Stability and convergence of numerical methods for ODEs

Consider the initial value problem for a system of ODEs

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

and the perturbed problem

$$\begin{cases} \dfrac{d\boldsymbol{z}}{dt} = \boldsymbol{f}(\boldsymbol{z}, t) + \boldsymbol{\delta}(t) \\ \boldsymbol{z}(0) = \boldsymbol{y}_0 + \boldsymbol{\delta}_0 \end{cases} \tag{2}$$

where $\boldsymbol{\delta}(t)$ is an integrable function and $\boldsymbol{\delta}_0 \in \mathbb{R}^n$. Note that we replaced $\boldsymbol{y}(t)$ with $\boldsymbol{z}(t)$ in (2) to emphasize the fact that the solutions to (1) and (2) are (in general) different.

Definition 1 (Stability of the Cauchy problem (see [1, 3]))**.** The Cauchy problem (1) is said to be *stable* within the time interval $[0, T]$ if for any perturbations $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}(t)$ such that[1]

$$\|\boldsymbol{\delta}_0\| \leq \epsilon, \quad \text{and} \quad \|\boldsymbol{\delta}(t)\| \leq \epsilon \quad \text{for all } t \in [0, T] \tag{3}$$

we have that

$$\|\boldsymbol{z}(t) - \boldsymbol{y}(t)\| \leq C\epsilon, \quad \text{for all } t \in [0, T], \tag{4}$$

where C is a finite constant that does not depend on ϵ.

Based on this definition, the Cauchy problem (1) is "stable" if the difference between the solutions of (1) and (2) is bounded in $[0, T]$ after we introduce a small perturbation $\boldsymbol{\delta}_0$ in the initial condition $\boldsymbol{y}_0$ and a perturbation $\boldsymbol{\delta}(t)$ in $\boldsymbol{f}(\boldsymbol{y}, t)$. The definition of stability also implies that the difference between the solutions of (1) and (2) goes to zero as $\epsilon \to 0$. In fact, from (4) it follows that

$$\lim_{\epsilon \to 0} \|\boldsymbol{z}(t) - \boldsymbol{y}(t)\| \leq C \lim_{\epsilon \to 0} \epsilon = 0. \tag{5}$$

The constant C appearing in (4) may not be small. This is consistent with the fact that a small perturbations in the Cauchy problem (1) can introduce large perturbations in its solution.

Hereafter we show that any well-posed initial value problem (1) is stable, i.e., robust to perturbations in the limit of small perturbations.

Theorem 1. Let $D \subseteq \mathbb{R}^n$ be an open set, $\boldsymbol{y}_0 \in D$. If $\boldsymbol{f}(\boldsymbol{y}, t)$ is Lipschitz continuous in D and $\boldsymbol{\delta}(t)$ is integrable then the initial value problem (1) is stable.

Proof. We need to show that for any $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}(t)$ the difference between the solutions of (1) and (2) is bounded in some time interval $[0, T]$ and that the difference goes to zero as we send ϵ to zero. First of all, we notice that if D is open and ϵ is small enough then the initial condition $(\boldsymbol{y}_0 + \boldsymbol{\delta}_0)$ is in D. If $\boldsymbol{f}(\boldsymbol{y}, t)$ is Lipschitz continuous

$$\|\boldsymbol{f}(\boldsymbol{z}, t) - \boldsymbol{f}(\boldsymbol{y}, t)\| \leq L \|\boldsymbol{z} - \boldsymbol{y}\| \quad \forall \boldsymbol{z}, \boldsymbol{y} \in D, \tag{6}$$

[1] Note that in (3) we are bounding $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}(t)$ with the same constant ϵ. Such a constant coincides with the radius of the largest ball centered at zero in $\mathbb{R}^n$ that includes both $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}(t)$ (for all $t \in [0, T]$).

and $\boldsymbol{\delta}(t)$ is integrable, then we have existence and uniqueness of the solution to both problems (1) and (2). Such problems can be equivalently written as

$$\boldsymbol{y}(t) = \boldsymbol{y}_0 + \int_0^t \boldsymbol{f}(\boldsymbol{y}(s), s)ds \tag{7}$$

$$\boldsymbol{z}(t) = \boldsymbol{y}_0 + \boldsymbol{\delta}_0 + \int_0^t \boldsymbol{f}(\boldsymbol{z}(s), s)ds + \int_0^t \boldsymbol{\delta}(s)ds \tag{8}$$

for all $t \in [0, T]$, where T is is the smallest "exit time", i.e., the time in which either $\boldsymbol{y}(t)$ or $\boldsymbol{z}(t)$ get out of D. Subtracting (7) from (8) and taking the norm yields

$$\begin{aligned}\|\boldsymbol{z}(t) - \boldsymbol{y}(t)\| =& \left\| \boldsymbol{\delta}_0 + \int_0^t \left[\boldsymbol{f}(\boldsymbol{z}(s), s) - \boldsymbol{f}(\boldsymbol{y}(s), s)\right] ds + \int_0^t \boldsymbol{\delta}(s)ds \right\| \\ \leq& \|\boldsymbol{\delta}_0\| + \int_0^t \|\boldsymbol{f}(\boldsymbol{z}(s), s) - \boldsymbol{f}(\boldsymbol{y}(s), s)\| \, ds + \int_0^t \|\boldsymbol{\delta}(s)\| \, ds \\ \leq& (1+t)\epsilon + L \int_0^t \|\boldsymbol{z}(s) - \boldsymbol{y}(s)\| \, ds \end{aligned} \tag{9}$$

where we used the triangle inequality, the inequalities (3), and the definition of Lipschitz continuity (6). At this point we use Grönwall's inequality[2] to conclude that

$$\begin{aligned}\|\boldsymbol{z}(t) - \boldsymbol{y}(t)\| \leq& (1+t)e^{tL}\epsilon \\ \leq& \underbrace{(1+T)e^{TL}}_{C} \epsilon. \end{aligned} \tag{13}$$

This proves that the Cauchy problem (1) is stable. Note that the constant C appearing in (13) does not depend on ϵ and it can be very big, depending on the Lipschitz constant L and the integration time T (integration time).

