tau}_{k+1}\| = \frac{\Delta t}{2} \left\| \frac{d^2 \boldsymbol{y}(\boldsymbol{\xi}_k)}{dt^2} \right\|. \tag{19}$$

Hence, the Euler forward method is *consistent with order one.* In Figure 1 we show the geometric meaning of the local truncation error for the Euler forward method.

- **Consistency of Crank-Nicolson:** The Crank-Nicolson scheme

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \frac{\Delta t}{2} \left[\boldsymbol{f}(\boldsymbol{u}_k, t_k) + \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1})\right] \tag{20}$$

is *consistent with order two.* To show this, let us substitute of the analytical solution of (1), denoted as $\boldsymbol{y}_k = \boldsymbol{y}(t_k)$, into (20) to obtain the local truncation error

$$\boldsymbol{\tau}_{k+1} = \frac{\boldsymbol{y}_{k+1} - \boldsymbol{y}_k}{\Delta t} + \frac{1}{2} \left[\boldsymbol{f}(\boldsymbol{y}_k, t_k) + \boldsymbol{f}(\boldsymbol{y}_{k+1}, t_{k+1})\right]. \tag{21}$$

[1] Equation (17) represents simultaneously n different Taylor series, one for each $y_j(t_k + \Delta t)$, $j = 1, \ldots, n$. This yields a remainder for each series in which the second derivative $d^2 y_j / dt^2$ is evaluated at some point ξ_{jk} within the time interval $[t_k, t_{k+1}]$. If we use a vector notation, this yields a vector $\boldsymbol{\xi}_k$ representing a point in the hyper-cube $[t_k, t_{k+1}]^n$.

Next, we expand $\boldsymbol{f}(\boldsymbol{y}_{k+1}, t_{k+1})$ in a Taylor series

$$\boldsymbol{f}(\boldsymbol{y}_{k+1}, t_{k+1}) = \boldsymbol{f}(\boldsymbol{y}_k, t_k) + \Delta t \left.\frac{d\boldsymbol{f}(\boldsymbol{y}(t), t)}{dt}\right|_{t=t_k} + \frac{\Delta t^2}{2} \left.\frac{d^2\boldsymbol{f}(\boldsymbol{y}(t), t)}{dt^2}\right|_{t=t_k} + \cdots . \tag{22}$$

Similarly, expand $\boldsymbol{y}_{k+1}$ in a Taylor series

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \Delta t \frac{d\boldsymbol{y}(t_k)}{dt} + \frac{\Delta t^2}{2}\frac{d^2\boldsymbol{y}(t_k)}{dt^2} + \frac{\Delta t^3}{6}\frac{d^3\boldsymbol{y}(t_k)}{dt^3} + \cdots . \tag{23}$$

A substution of (22)-(23) into (21) yields[2] after simple algebraic simplifications

$$\|\boldsymbol{\tau}_{k+1}\| = \frac{2}{3}\Delta t^2 \left\| \frac{d^3\boldsymbol{y}(t_k)}{dt^3} \right\| + o(\Delta t^2) \tag{24}$$

for any vector norm $\|\cdot\|$. Equation (24) shows that the local truncation error $\|\boldsymbol{\tau}_{k+1}\|$ goes to zero as Δt^2 and therefore the Crank-Nicolson method is *consistent with order two.*

General conditions for consistency

We have seen in previous section that the calculation of the local truncation error can be effectively carried out for any given numerical scheme by using Taylor series. Next, we derive a set conditions guaranteeing that the local truncation method of any numerical scheme of the form (2) goes to zero as $\Delta t \to 0$. To this end, let us expand $\boldsymbol{y}_{k+j} = \boldsymbol{y}(t_k + j\Delta t)$ in equation (9) in a first-order Taylor series

$$\boldsymbol{y}_{k+j} = \boldsymbol{y}_k + j\Delta t \frac{d\boldsymbol{y}(t_k)}{dt} + \cdots . \tag{25}$$

Substituting (25) into (9) yields

$$\boldsymbol{\tau}_{k+q} = \frac{\boldsymbol{y}_k}{\Delta t}\sum_{j=0}^{q}\alpha_j + \frac{d\boldsymbol{y}(t_k)}{dt}\sum_{j=0}^{q} j\alpha_j - \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{y}_{k+q}, \ldots, \boldsymbol{y}_k, t_k, \Delta t) + \cdots . \tag{26}$$

