

This last quantity goes to zero as $k \rightarrow 0$ since $e^{|\lambda|T}$, q'' , T are bounded. We have therefore established that

$$(6.25) \quad \lim_{\substack{k \rightarrow 0 \\ kN=T}} E^N = 0$$

and that the forward Euler algorithm is convergent in exact arithmetic.

6.2. The effect of inexact arithmetic and truncation error - stability.

The forward Euler method is convergent in exact arithmetic but this does not necessarily hold true when inexact, computer arithmetic is used. We have shown that the first term in (6.17) goes to zero if $E^0 = 0$

$$(6.26) \quad \lim_{\substack{k \rightarrow 0 \\ kN=T}} (1 + k\lambda)^N E^{(0)} = 0$$

but if $E^0 \neq 0$ then $(1 + k\lambda)^N E^{(0)}$ could lead to an error bounded by $e^{|\lambda|T} E^0$. Clearly, just establishing convergence for exact arithmetic is not sufficient for practical purposes. We must establish conditions such that inherent computer arithmetic errors can be controlled. We say that algorithms that permit us to control error growth are *stable* while those in which errors grow without bound are *unstable*.

These intuitive definitions of stability must be made more precise. We take our cue from the continuous dependence of the solution to an IVP on the initial conditions. Suppose we could exactly compute the solution to the linear ODE $q' = \lambda q + \sigma$ with slightly different initial conditions $\tilde{q}_0 = q_0 + E^0$. Here $E^0 \neq 0$ is intended to represent any initial error due to inexact computer arithmetic. From (??) we know that

$$(6.27) \quad |q(t; q_0) - q(t; \tilde{q}_0)| \leq e^{Lt} |q_0 - \tilde{q}_0| .$$

For the model problem this result can be sharpened since we know the exact solution from (6.3), and we have

$$(6.28) \quad |q(t; q_0) - q(t; \tilde{q}_0)| = e^{\lambda t} E^0 .$$

This means that if an initial error E^0 is unavoidable, and all subsequent computations are done exactly, we can expect an error $E(T) = e^{\lambda T} E^0$ at time T . Clearly we cannot hope to better this using approximate computations, so we should recognize that algorithms that are capable of reproducing this sort of behavior are good candidates for further study. Such an algorithm should reproduce the error growth of the exact solutions in the limit of $k \rightarrow 0$ though it might have different behavior when $k \neq 0$. Note that this discussion in terms of propagation of an initial error is also applicable to later stages of the algorithm. We can envisage that the truncation error of the algorithm introduces a certain error at each step. We would like this error to be kept under control as the algorithm progresses.

DEFINITION 3 (intuitive). *An algorithm \mathcal{A} producing the approximations $\{Q^n\}_{n=0,1,\dots}$ to the exact solution of $q' = f(q)$, $q(0) = q_0$ is said to be **zero-stable** if $|Q^n - q(t^n)|$ is bounded in the limit of $k \rightarrow 0$, $n = 1, 2, \dots, N = T/k$.*

DEFINITION 4 (exact). *Let S be the vector with r components of initial errors in an r -step algorithm that produce the approximations $\{Q^n\}_{n=r,r+1,\dots}$ to the exact solution of the IVP $q' = f(q)$, $q(0) = q_0$*

$$(6.29) \quad S = [Q^0 - q(t^0), Q^1 - q(t^1), \dots, Q^{r-1} - q(t^{r-1})] .$$

The algorithm \mathcal{A} is said to be **zero-stable** if for any $\varepsilon > 0$ there exist two positive constants δ, c such that when $\|S\|_\infty < \varepsilon$, $k < \delta$ we have

$$(6.30) \quad \|Q^n - q(t^n)\|_\infty < c \varepsilon$$

with $n = r, r + 1, \dots, N = T/k$.

Note that in the above definition we require bounded errors at later times when two conditions are met:

- (1) The initial errors are small, i.e. $\|S\|_\infty < \varepsilon$,
- (2) The step size is small $k < \delta$.