□

Remark: If we replace the initial value problem (1) by a numerical scheme we introduce errors that can accumulate in time. Such errors can be considered as perturbations in the ODE (1). Just think about representing the discrete numerical solution $\boldsymbol{u}_k$ in terms of some local interpolant $\boldsymbol{z}(t)$ (differentiable in time) and substituting it into (1). This yields an ODE in the form (2). If the initial value problem (1) is not stable, i.e., robust to small perturbations, then there is no hope for any numerical method to approximate the solution.

Stability and zero-stability of numerical methods for ODEs. The concept of stability we discussed in previous section for continuous-time dynamical systems can be extended to discrete-time dynamical

[2]The Grönwall's inequality (see [3, Lemma 11.1]) can be stated as follows. Let $u(t)$, $g(t)$ and $p(t)$ such that

$$u(t) \leq g(t) + \int_0^t p(s)u(s)ds. \tag{10}$$

If $g(t)$ is non-decreasing and $p(s)$ is strictly positive then

$$u(t) \leq g(t) \exp\left[\int_0^t p(s)u(s)ds\right]. \tag{11}$$

In the case of equation (9) we have

$$u(t) = \|\boldsymbol{z}(t) - \boldsymbol{y}(t)\|, \qquad g(t) = (1+t)\epsilon, \qquad p(s) = L. \tag{12}$$

systems, i.e., to numerical schemes aiming at computing the approximate solution of the initial value problem (1). we have seen that such schemes can be written in the general form[3]

$$\begin{cases} \displaystyle\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = \Delta t \boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_{k+q}, \ldots, \boldsymbol{u}_k, t_k, \Delta t), \\ \text{given} \quad \{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{q-1}\} \end{cases} \tag{14}$$

where $\Phi_{\boldsymbol{f}}$ is some iteration function. By taking a perturbation of the initial condition $\boldsymbol{u}_0$ and a "time-dependent" perturbation of $\Phi_{\boldsymbol{f}}$ in (14) we obtain

$$\begin{cases} \displaystyle\sum_{j=0}^{q} \alpha_j \boldsymbol{z}_{k+j} = \Delta t \left[\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{z}_{k+q}, \ldots, \boldsymbol{z}_k, t_k, \Delta t) + \boldsymbol{\delta}_{k+q}\right], \\ \text{given} \quad \{\boldsymbol{z}_0 = \boldsymbol{u}_0 + \boldsymbol{\delta}_0, \ldots, \boldsymbol{z}_{q-1} = \boldsymbol{u}_{q-1} + \boldsymbol{\delta}_{q-1}\} \end{cases} \tag{15}$$

where $\{\boldsymbol{\delta}_0, \ldots, \boldsymbol{\delta}_{k+q}, \ldots\}$ is a sequence of vectors in $\mathbb{R}^n$ bounded by some constant ϵ, i.e.,

$$\|\boldsymbol{\delta}_j\| \leq \epsilon \quad \text{for all} \quad j = 0, 1, \ldots. \tag{16}$$

The perturbations $\boldsymbol{\delta}_j$ can arise, e.g., because of round-off or truncation errors when performing floating point operations using double precision arithmetic. Clearly, the orbits generated by (14) and (15), i.e.,

$$\{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_N\} \quad \text{and} \quad \{\boldsymbol{z}_0, \ldots, \boldsymbol{z}_N\} \tag{17}$$

are (in general) different. For any given iteration function $\boldsymbol{\Phi}_{\boldsymbol{f}}$ and any given Δt we can provide a definition of stability for the numerical scheme (14) that closely resembles Definition 1 for continuous time dynamical systems. To this end, let T be the period of integration, and N be the number of time steps, i.e.,

$$\Delta t^* = \frac{T}{N} \tag{18}$$

Of particular interest when performing convergence analysis, is the behavior of the scheme for small Δt, i.e., for all Δt smaller than Δt^*.

Definition 2 (Zero-stability)**.** We say that the numerical scheme (14) is *zero-stable* if there exists a $\Delta t^* > 0$ such that *for all* $\Delta t \leq \Delta t^*$ and for any perturbations $\boldsymbol{\delta}_j$ $(j = 0, \ldots, N)$ such that

$$\|\boldsymbol{\delta}_k\| \leq \epsilon, \quad \text{for all } j = 0, \ldots, N \tag{19}$$

we have that

$$\|\boldsymbol{z}_k - \boldsymbol{u}_k\| \leq C\epsilon, \quad \text{for all } j = 0, \ldots, N, \tag{20}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{z}_k$ are defined by (14) and (15), and C is a finite constant that does not depend on ϵ[4].