In the limit $\Delta t \to 0$ we have that $\boldsymbol{y}_{k+j} \to \boldsymbol{y}_k$ for all $j = 0, \ldots, q$. Assuming that $\boldsymbol{y}_k$ is arbitrary, the previous equation yields the consistency conditions

$$1. \qquad \sum_{j=0}^{q} \alpha_j = 0, \tag{27}$$

$$2. \qquad \frac{\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{y}_k, \ldots, \boldsymbol{y}_k, t_k, 0)}{\displaystyle\sum_{j=0}^{q} j\alpha_j} = \boldsymbol{f}(\boldsymbol{y}_k, t_k). \tag{28}$$

At this point it is convenient to define the *first characteristic polynomial* associated with the numerical method (2)

$$\rho(z) = \sum_{j=0}^{q} \alpha_j z^j \qquad \text{(first characteristic polynomial).} \tag{29}$$

[2] To derive (24) we recall that

$$\frac{d^2\boldsymbol{y}}{dt^2} = \frac{d\boldsymbol{f}(\boldsymbol{y}, t)}{dt} \qquad \frac{d^3\boldsymbol{y}}{dt^3} = \frac{d^2\boldsymbol{f}(\boldsymbol{y}, t)}{dt^2}$$

By using ρ we can write the *consistency conditions* in a more compact form as

$$1. \qquad \rho(1) = 0, \tag{30}$$

$$2. \qquad \frac{\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{y}_k, \dots, \boldsymbol{y}_k, t_k, 0)}{\rho'(1)} = \boldsymbol{f}(\boldsymbol{y}_k, t_k), \tag{31}$$

where $\rho'(z) = d\rho(z)/ds$.

We emphasize that the consistency conditions (27)-(28) (or the equivalent ones (30)-(31)) do not provide *any information on the order of consistency*, but simply allow us to check whether a numerical scheme is consistent or not. Let us provide a few examples.

- **One-step methods**: Consider a general one-step method[3] in the form

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \Delta t \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_{k+1}, \boldsymbol{u}_k, t_k, \Delta t). \tag{32}$$

The first characteristic polynomial associated with (32) is

$$\rho(z) = z - 1. \tag{33}$$

In fact, the scheme (32) can be written in the form (2) if we set $q = 1$, $\alpha_1 = 1$ and $\alpha_0 = -1$. By evaluating $\rho(z)$ and $\rho'(z)$ at $z = 1$ we obtain

$$\rho(1) = 0, \qquad \rho'(1) = 1 \tag{34}$$

Hence, (30) is always satisfied. The second condition, i.e. (31), can be written as

$$\Phi_{\boldsymbol{f}}(\boldsymbol{y}_k, \boldsymbol{y}_k, t_k, 0) = \boldsymbol{f}(\boldsymbol{y}_k, t_k) \quad \text{for all} \quad \boldsymbol{y}_k \in \mathbb{R}^n. \tag{35}$$

This condition is clearly satisfied, e.g., by the Crank-Nicolson method (see the iteration function (5)), by the Euler forward and backward methods, and by the implicit midpoint method. In fact, recall that the iteration function of the implicit midpoint method is

$$\Phi_{\boldsymbol{f}}(\boldsymbol{u}_{k+1}, \boldsymbol{u}_k, t_k, \Delta t) = \boldsymbol{f}\left(\frac{\boldsymbol{u}_{k+1} + \boldsymbol{u}_k}{2}, t_k + \Delta t\right) \quad \Rightarrow \quad \Phi_{\boldsymbol{f}}(\boldsymbol{y}_k, \boldsymbol{y}_k, t_k, 0) = \boldsymbol{f}(\boldsymbol{y}_k, t_k). \tag{36}$$

Regarding Runge-Kutta methods, we recall that their iteration function can be written as

$$\boldsymbol{\Phi}_{\boldsymbol{f}}\left(\boldsymbol{u}_k, \boldsymbol{u}_{k+1}, t_k, \Delta t\right) = \sum_{i=1}^{s} b_i \boldsymbol{K}_i \qquad \boldsymbol{K}_i = \boldsymbol{f}\left(\boldsymbol{u}_k + \Delta t \sum_{j=1}^{s} a_{ij} \boldsymbol{K}_j, t_k + c_i \Delta t\right) \tag{37}$$

Hence, condition (35) implies that Runge-Kutta methods are consistent if and only if

$$\sum_{i=1}^{s} b_i = 1. \tag{38}$$

- **Adams methods**: The general form of a q-step Adams method is

$$\boldsymbol{u}_{k+q} = \boldsymbol{u}_{k+q-1} + \Delta t \sum_{j=0}^{q} \beta_j f(\boldsymbol{u}_{k+j}, t_{k+j}). \tag{39}$$

[3] Recall that all Runge-Kutta methods (implicit and explicit) are in the form (32).