By this definition the forward Euler method is zero-stable. It is a one step method requiring just one starting value Q^0 so S is a scalar $S = E^0 = Q^0 - q_0$. Suppose we know that error in approximating the starting value (due to computer imprecise arithmetic) is less than ε , $|E^0| < \varepsilon$. By the argument from the previous section we know that

$$(6.31) \quad |E^n| \leq A |E^0| + B k$$

with $A = e^{|\lambda|t^n}$, $B = e^{|\lambda|t^n} t^n |q''(\xi)|/2$. Both A and B are positive finite quantities for any finite time t^n . We can write

$$(6.32) \quad |E^n| \leq A |E^0| + B k < A\varepsilon + B\delta = \left(A + B\frac{\delta}{\varepsilon} \right) \varepsilon$$

and choosing $\delta = \varepsilon$ and $c = A + B$ we verify that

$$(6.33) \quad |E^n| < c\varepsilon$$

so Euler's method is zero stable for the model problem. This means that even if an initial error is introduced, the algorithm produces an approximation in which the error has not grown faster than we would have expected from solving the exact equation with perturbed initial conditions provided we use a small enough step size $k < \delta$. The above is also a proof of convergence for Euler's method, i.e. for sufficiently small initial error $|E^0| < \varepsilon$ and sufficiently small step sizes $k < \delta$ the error at time step n can be bounded by $|E^n| < c\varepsilon$ and therefore be made as small as we'd like. Thus zero-stability is a necessary and sufficient condition for convergence of Euler's method.

An immediate question that arises is the effect of using a finite step size $k > 0$. In the above proof of zero-stability for Euler's method we used $\delta = \varepsilon$ for any $\varepsilon > 0$. This essentially means that we considered that we can take as small a step size as we'd like in order to establish the bound (6.33). Practical considerations might dictate otherwise though since decreasing the step size increases the number of steps that must be computed. Of course even when using finite step sizes we'd still want to obtain suitable approximations, in which errors have not grown so much that they completely mask the true solution. Let's look again at (6.17) which we repeat here for some intermediate time $t^{(n)}$

$$(6.34) \quad E^{(n)} = (1 + k\lambda)^n E^{(0)} - \sum_{j=0}^{n-1} (1 + k\lambda)^{n-1-j} \omega^{(j)} .$$

When we take a large number of steps n the initial error gets amplified by $(1 + k\lambda)^n$ and that introduced in the previous step j is amplified by $(1 + k\lambda)^{n-1-j}$. An immediate condition that suggests itself is to impose

$$(6.35) \quad |1 + k\lambda| \leq 1$$

so that the error $E^{(n)}$ does not grow as n increases. The quantity $A(z) = 1 + z$, $z = k\lambda$ is known as the *amplification factor* for Euler's method, and (6.35) says that the absolute value of the amplification factor should be less than one to have a stable computation.

DEFINITION 5. *An algorithm \mathcal{A} using step size k whose amplification factor $A(z)$ satisfies $|A(z)| \leq 1$ for the model problem $q' = \lambda q$ is said to be **absolutely stable**. The region in $z = k\lambda$ over which $|A(z)| \leq 1$ is called the region of **absolute stability** of the algorithm.*

In the above definition the “absolute” epithet is suggested by the taking of the absolute value of $A(z)$. The inequality (6.35) defines the region of absolute stability for Euler's method. Within the region of absolute stability Euler's method is convergent in the sense that

$$(6.36) \quad \lim_{n \rightarrow \infty} E^n = 0$$

which means that as we integrate towards ever larger times t^n the error between the approximation and the exact solution $E^n = Q^n - q(t^n)$ goes to zero. Note the rather subtle difference between zero-stability and absolute stability for Euler's method. Zero-stability establishes that the error of Euler's method at any given, finite time T that is attained after $N = T/k$ steps goes to zero as the step size k goes to zero. Absolute stability shows that the asymptotic error E^n at $t^n = nk$ goes to zero as $n \rightarrow \infty$ with k finite.

6.3. Convergence on the model problem for other algorithms. From the study of Euler's algorithm we have been able to establish the concepts of convergence, zero-stability, absolute stability and amplification factor. For these to be useful they must be applicable to other algorithms. We have seen that zero-stability ensures convergence at any fixed time as the step size goes to zero for Euler's method. Similarly, absolute stability ensures asymptotic convergence as $t \rightarrow \infty$ for Euler's method. We would like to know whether this holds for all algorithms or whether additional conditions are required. Also, it would be useful to find a quicker way to establish convergence than the bounding, “ $\delta - \varepsilon$ ” proofs used above. We now turn to these tasks, still concentrating on the model problem $q' = f(q) = \lambda q$, $q(0) = q_0$.