The definition "zero-stability" follows from the fact that we require $\|\boldsymbol{z}_k - \boldsymbol{u}_k\| \leq C\epsilon$ for all $\Delta t \leq \Delta t^*$, and in particular for $\Delta t \to 0$. Hence the "zero" part in "zero-stability" refers to the stability of the scheme in the limit $\Delta t \to 0$.

- Zero stability is a property of the numerical scheme, not of the ODE system (1). We have seen, in fact, that a well-posed Cauchy problem is always stable.

[3] In (14) we set $\alpha_q = 1$ to remove the non-uniqueness of α_j and β_j due to possible rescaling by a constant.
[4] The constant C in (20) may depend also on T, Δt or other constants, but it cannot depend on k or ϵ.

- Numerical methods that are not zero-stable have no hope to reliably approximate the solution of (1). In fact, even if the method is consistent, i.e., if the truncation error goes to zero as $\Delta t \to 0$, we have that perturbations due to finite-arithmetic may rapidly propagate in schemes that are not zero-stable, and therefore generate instabilities. In other words, consistent schemes that are not zero-stable may not converge as $\Delta t \to 0$. For example, the numerical scheme in equation (32) hereafter is consistent but not zero stable. Another example of a consistent scheme that is not zero stable is discussed in [2, p. 32].

The root condition and zero-stability. The numerical method (14) is said to satisfy the *root condition* if all roots of the first characteristic polynomial

$$\rho(z) = \sum_{j=0}^{q} \alpha_j z^j \tag{21}$$

are within the unit circle, and those of modulus one (i.e., the ones on the unit circle) are simple. The following fundamental theorem relates zero stability of the numerical method (14) to the root condition.

Theorem 2. The numerical method (14) is zero-stable if and only if it satisfies the root condition.

A detailed proof of this theorem is provided at the end of this note[5]. Recall that a necessary condition for consistency is that $\rho(1) = 0$, i.e., $z = 1$ is a root of (21). Such a root must be simple in order for the method to satisfy the root condition. Let us now study zero-stability of all schemes we have considered so far.

- ***One-step methods:*** The most general form of a one-step method is

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \Delta t \Phi_{\boldsymbol{f}}(\boldsymbol{u}_{k+1}, \boldsymbol{u}_k, t_k, \Delta t). \tag{22}$$

  The characteristic polynomial for this class of methods is

$$\rho(z) = z - 1 \tag{23}$$

  Clearly, $\rho(z)$ has a simple root at $z = 1$ and therefore (22) satisfies the root condition. This implies that *all one-step methods are zero-stable*. Recall that all Runge-Kutta methods are one-step methods.

- ***Adams-Bashforth and Adams-Moulton methods:*** A q-step Adams method can be written in the general form

$$\boldsymbol{u}_{k+q} = \boldsymbol{u}_{k+q-1} + \Delta t \sum_{j=0}^{q} \beta_j \boldsymbol{f}\left(\boldsymbol{u}_{k+j}, t_{k+j}\right). \tag{24}$$

  For Adams-Bashforth methods (explicit) we have $\beta_q = 0$; for Adams-Moulton (implicit) $\beta_q \neq 0$. The characteristic polynomial associated with (24) is

$$\rho(z) = z^q - z^{q-1} = z^{q-1}(z-1). \tag{25}$$

  This polynomial has as a simple root at $z = 1$ and a root with algebraic multiplicity $q - 1$ at $z = 0$. Therefore it satisfies the root condition. By Theorem (2) we have that *all Adams-Bashforth and all Adams-Moulton methods are zero-stable.*

[5] For whatever reason, none of the books I came across in my career provides concise and direct proof of Theorem 2 in the general case we are considering here, i.e., for vector-valued ODEs and numerical methods of the form (14). Hence, I decided to provide my own version of the proof.

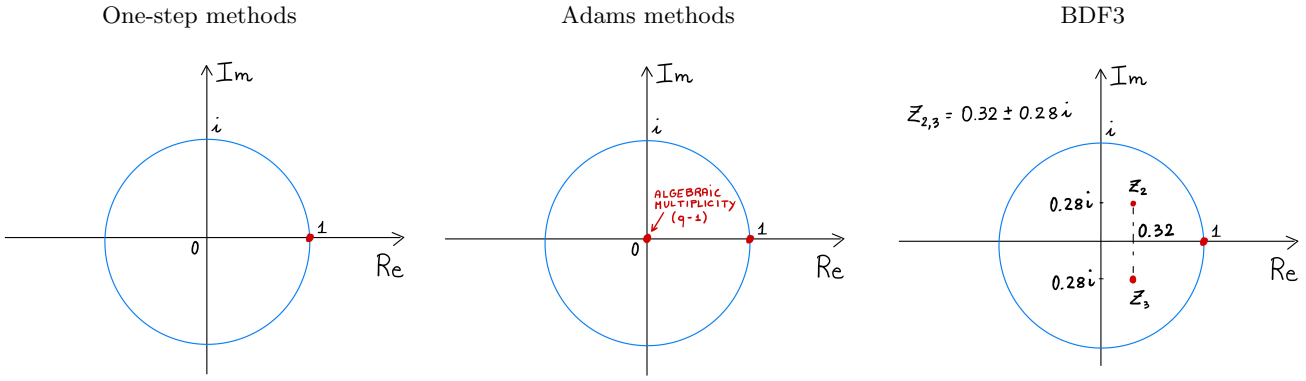

Figure 1: Roots of the characteristic polynomial (21). If all roots are within the unit circle and those modulus one (i.e., the ones on the unit circle) are simple (i.e., they have algebraic multiplicity one) then the method is zero-stable. All methods sketched in this figure are zero-stable.