If $\beta_q = 0$ then the method is explicit (Adams-Bashforth). Otherwise it is implicit (Adams-Moulton). The first characteristic polynomial and the iteration function of a q-step Adams method are, respectively

$$\rho(z) = z^q - z^{q-1}, \qquad \boldsymbol{\Phi}_{\boldsymbol{f}}\left(\boldsymbol{u}_{k+q}, \dots, \boldsymbol{u}_k, t_k, \Delta t\right) = \sum_{j=0}^{q} \beta_j \boldsymbol{f}(\boldsymbol{u}_{k+j}, t_k + j\Delta t). \tag{40}$$

Clearly,

$$\rho(1) = 0 \qquad \text{and} \qquad \rho'(z) = qz^{q-1} - (q-1)z^{q-2} \quad \Rightarrow \quad \rho'(1) = 1. \tag{41}$$

Hence, the first consistency condition (30) is always satisfied. The second condition (31) can be written as

$$\Phi_{\boldsymbol{f}}(\boldsymbol{y}_k, \dots, \boldsymbol{y}_k, t_k, 0) = \boldsymbol{f}(\boldsymbol{y}_k, t_k) \sum_{j=0}^{q} \beta_j = \boldsymbol{f}(\boldsymbol{y}_k, t_k), \tag{42}$$

and it is satisfied for all $y_k \in \mathbb{R}^n$ if and only if

$$\sum_{j=0}^{q} \beta_j = 1. \tag{43}$$

- **BDF methods**: The general form of a q-step BDF method is

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = c\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+q}, t_{k+q}). \tag{44}$$

where the constant c takes care of the fact that we set $\alpha_q = 1$. Equivalently, we can say that we set $c = \beta_q$. For BDF methods we have

$$\rho(z) = \sum_{j=0}^{q} \alpha_j z^j, \qquad \boldsymbol{\Phi}_{\boldsymbol{f}}\left(\boldsymbol{u}_{k+q}, t_k, \Delta t\right) = c\boldsymbol{f}(\boldsymbol{u}_{k+q}, t_k + q\Delta t). \tag{45}$$

Therefore the consistency conditions (30)-(31) reduce to

$$\rho(1) = 0, \qquad \rho'(1) = c. \tag{46}$$

For example, the BDF2 method can be written in the form (44) as

$$\boldsymbol{u}_{k+2} - \frac{4}{3}\boldsymbol{u}_{k+1} + \frac{1}{3}\boldsymbol{u}_k = \frac{2}{3}\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+2}, t_{k+2}), \quad \text{i.e.} \quad c = \frac{2}{3}. \tag{47}$$

The first characteristic polynomial is

$$\rho(z) = z^2 - \frac{4}{3}z + \frac{1}{3} \qquad \rho'(z) = 2z - \frac{4}{3}. \tag{48}$$

Clearly, $\rho(1) = 0$ and $\rho'(1) = 2/3 = c$, which implies that BDF2 is consistent.

- **Linear multistep methods (LMMs)**: We have seen that the general form of a linear q-step method is

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = \Delta t \sum_{j=0}^{q} \beta_j \boldsymbol{f}(\boldsymbol{u}_{k+j}, t_{k+j}). \tag{49}$$

In this case the consistency conditions (30)-(31) can be written, respectively, as

$$\rho(1) = \sum_{j=0}^{q} \alpha_j = 0, \qquad \text{and} \qquad \sum_{j=0}^{q} (j\alpha_j - \beta_j) = 0. \tag{50}$$

Order of consistency of general linear multistep methods (LMMs)

We have seen in previous section that there is a simple criterion to check whether a numerical scheme of the form (2) is consistent or not. The criterion is summarized by the conditions (27)-(28), or the equivalent ones (30)-(31) involving the first characteristic polynomial of the scheme. The consistency conditions (27)-(28), however, do not provide any indication on the order of consistency. In this section we derive a theory that allows us to determine the order of consistency for general linear multistep methods of the form