6.3.1. Consistency. Suppose that a LMM is proposed for the model problem which leads to the recurrence relation

$$(6.37) \quad \frac{1}{k} \sum_{j=0}^r a_j Q^{n+j} = \sum_{j=0}^r b_j f(Q^{n+j}) = \lambda \sum_{j=0}^r b_j Q^{n+j} .$$

in which some linear combination of values on the lhs is intended to evaluate an average of the derivative on the rhs. In operator form this can be written as

$$(6.38) \quad \tilde{D}Q^n = DQ^n$$

$$(6.39) \quad \tilde{D} = \frac{1}{k} \sum_{j=0}^r a_j E^j, \quad D = \sum_{j=0}^r b_j f(E^j).$$

Here we interpret \tilde{D} as the truncation of some infinite series that would give as its exact sum the exact operator D . This relation is intended to approximate the initial ODE so we expect that

$$(6.40) \quad \frac{1}{k} \sum_{j=0}^r a_j q(t^{n+j}) \cong \sum_{j=0}^r b_j q'(t^{n+j})$$

By analogy with the definition of truncation error for Euler's method we define the truncation error for the LMM (6.37) as

$$(6.41) \quad \tau^n = (\tilde{D} - D) q(t^n)$$

Taylor series expansion around t^n leads to

$$(6.42) \quad \tau^n = \frac{1}{k} \left\{ \sum_{j=0}^r a_j \right\} q(t^n) + \left\{ \sum_{j=0}^r (j a_j - b_j) \right\} q'(t^n) + \dots$$

$$(6.43) \quad + k^{p-1} \left\{ \sum_{j=0}^r \left[\frac{j^p}{p!} a_j - \frac{j^{p-1}}{(p-1)!} b_j \right] \right\} q^{(p)}(t^n) + \dots$$

It is clear that for the method to converge a necessary condition is that

$$(6.44) \quad \lim_{k \rightarrow 0} \tau^n = 0.$$

DEFINITION 6. *An algorithm is said to be consistent if its truncation error goes to zero as the step size goes to zero.*

Euler's method has a truncation error of

$$(6.45) \quad \tau^n = \frac{k q''(\xi)}{2}$$

and is consistent. A general LMM is consistent if

$$(6.46) \quad \sum_{j=0}^r a_j = 0, \quad \sum_{j=0}^r (j a_j - b_j) = 0$$

which may be concisely stated in terms of the characteristic polynomials $\rho(\zeta)$, $\sigma(\zeta)$ as

$$(6.47) \quad \rho(1) = 0, \quad \rho'(1) - \sigma(1) = 0.$$

The definition of consistency can be applied to other classes of methods also. Taylor's series methods (4.4-4.6) are obviously consistent and consistency conditions are imposed in order to determine the coefficients in Runge-Kutta methods.

6.3.2. *Operational definitions of stability.* The true practical value of the characteristic polynomials is that they allow us to quickly establish whether a LMM is stable. Here are the principal results of the theory.

DEFINITION 7. *The roots ζ_j of a polynomial are said to satisfy the **stability condition** if either $|\zeta_j| < 1$ or when $|\zeta_j| = 1$, ζ_j is a simple root.*

PROPOSITION 1. *The LMM (6.37) is zero-stable if the roots of of the lhs characteristic polynomial $\rho(\zeta)$ satisfy the stability condition.*

PROPOSITION 2. *The LMM (6.37) is absolutely stable if the roots of $\pi(\zeta; z) = \rho(\zeta) - z\sigma(\zeta)$ satisfy the stability condition. The region of absolute stability is the region of values of $z = k\lambda$ for which the roots of $\pi(\zeta; z)$ satisfy the stability condition.*

These results are proved using the properties of finite difference equations to establish “ $\delta - \varepsilon$ ” proofs as was done above for Euler’s method. The proofs shall be omitted here but are given by Henrici. There is an analogy which may be made with standard calculus. The definition of a derivative $f'(x_0)$ is

$$(6.48) \quad \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

but we rarely use that definition in practice to compute the derivative of $\sin(\cos x)$ say. Rather we establish differentiation rules and tables to more quickly evaluate derivatives. Likewise the definitions of stability involve bounding arguments which are inconvenient to redo for every new method. Finding the roots of characteristic polynomials is more straightforward.