- ***BDF methods:*** We know that a q-step BDF method can be written in the form

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = c\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+q}, t_{k+q}). \tag{26}$$

  The characteristic polynomial associated with (26) is

$$\rho(z) = z^q + \alpha_{q-1} z^{q-1} + \cdots \alpha_0. \tag{27}$$

  It can be shown that a q-step BDF method satisfies the root condition and therefore it is zero-stable if and only if $q \leq 6$.

- ***2-step midpoint method:*** The 2-step midpoint method

$$\boldsymbol{u}_{k+2} = \boldsymbol{u}_k + 2\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1}) \tag{28}$$

  satisfies the root condition and therefore it is zero-stable. In fact, the characteristic polynomial associated with (28) is

$$\rho(z) = z^2 - 1. \tag{29}$$

  The roots $z = \pm 1$ are both simple and sitting at the boundary of the unit circle in the complex plane. As we will see, a scheme that satisfies the root condition with simple eigenvalues at boundary of the unit circle is theoretically zero-stable, but in practical applications it can generate instabilities.

- ***2-step LMM method:*** The following two-step explicit linear multi-step method[6]

$$\boldsymbol{u}_{k+2} - 4\boldsymbol{u}_{k+1} + 3\boldsymbol{u}_k = -2\Delta t \boldsymbol{f}(\boldsymbol{u}_k, t_k) \tag{32}$$

  is consistent but *not zero-stable*. The characteristic polynomial is

$$\rho(z) = z^2 - 4z + 3. \tag{33}$$

[6]The method (32) is *not* a BDF method, and is obtained by approximating $d\boldsymbol{y}(t_k)/dt$ with a second-order forward finite difference formula:

$$\frac{d\boldsymbol{y}(t_k)}{dt} \simeq \frac{-3\boldsymbol{y}(t_k) + 4\boldsymbol{y}(t_k + 1) - \boldsymbol{y}(t_{k+2})}{2\Delta t}. \tag{30}$$

and setting the equality

$$\frac{-3\boldsymbol{u}_k + 4\boldsymbol{u}_{k+1} - \boldsymbol{u}_{k+2}}{2\Delta t} = f(\boldsymbol{u}_k, t_k). \tag{31}$$

Consistency can be checked immediately, since (see course note number 3)

$$\rho(1) = 0 \qquad \frac{\boldsymbol{\Phi}(\boldsymbol{u}_k, t_k, 0)}{\rho'(z)} = f(\boldsymbol{u}_k, t_k). \tag{34}$$

The polynomial (33) has roots $z = 1$ and $z = 3$. Therefore the method (32) is not zero-stable.

- ***General LMM methods:*** We have seen in the course note 3 that the maximal order of consistency of a linear q-step method of the form

$$\sum_{j=0}^{q} \alpha_j \boldsymbol{u}_{k+j} = \Delta t \sum_{j=0}^{q} \beta_j \boldsymbol{f}(\boldsymbol{u}_{k+j}, t_{k+1}), \tag{35}$$

is $2q$ (implicit methods) or $2q-1$ (explicit methods). At this point we notice that such maximal order LMM methods are, in general, *zero-unstable*, i.e., they do not satisfy the root-condition (see [2, §3.4]). It fact the following theorem holds true.

Theorem 3 (First Dahlquist barrier - 1956)**.** There is no zero-stable linear q-step method with consistency order exceeding $q+1$ (q odd) or $q+2$ (q even).

Zero-stable linear q-step implicit methods with order $q+2$ are called *optimal*. These methods have all roots with algebraic multiplicity one sitting on the boundary of the unit circle. This can yields stability issues.

Convergence. Let $T = N\Delta t$ be period of integration. We say that the scheme (14) is convergent if the error (in any norm)

$$\max_{k \in 0,\dots,N} \|\boldsymbol{u}_k - \boldsymbol{y}_k\| \tag{36}$$

goes to zero as $\Delta t \to 0$. Here $\boldsymbol{y}_k = \boldsymbol{y}(t_k)$ represents the analytical solution of the ODE system (1) evaluated at $t = t_k$, while $\boldsymbol{u}_k$ is the numerical solution produced by the scheme (14). If the error decreases as Δt^p then we say that the scheme converges with order p.

If a numerical scheme is convergent then the order of convergence is the same as the order of consistency (see the proof of theorem 4 at the end of this note). Indeed the error (36) can be bounded by the norm of the global truncation error, which goes to zero to some order in Δt (if the scheme is consistent) The following fundamental theorem provides necessary and sufficient conditions for convergence of numerical method for a system of ODEs.

Theorem 4 (Convergence)**.** The numerical method (14) is convergent if and only if is consistent and zero-stable. In other words,

$$\text{convergence} \;\Leftrightarrow\; \text{consistency} \;+\; \text{zero stability}. \tag{37}$$

Moreover, the convergence order coincides with the consistency order.