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = \Delta t \sum_{j=0}^{q} \beta_j \boldsymbol{f}(\boldsymbol{u}_{k+j}, t_{k+j}). \tag{51}$$

As we know, this class of methods includes Adams-Bashforth methods, Adams-Moulton methods and BDF methods. We define the following polynomials associated with (51)

$$\rho(z) = \sum_{j=0}^{q} \alpha_j z^j \qquad \text{(First characteristic polynomial)}, \tag{52}$$

$$\sigma(z) = \sum_{j=0}^{q} \beta_j z^j \qquad \text{(Second characteristic polynomial)}. \tag{53}$$

The local truncation error of the linear multistep scheme (51) is

$$\boldsymbol{\tau}_{k+q} = \frac{1}{\Delta t} \sum_{j=0}^{q} \alpha_j \boldsymbol{y}_{k+j} - \sum_{j=0}^{q} \beta_j \boldsymbol{f}(\boldsymbol{y}_{k+j}, t_{k+j}), \tag{54}$$

where $\boldsymbol{y}_k = \boldsymbol{y}(t_k)$ denotes the analytical solution to (1) evaluated at t_k. On the other hand, by evaluating the ODE (1) at t_{k+j} we obtain

$$\boldsymbol{f}(\boldsymbol{y}_{k+j}, t_{k+j}) = \frac{d\boldsymbol{y}(t_{k+j})}{dt}. \tag{55}$$

Substituting (55) into (54) yields

$$\boldsymbol{\tau}_{k+q} = \frac{1}{\Delta t} \underbrace{\sum_{j=0}^{q} \left(\alpha_j \boldsymbol{y}_{k+j} - \Delta t \beta_j \frac{d\boldsymbol{y}(t_{k+j})}{dt} \right)}_{\mathcal{L}\boldsymbol{y}(t_k)}, \tag{56}$$

where we defined the linear difference operator $\mathcal{L}$ as

$$\mathcal{L}\boldsymbol{y}(s) = \sum_{j=0}^{q} \left(\alpha_j \boldsymbol{y}(s + j\Delta t) - \Delta t \beta_j \frac{d\boldsymbol{y}(s + j\Delta t)}{dt} \right). \tag{57}$$

Assuming that $\boldsymbol{y}(t)$ is differentiable with respect to t as many times as we need, we compute the Taylor series

$$\boldsymbol{y}(t_k + j\Delta t) = \boldsymbol{y}(t_k) + j\Delta t \frac{d\boldsymbol{y}(t_k)}{dt} + \frac{(j\Delta t)^2}{2} \frac{d^2\boldsymbol{y}(t_k)}{dt^2} + \cdots \tag{58}$$

$$\frac{d\boldsymbol{y}(t_k + j\Delta t)}{dt} = \frac{d\boldsymbol{y}(t_k)}{dt} + j\Delta t \frac{d^2\boldsymbol{y}(t_k)}{dt^2} + \frac{(j\Delta t)^2}{2} \frac{d^3\boldsymbol{y}(t_k)}{dt^3} + \cdots \tag{59}$$

A substitution of (58)-(59) into (57) yields the following Taylor series

$$\begin{aligned} \mathcal{L}\boldsymbol{y}(t_k) =& \sum_{j=0}^{q} \alpha_j \left(\boldsymbol{y}(t_k) + j\Delta t \frac{d\boldsymbol{y}(t_k)}{dt} + \cdots \right) - \Delta t \beta_j \left(\frac{d\boldsymbol{y}(t_k)}{dt} + j\Delta t \frac{d^2\boldsymbol{y}(t_k)}{dt^2} + \cdots \right) \\ =& C_0 \boldsymbol{y}(t_k) + C_1 \Delta t \frac{d\boldsymbol{y}(t_k)}{dt} + C_2 \Delta t^2 \frac{d^2\boldsymbol{y}(t_k)}{dt^2} + \cdots, \end{aligned} \tag{60}$$

where the coefficients C_p are defined by collecting all terms multiplying $\Delta t^p d^k \boldsymbol{y}(t_k)/dt$, i.e.,

$$C_0 = \sum_{j=0}^{q} \alpha_j = \rho(1) \tag{61}$$

$$C_1 = \sum_{j=0}^{q} (j\alpha_j - \beta_j) = \rho'(1) - \sigma(1) \tag{62}$$

$$\vdots$$

$$C_s = \frac{1}{s!} \sum_{j=0}^{q} \left(j^s \alpha_j - s j^{s-1} \beta_j \right) \qquad s = 2, 3, \dots \tag{63}$$