6.3.3. Stability of common LMM’s.

Forward Euler.

$$(6.49) \quad Q^{n+1} = Q^n + \lambda k Q^n$$

$$(6.50) \quad \rho(\zeta) = \zeta - 1, \quad \sigma(\zeta) = 1$$

$$(6.51) \quad \pi(\zeta; z) = \zeta - 1 - z$$

The root of $\rho(\zeta)$ is $\zeta_1 = 1$ which satisfies the stability condition since it is a simple root of absolute value 1. The method is zero stable. The root of $\pi(\zeta)$ is

$$(6.52) \quad \zeta = 1 + z$$

so the region of absolute stability is

$$(6.53) \quad |1 + k\lambda| \leq 1$$

Backward Euler.

$$(6.54) \quad Q^{n+1} = Q^n + \lambda k Q^{n+1}$$

$$(6.55) \quad \rho(\zeta) = \zeta - 1, \quad \sigma(\zeta) = \zeta$$

$$(6.56) \quad \pi(\zeta; z) = \zeta - 1 - z\zeta$$

The root of $\rho(\zeta)$ is $\zeta_1 = 1$ so the method is zero stable. The region of absolute stability is

$$(6.57) \quad |\zeta_1| = \frac{1}{|1 - z|} \leq 1 \Rightarrow |1 - k\lambda| \geq 1$$

Trapezoidal method.

$$(6.58) \quad Q^{n+1} = Q^n + \frac{k\lambda}{2} (Q^n + Q^{n+1})$$

$$(6.59) \quad \rho(\zeta) = \zeta - 1, \quad \sigma(\zeta) = \frac{1}{2}(\zeta + 1)$$

The method is zero stable. The region of absolute stability is defined by

$$(6.60) \quad \left| \frac{2+z}{2-z} \right| \leq 1.$$

Leapfrog method.

$$(6.61) \quad Q^{n+2} = Q^n + 2k\lambda Q^{n+1}$$

$$(6.62) \quad \rho(\zeta) = \zeta^2 - 1, \quad \sigma(\zeta) = 2\zeta$$

The roots of $\rho(\zeta)$ are $\zeta_1 = 1$, $\zeta_2 = -1$ so the method is zero-stable. The absolute stability region is defined by

$$(6.63) \quad \left| z \pm \sqrt{z^2 + 1} \right| \leq 1$$

6.3.4. *Boundary locus method.* Regions of absolute stability defined by conditions such as (6.60) or (6.63) are difficult to analyze. A useful tool to simplify the analysis is the *boundary locus method*. We are interested in regions of the values of z for which the roots of $\pi(\zeta; z)$ satisfy the stability condition. In general z can take complex values. As z varies a particular root of π takes different values; the interesting event is when the absolute value of a root is unity. We can therefore consider the problem of determining the curve in the complex z -plane where one of the roots of $\pi(\zeta; z)$ has absolute value unity. If $|\zeta| = 1$ then we can write $\zeta = e^{i\theta}$ and

$$(6.64) \quad \pi(e^{i\theta}; z) = \rho(e^{i\theta}) - z\sigma(e^{i\theta}) = 0$$

from whence

$$(6.65) \quad z(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}.$$

The curve defined by (6.65) is known as the *boundary locus*. In the boundary locus method we trace $z(\theta)$. This delimits the complex z -plane into regions. We then take an arbitrary point z^* within each region and find the roots of $\pi(\zeta; z^*) = 0$. If the roots satisfy the stability condition then that region of the complex z -plane is a region of absolute stability; if not it is a region of absolute instability.

EXAMPLE 4. *For the trapezoidal method we have*

$$(6.66) \quad z(\theta) = 2 \frac{e^{i\theta} - 1}{e^{i\theta} + 1} = 2i \tan \frac{\theta}{2}$$

so the boundary locus is the imaginary axis. We have two regions. To the left of the imaginary axis the root of $\pi(\zeta; z = -1)$ is $\zeta_1 = 1/3$ which satisfies the stability condition. To the right the root of $\pi(\zeta; 1)$ is $\zeta_1 = 3$ which does not satisfy the stability condition. We conclude that $\text{Im } z \leq 0$ is the region of absolute stability for the trapezoidal method.

EXAMPLE 5. *For the leapfrog method we have*

$$(6.67) \quad z(\theta) = \frac{e^{2i\theta} - 1}{e^{i\theta}} = 2i \sin \theta$$

so the boundary locus is the segment from $-i$ to i along the imaginary axis. There are two regions. One is outside of this segment. The roots of $\pi(\zeta; z = 1)$ are $\zeta_{1,2} = 1 \pm \sqrt{2}$ that do not satisfy the stability condition. The other region is the $-i$ to i segment. The roots of $\pi(\zeta; z = 0)$ are $\zeta_{1,2} = \pm 1$ which satisfy the stability condition. The region of absolute stability for the leapfrog method is the segment from $-i$ to i along the imaginary axis.