The proof of this theorem follows exactly the same steps as the proof of theorem 2, and it is briefly discussed at the end of this note. This theorem has several corollaries. For instance, we have just seen that all one-step methods are zero-stable and therefore we have that:

Corollary 1. A one-step method is convergent if and only if it is consistent.

This means that in order to prove convergence of a one-step method it is necessary and sufficient to prove consistency. Hence, in the case of RK methods a necessary and sufficient condition for convergence is

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

Corollary 2. Adams methods are convergent if and only if they are consistent.

In fact, we have seen that Adams methods are always zero-stable and therefore consistency implies convergence. Recall that Adams-Bashforth and Adams-Moulton methods are consistent if and only if

$$\sum_{j=0}^{q} \beta_j = 1. \tag{39}$$

Hence, if (39) is satisfied then Adams-Bashforth ($\beta_q = 0$) or Adams-Moulton ($\beta_q \neq 0$) methods are convergent.

Example: The numerical scheme (32) is not convergent. In fact, it is consistent, but not zero-stable.

Estimating the convergence order of a numerical method. To estimate the convergence order of the scheme (14) numerically it is sufficient to compute the error $\|\boldsymbol{y}(t_k) - \boldsymbol{u}_k\|$ (in any norm) relative to an analytical solution $\boldsymbol{y}(t)$ for various (sufficiently small) Δt, and then plot

$$\max_{k=1,\ldots,N} \|\boldsymbol{y}(t_k) - \boldsymbol{u}_k\|$$

versus Δt in a logarithmic scale. The slope of the line obtained in this way represents the order of the method. In fact, suppose that for sufficiently small Δt we have

$$\max_{k=1,\ldots,N} \|\boldsymbol{y}(t_k) - \boldsymbol{u}_k\| \simeq C\Delta t^p. \tag{40}$$

Taking the logarithm yields

$$\log\left(\max_{k=1,\ldots,N} \|\boldsymbol{y}(t_k) - \boldsymbol{u}_k\|\right) \simeq \log(C) + p\log(\Delta t) \tag{41}$$

which represents a line with slope p in a log-log plot. To compute the error, we need of course the analytical solution to the initial value problem (1), which is not always available. However, it is very easy to manufacture an ODE with a time-dependent right hand side that has any desired solution $\boldsymbol{y}(t)$. To this end, choose any continuously differentiable vector $\boldsymbol{y}(t)$ and any Lipschitz continuous function $\boldsymbol{f}(\boldsymbol{y})$. Compute the time forcing term

$$\boldsymbol{h}(t) = \frac{d\boldsymbol{y}(t)}{dt} - \boldsymbol{f}(\boldsymbol{y}(t)). \tag{42}$$

Then the chosen $\boldsymbol{y}(t)$ is the analytical solution to the initial value problem

$$\begin{cases} \dfrac{d\boldsymbol{y}}{dt} = \boldsymbol{f}(\boldsymbol{y}) + \boldsymbol{h}(t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases} \tag{43}$$

In this way, for each given Δt we can solve (43) using the numerical method (14) and compute the error $\|\boldsymbol{u}_k - \boldsymbol{y}(t_k)\|$.

Proof of Theorem 2. Let us consider the m-th component of the perturbed scheme (14)

$$\sum_{j=0}^{q} \alpha_j z_{k+j}^m = \Delta t \left[\Phi_{\boldsymbol{f}}^m(\boldsymbol{z}_{k+q}, \ldots, \boldsymbol{z}_k, t_k, \Delta t) + \delta_{q+k}^m \right]. \tag{44}$$

and the unperturbed one

$$\sum_{j=0}^{q} \alpha_j u_{k+j}^m = \Delta t \Phi_{\boldsymbol{f}}^m(\boldsymbol{u}_{k+q}, \ldots, \boldsymbol{u}_k, t_k, \Delta t). \tag{45}$$

Subtracting (44) from (44) yields

$$\sum_{j=0}^{q} \alpha_j e_{k+j}^m = \Delta t \left[\Phi_{\boldsymbol{f}}^m(\boldsymbol{z}_{k+q}, \ldots, \boldsymbol{z}_k, t_k, \Delta t) - \Phi_{\boldsymbol{f}}^m(\boldsymbol{u}_{k+q}, \ldots, \boldsymbol{u}_k, t_k, \Delta t) + \delta_{q+k}^m \right], \tag{46}$$

where

$$e_{k+j}^m = z_{k+j}^m - u_{k+j}^m. \tag{47}$$

Upon definition of

$$\boldsymbol{e}_k^m = \begin{bmatrix} e_k^m \\ e_{k+1}^m \\ \vdots \\ e_{k+q-1}^m \end{bmatrix}, \quad \boldsymbol{b}_k^m = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \Phi_{\boldsymbol{f}}^m(\boldsymbol{z}_{k+q}, \ldots, \boldsymbol{z}_k, t_k, \Delta t) - \Phi_{\boldsymbol{f}}^m(\boldsymbol{u}_{k+q}, \ldots, \boldsymbol{u}_k, t_k, \Delta t) \end{bmatrix}, \quad \boldsymbol{d}_k^m = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \delta_{q+k}^m \end{bmatrix} \tag{48}$$

we see that we can write (46) in a compact form as as[7]

$$\boldsymbol{e}_{k+1}^m = \boldsymbol{A}\boldsymbol{e}_k^m + \Delta t \left(\boldsymbol{b}_k^m + \boldsymbol{d}_k^m \right), \tag{49}$$

where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \cdots & -\alpha_{q-1} \end{bmatrix}. \tag{50}$$