Dividing $\boldsymbol{\mathcal{L}y}(t_k)$ by Δt (see (56)) yields the following series expansion of the local truncation error

$$\boldsymbol{\tau}_{k+q} = \frac{C_0}{\Delta t} \boldsymbol{y}(t_k) + C_1 \frac{d\boldsymbol{y}(t_k)}{dt} + C_2 \Delta t \frac{d^2 \boldsymbol{y}(t_k)}{dt^2} + C_3 \Delta t^2 \frac{d^3 \boldsymbol{y}(t_k)}{dt^3} \cdots . \tag{64}$$

We have seen that a necessary condition for consistency is that $\rho(1) = 0$ (see Eq. (30)), and therefore $C_0 = 0$. For consistency we also need to have $C_1 = 0$ (otherwise the truncation error (64) does not go to zero as $\Delta t \to 0$). Clearly, if for a certain scheme the coefficients $C_1, \dots, C_p$ are all zero and $C_{p+1} \neq 0$ then we see that the linear multistep method is *consistent with order p*. In fact, in this case we have (to leading order in Δt)

$$\|\boldsymbol{\tau}_{k+q}\| \leq C_{p+1} \Delta t^p \left\| \frac{d^{p+1} \boldsymbol{y}(t_k)}{dt^{p+1}} \right\| + O(\Delta t^{p+1}). \tag{65}$$

- **Order of consistency of AB3**: The AB3 scheme can be written as

$$\boldsymbol{u}_{k+3} = \boldsymbol{u}_{k+2} + \frac{\Delta t}{12} \left(23 \boldsymbol{f}_{k+2} - 16 \boldsymbol{f}_{k+1} + 5 \boldsymbol{f}_k \right). \tag{66}$$

  The characteristic polynomials (52)-(53) associated with (66) are

$$\rho(z) = z^3 - z^2 \qquad \sigma(z) = \frac{23}{12} z^2 - \frac{4}{3} z + \frac{5}{12}. \tag{67}$$

  By using (61)-(64) we obtain

$$C_0 = \rho(1) = 0 \tag{68}$$

$$C_1 = \rho'(1) - \sigma(1) = 1 - \frac{23}{12} + \frac{16}{12} - \frac{5}{12} = 0, \tag{69}$$

$$C_2 = \frac{1}{2} \left[3^2 - 2^2 - 2 \left(-\frac{16}{12} + 2 \frac{23}{12} \right) \right] = 0, \tag{70}$$

$$C_3 = \frac{1}{6} \left[3^3 - 2^3 - 3 \left(-\frac{16}{12} + 2^2 \frac{23}{12} \right) \right] = 0, \tag{71}$$

$$C_4 = \frac{1}{24} \left[3^4 - 2^4 - 4 \left(-\frac{16}{12} + 2^3 \frac{23}{12} \right) \right] = \frac{9}{24}. \tag{72}$$

  Therefore AB3 is consistent with order 3.

- **Order of consistency of BDF2**: The BDF2 scheme can be written as

$$\boldsymbol{u}_{k+2} - \frac{4}{3} \boldsymbol{u}_{k+1} + \frac{1}{3} \boldsymbol{u}_k = \frac{2}{3} \Delta t \boldsymbol{f}_{k+2}. \tag{73}$$

The characteristic polynomials (52)-(53) associated with (66) are

$$\rho(z) = z^2 - \frac{4}{3}z + \frac{1}{3} \qquad \sigma(z) = \frac{2}{3}z^2. \tag{74}$$

By using (61)-(64) we obtain

$$C_0 = \rho(1) = 0 \tag{75}$$

$$C_1 = \rho'(1) - \sigma(1) = \frac{2}{3} - \frac{2}{3} = 0, \tag{76}$$

$$C_2 = \frac{1}{2}\left[2^2 - 1^2\frac{4}{3} - 4\frac{2}{3})\right] = 0, \tag{77}$$

$$C_3 = \frac{1}{3}\left[2^3 - 1^3\frac{4}{3} - 12\frac{2}{3})\right] = -\frac{4}{9}. \tag{78}$$

Therefore BDF2 is consistent with order 2.