By using the discrete variation of constant formula (in which we treat $\Delta t \left(\boldsymbol{b}_k^m + \boldsymbol{d}_k^m \right)$ as a "forcing term") we write the formal solution of (49) as

$$\boldsymbol{e}_{k+1}^m = \boldsymbol{A}^{k+1} \boldsymbol{e}_0^m + \Delta t \sum_{p=0}^{k} \boldsymbol{A}^{k-p} \left(\boldsymbol{b}_p^m + \boldsymbol{d}_p^m \right), \tag{51}$$

As we shall see hereafter, the zero-stability of the numerical scheme is essentially determined by the properties of the matrix $\boldsymbol{A}$, in particular by the behavior of the matrix powers $\boldsymbol{A}^k$ as k is increased. If the norm of the matrix powers can be bounded by a constant that is independent of k then zero-stability follows rather straightforwardly. The properties of the matrix powers $\boldsymbol{A}^k$ are fully determined by the roots of the first characteristic polynomial (21).

[7] Recall that $\alpha_q = 1$.

Lemma 1. Let $\|\cdot\|$ be any matrix norm compatible with a vector norm. Then $\left\|\boldsymbol{A}^k\right\|$ can be bounded by a quantity M that does not depend on k, i.e.,

$$\left\|\boldsymbol{A}^k\right\| \leq M \quad \text{for all } k \in \mathbb{N} \tag{52}$$

if and only if the root condition is satisfied.

Proof. The matrix $\boldsymbol{A}$ in (50) is the transpose of the companion matrix associated with the characteristic polynomial (21). This means that the eigenvalues of $\boldsymbol{A}$ coincide with the roots of the polynomial (21). Moreover, companion matrices are *non-derogatory*, which means that there exists only one eigenvector corresponding to each eigenvalue λ. Such eigenvector is explicitly obtained as

$$\boldsymbol{h} = \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{q-1} \end{bmatrix}. \tag{53}$$

The non-derogatory property of $\boldsymbol{A}$ implies that if there exists any eigenvalue with algebraic multiplicity $r_j > 1$, then the corresponding eigenspace has dimension $r_j - 1$. This means that the matrix $\boldsymbol{A}$ is diagonalizable (i.e., similar to a diagonal matrix), if and only of all the eigenvalues are simple. If there exist any eigenvalue with multiplicity larger than one then the matrix $\boldsymbol{A}$ is similar to a (block-diagonal) Jordan matrix $\boldsymbol{J}$

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1} \tag{54}$$

where $\boldsymbol{P}$ is the matrix that has the generalized eigenvectors of $\boldsymbol{A}$ columnwise and

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{J}_l \end{bmatrix}, \qquad \boldsymbol{J}_i = \lambda_i \boldsymbol{I}_{r_i} + \boldsymbol{B}_{r_i}. \tag{55}$$

In this equation, $\boldsymbol{J}_i$ denotes the Jordan block corresponding to the eigenvalue λ_i (which has algebraic multiplicity r_i), $\boldsymbol{I}_{r_i}$ is a $r_i \times r_i$ identity matrix and $\boldsymbol{B}_{r_i}$ is a $r_i \times r_i$ matrix with ones above the main diagonal. For instance, if λ_i has algebraic multiplicity $r_i = 3$ then the geometric multiplicity is 2 and we have

$$\boldsymbol{J}_i = \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix}, \qquad \boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{B}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{56}$$

The matrix power $\boldsymbol{A}^k$ can be written as

$$\boldsymbol{A}^k = \boldsymbol{P}\boldsymbol{J}^k\boldsymbol{P}^{-1}, \tag{57}$$

where

$$\boldsymbol{J}^k = \begin{bmatrix} \boldsymbol{J}_1^k & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_2^k & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{J}_l^k \end{bmatrix}, \qquad \boldsymbol{J}_i^k = (\lambda_i \boldsymbol{I}_{r_i} + \boldsymbol{B}_{r_i})^k \tag{58}$$

Let us compute J_i^k for $r_i = 1$ (simple eigenvalue)

$$J_i^k = \lambda_i^k. \tag{59}$$

On the other hand, for $r_i = 2$, (eigenvalue with algebraic multiplicity 2 and geometric multiplicity 2) we have have

$$J_i = \lambda_i \boldsymbol{I}_2 + \boldsymbol{B}_2, \tag{60}$$

$$J_i^2 = (\lambda_i \boldsymbol{I}_2 + \boldsymbol{B}_2)^2 = \lambda_i^2 \boldsymbol{I}_2 + 2\lambda_i \boldsymbol{B}_2, \tag{61}$$

$$\vdots$$

$$J_i^k = (\lambda_i \boldsymbol{I}_2 + \boldsymbol{B}_2)^k = \lambda_i^k \boldsymbol{I}_2 + k\lambda_i^{k-1} \boldsymbol{B}_2, \tag{62}$$

where we used the fact that $\boldsymbol{B}_2^k = 0$ for all $\boldsymbol{k} \geq 2$. Similarly, for $r_i = 3$ it can be shown that $\boldsymbol{B}_3^k = 0$ for all $k \geq 3$, which yields

$$J_i^k = \lambda_i^k \boldsymbol{I}_3 + k\lambda_i^{k-1} \boldsymbol{B}_3 + k\lambda_i^{k-2} \boldsymbol{B}_3^2 \qquad k \geq 3. \tag{63}$$

By taking the norm of (57) we obtain

$$\left\| \boldsymbol{A}^k \right\| \leq K \left\| \boldsymbol{J}^k \right\|, \qquad K = \|\boldsymbol{P}\| \left\| \boldsymbol{P}^{-1} \right\|. \tag{64}$$

At this point we recall that for any matrix norm compatible with a vector norm and for any block-diagonal matrix such as $\boldsymbol{J}$ or $\boldsymbol{J}^k$ we have

$$\left\| \boldsymbol{J}^k \right\| = \max \left\{ \left\| \boldsymbol{J}_1^k \right\|, \ldots, \left\| \boldsymbol{J}_l^k \right\| \right\}. \tag{65}$$

If the eigenvalues of $\boldsymbol{A}$ sitting at the boundary of the unit circle are *simple* then, by equation (59) we have

$$\left\| \boldsymbol{J}_i^k \right\| = 1. \tag{66}$$

On the other hand, if $|\lambda_i| < 1$ (eigenvalue within the unit circle or arbitrary multiplicity) then by equation (62) or (63) we have that

$$\left\| \boldsymbol{J}_i^k \right\| \to 0 \quad \text{for } k \to \infty. \tag{67}$$

Since $\left\| \boldsymbol{J}_i^k \right\|$ if finite for all k, there exists a finite M such that $\left\| \boldsymbol{J}_i^k \right\| \leq M$ for all k.

Finally, if there exists a non-simple eigenvalue λ_i (eigenvalue with algebraic multiplicity larger than one) at the boundary of the unit circle then we can no longer guarantee that $\left\| \boldsymbol{J}_i^k \right\|$ is bounded independently of k. In fact, suppose that the algebraic multiplicity of the eigenvalue λ_i at the boundary of the unit circle (i.e., $|\lambda_i| = 1$) is $r_i = 2$. Then by using (62) we see that

$$\left\| \boldsymbol{J}_i^k \right\|_1 = |\lambda_i|^k + k\,|\lambda_i|^{k-1} = 1 + k \qquad \text{for all } k \geq 2. \tag{68}$$

In summary, if the root condition is satisfied, i.e., if all the eigenvalues of $\boldsymbol{A}$ are within the unit circle with the exception of a finite number of *simple* eigenvalues sitting at the boundary of the unit circle then

$$\left\| \boldsymbol{A}^k \right\| \leq K \left\| \boldsymbol{J}^k \right\| \leq M \qquad \text{for all } k \in \mathbb{N}, \tag{69}$$

where $M > 0$ is independent of k. This completes the proof of Lemma 1.

□

We now have all elements to show that if a scheme satisfies the root condition then it is zero-stable. To this end, let us take the infinity norm of (51), and use (52) (or (69)) to obtain

$$\left| e_{k+q}^m \right| \leq \left\| \boldsymbol{e}_{k+1}^m \right\|_\infty \leq M \left(\left\| \boldsymbol{e}_0^m \right\|_\infty + \Delta t \sum_{p=0}^{k} \left\| \boldsymbol{b}_p^m \right\|_\infty + \Delta t \sum_{p=0}^{k} \left\| \boldsymbol{d}_p^m \right\|_\infty \right). \tag{70}$$

By using definition (48) and (47) we see that

$$\begin{aligned}\sum_{m=1}^{n}\left\|\boldsymbol{b}_p^m\right\|_\infty &= \sum_{m=1}^{n}\left|\Phi_{\boldsymbol{f}}^m(\boldsymbol{z}_{p+q},\ldots,\boldsymbol{z}_p,t_p,\Delta t)-\Phi_{\boldsymbol{f}}^m(\boldsymbol{u}_{p+q},\ldots,\boldsymbol{u}_p,t_p,\Delta t)\right| \\ &= \left\|\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{z}_{p+q},\ldots,\boldsymbol{z}_p,t_p,\Delta t)-\boldsymbol{\Phi}_{\boldsymbol{f}}(\boldsymbol{u}_{p+q},\ldots,\boldsymbol{u}_p,t_p,\Delta t)\right\|_1 \\ &\leq L\sum_{s=0}^{q}\left\|\boldsymbol{z}_{p+s}-\boldsymbol{u}_{p+s}\right\|_1 \\ &= L\sum_{s=0}^{q}\sum_{m=1}^{n}\left|z_{p+s}^m-u_{p+s}^m\right| \\ &= L\sum_{s=0}^{q}\sum_{m=1}^{n}\left|e_{p+s}^m\right|, \end{aligned} \tag{71}$$

where we assumed that $\boldsymbol{\Phi}_{\boldsymbol{f}}$ is Lischitz continuous. Next, define

$$g_{k+q}=\sum_{m=1}^{n}\left|e_{k+q}^m\right|. \tag{72}$$

Note that g_{k+q} is the 1-norm of the vector $\boldsymbol{z}_{k+q}-\boldsymbol{u}_{k+q}$ (see Eq. (47)). Substituting (71) into the inequality (70) (summed-up in m) yields

$$\begin{aligned} g_{k+q} &\leq M\left(\sum_{m=1}^{n}\left\|\boldsymbol{e}_0^m\right\|_\infty + L\Delta t\sum_{s=0}^{k+q}g_s+\Delta t\sum_{p=0}^{k}\sum_{m=1}^{n}\left\|\boldsymbol{d}_p^m\right\|_\infty\right) \\ &\leq Mn\epsilon(1+k\Delta t)+ML\Delta t\sum_{s=0}^{k+q}g_s. \end{aligned} \tag{73}$$