Maximum order of consistency of LMMs. What is the maximum order of consistency attainable by a q-step linear multistep method (LMM)? The scheme (51) is fully determined by the $2q+1$ coefficients (recall that we set $\alpha_q = 1$)

$$\{\alpha_{q-1}, \ldots, \alpha_0, \beta_q, \ldots, \beta_0\}. \tag{79}$$

If the method is consistent with order p then we $p+1$ conditions (see Eq. (64))

$$C_0 = 0, \quad C_1 = 0, \quad \cdots, \quad C_p = 0. \tag{80}$$

By setting $2q+1 = p+1$ we see that

Theorem 1. The maximum order of consistency attainable by q-step linear method of the form (51) is

$$p = 2q \quad \text{(implicit LMM methods)}, \qquad p = 2q-1 \quad \text{(explicit LMM methods)} \tag{81}$$

In particular, for Adams-Bashforth and Adams-Moulton methods we have the following result.

Theorem 2. The maximum order of consistency attainable by a q-step Adams-Bashforth method is $p = q$. Similarly, the maximum order of consistency attainable by a q-step Adams-Moulton method is $p = q+1$

In fact, the condition $\alpha_q = -\alpha_{q-1}$ automatically guarantees that $\rho(1) = C_0 = 0$. Therefore a q-step Adams-Bashforth has q free parameters $\{\beta_0, \beta_1, \ldots, \beta_{q-1}\}$, which can be chosen to satisfy the conditions $C_1 = 0$, $C_2 = 0$ up to $C_p = 0$ (p equations). This implies that a q-step Adams-Bashforth method has maximal order of consistency $p = q$.

As we will see in the next course note, Adams-Bashforth and Adams-Moulton methods are all zero-stable, and therefore $p = q$ and $p = q+1$ is actually the convergence order of these methods. On the other hand, for general LMM methods, it is possible to prove that LMM methods with order of consistency exceeding $p = q+1$ (q odd) or $p = q+2$ (q even) are all *zero-unstable* and therefore not convergent. This result is known as *first Dahlquist barrier* for LMM methods.

Order of consistency of RK methods

The order of an RK method (like the order of any other method) can be determined by using a Taylor series analysis of the local truncation error (see the examples at the beginning of this note). On the other hand, if we are interested in developing an explicit or implicit RK method with a maximal order of consistency, we can just expand the RK method in a Taylor series and then try to match as many powers of Δt as possible relative to a power series expansion of the exact solution. We have already seen one of such calculations when we derived the one-parameter family of explicit two-stage RK methods. Obtaining similar results for RK methods with a larger number of stages is quite cumbersome[4], and also yields surprising results. In general, it can be shown that:

Theorem 3. An s-stage explicit RK method cannot have order greater than s.

This theorem establishes an upper bound for the maximum order attainable by explicit RK methods. However, determining the maximum attainable order for a fixed number of stages is not a trivial problem. Order conditions similar to those derived for LMM methods, i.e., (61)-(63) can be derived for RK methods using Butcher's theory. For instance, it can be shown that for the three-stage explicit RK method

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & a_{21} & 0 & 0 \\ c_3 & a_{31} & a_{32} & 0 \\ \hline & b_1 & b_2 & b_3 \end{array}$$

to be of order 3 the following *stage-order conditions* need to be satisfied

$$\begin{cases} b_1 + b_2 + b_3 = 1 \\ b_2 c_2 + b_3 c_3 = \dfrac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 = \dfrac{1}{3} \\ b_3 a_{32} c_2 = \dfrac{1}{6} \end{cases} \tag{82}$$

The solution to this algebraic nonlinear system yields a one two-parameter family of solutions and two one-parameter families of solutions (see [1, p. 178]). A similar calculation performed on a five-stage explicit RK method yields a system of order conditions that has no solution (see [1, p. 181]). In other words:

Theorem 4. There exist no five-stage explicit RK method of order 5.

The following table summarizes the maximum order attainable by an *explicit* RK method with s stages:

order of consistency	1	2	3	4	5	6	7	8
Minimum number of stages	1	2	3	4	6	7	9	11

Regarding *implicit* RK methods, the highest attainable order is $2s$ (Gauss-RK methods). Similarly, Gauss-Radau and Gauss-Lobatto RK methods can attain order of consistency $2s-1$ and $2s-2$, respectively.

References

[1] J. D. Lambert. *Numerical methods for ordinary differential systems: the initial value problem.* Wiley, 1991.

[4] The Taylor series analysis can be simplified substantially by using by using Butcher's theory (see [1, §5.6]), which relies on rooted trees (graphs) techniques.