Now we can use the discrete Grönwall lemma (see, e.g., [3, Lemma 11.2]) to conclude that

$$g_{k+q}\leq\epsilon\left(nM(1+(k+1)\Delta t)e^{ML(k+q+1)\Delta t}\right)\leq\epsilon\underbrace{nM(1+T)e^{MLT}}_{C}, \tag{74}$$

where $T\geq(q+k+1)\Delta t$ is some integration period. Recalling that g_{k+p} is the 1-norm of the vector $\boldsymbol{z}_{k+q}-\boldsymbol{u}_{k+q}$ (see Eq. (72)) we see that (74) can be written as

$$\left\|\boldsymbol{z}_{k+q}-\boldsymbol{u}_{k+q}\right\|_1\leq C\epsilon, \tag{75}$$

for all k such that $(q+k+1)\Delta t\leq T$. Alternatively, if we set a maximum number of time steps steps $N\geq k$ and an integration period T then (75) holds for all $k\leq N$ and for all $\Delta t\leq T/(N+q)=\Delta t^*$. This is were the definition of zero-stability kicks in, i.e., conditions (74) and (75) are satisfied for all $\Delta t\leq T/(N+q)=\Delta t^*$. Based on definition (19) we conclude that the root condition implies zero-stability. The converse statement, i.e., zero-stability implies root condition, is straightforward. Indeed, if the scheme is zero stable then (20) is satisfied for all ϵ. This implies that (see Euation 51)

$$\left\|\boldsymbol{A}^{k+1}\boldsymbol{e}_0^m+\Delta t\sum_{p=0}^{k}\boldsymbol{A}^{k-p}\left(\boldsymbol{b}_p^m+\boldsymbol{d}_p^m\right)\right\|_\infty\leq C\epsilon \tag{76}$$

Recalling that C must be independent of k, this condition can be satisfied for all ϵ if and only if $\left\|\boldsymbol{A}^k\right\|\leq M$.

Proof of theorem 4. Let $\boldsymbol{y}_k = \boldsymbol{y}(t_k)$ be the solution of the ODE (1) evaluated at $t = t_k$. A substitution of such solution into the scheme (14) yields the truncation error $\boldsymbol{\tau}_{k+q}$, hereafter written in a componentwise form $(m = 1, ..., n)$

$$\sum_{j=0}^{q} \alpha_j y_{k+j}^m = \Delta t \left(\Phi_{\boldsymbol{f}}^m(\boldsymbol{y}_{k+q}, \ldots, \boldsymbol{y}_k, t_k, \Delta t) + \tau_{q+k}^m \right). \tag{77}$$

Similarly, the numerical solution $\boldsymbol{u}_k$ satisfies

$$\sum_{j=0}^{q} \alpha_j u_{k+j}^m = \Delta t \Phi_{\boldsymbol{f}}^m(\boldsymbol{u}_{k+q}, \ldots, \boldsymbol{u}_k, t_k, \Delta t). \tag{78}$$

Subtracting (78) from (77) yields

$$\sum_{j=0}^{q} \alpha_j e_{k+j}^m = \Delta t \left[\Phi_{\boldsymbol{f}}^m(\boldsymbol{y}_{k+q}, \ldots, \boldsymbol{y}_k, t_k, \Delta t) - \Phi_{\boldsymbol{f}}^m(\boldsymbol{u}_{k+q}, \ldots, \boldsymbol{u}_k, t_k, \Delta t) + \tau_{q+k}^m \right], \tag{79}$$

where

$$e_{k+j}^m = y_{k+j}^m - u_{k+j}^m. \tag{80}$$

By following exactly same steps that took as from equation (46) to (75) in the proof of theorem 2 we obtain the error bound

$$\|\boldsymbol{y}(t_k) - \boldsymbol{u}_k\|_1 \leq MTe^{MLT} \|\boldsymbol{\tau}(\Delta t)\|_1, \tag{81}$$

where the global truncation error $\|\boldsymbol{\tau}\|$ is a function of Δt. To obtain (81) we replaced $(q + k)\Delta t$ with T, which implies that (81) holds for all $\Delta t \leq T/(N + q)$ (this is where zero stability comes in) where N any fixed number larger or equal than $(k + q)$.

Moreover, for Δt small enough, we have seen that $\|\boldsymbol{\tau}\|_1$ goes to zero as some power of Δt (otherwise the method is not consistent). Equation (81) says that the convergence order of the method is the same as the order of consistency.

To obtain the bound (81) we assumed that the initial condition has no error, and that the numerical computation of $\boldsymbol{\Phi_f}$ and all arithmetic operations in the schemes are exact. Clearly this is not the case in practice. It is possible to repeat the proof above, by assuming that all these numerical inaccuracies are bounded, e.g., as a function of the machine precision ϵ, and develop a more detailed bound that depends on ϵ.

References

[1] W. Hahn. *Stability of motion.* Springer, 1967.

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

[3] A. Quarteroni, R. Sacco, and F. Salieri. *Numerical mathematics.* Springer, 2007.