Numerical Methods
in Engineering and Applied Science

Lecture 9. Stability of numerical methods for initial-value problems.

Numerical stability

When it comes to numerical methods for ordinary differential equations, there are two kinds of stability:

- at constant $t_N = T$, is the value of u_N (approximation to the exact solution $u(T)$) bounded as $h \to 0$?
- at a constant $h > 0$, is the sequence u_n bounded as $n \to \infty$ $(t \to \infty)$?

We begin with the first question and note that the construction of the multi-step methods which ensures consistency does not guarantee the positive answer to this question.

Stability of linear multi-step methods.

Example. A two-step method,

$$u_{n+1} = -4u_n + 5u_{n-1} + h(4f_n + 2f_{n-1}). \tag{1}$$

Using the coefficients of the scheme, $\alpha_0 = -5$, $\alpha_1 = 4$, $\alpha_2 = 1$, $\beta_0 = 2$, $\beta_1 = 4$ and $\beta_2 = 0$, we calculate

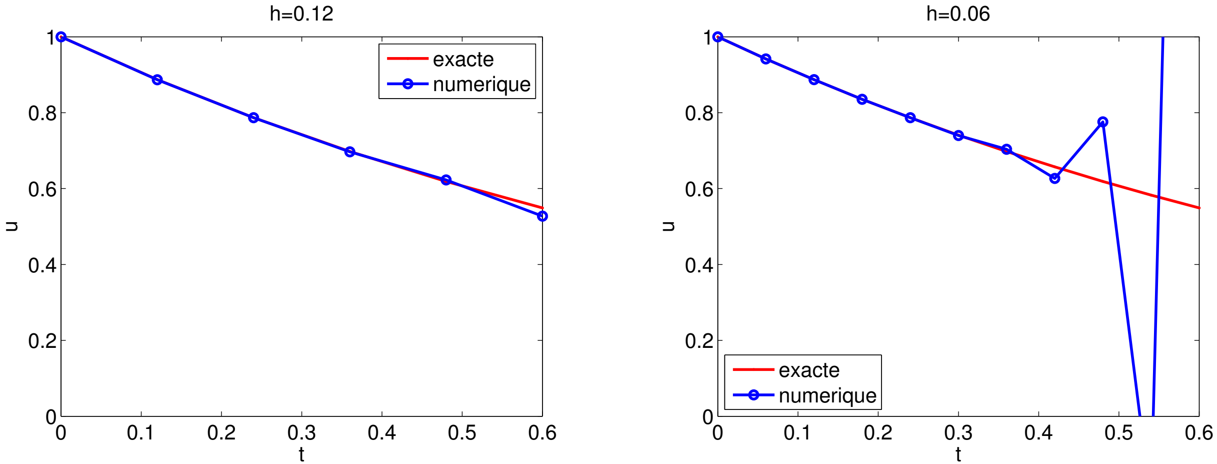
$$\begin{aligned}
C_0 &= \alpha_0 + \alpha_1 + \alpha_2 = -5 + 4 + 1 = 0, \\
C_1 &= (\alpha_1 + 2\alpha_2) - (\beta_0 + \beta_1) = (4 + 2) - (2 + 4) = 0, \\
C_2 &= \tfrac{1}{2}(\alpha_1 + 4\alpha_2) - \beta_1 = \tfrac{1}{2}(4 + 4) - 4 = 0, \\
C_3 &= \tfrac{1}{6}(\alpha_1 + 8\alpha_2) - \tfrac{1}{2}\beta_1 = \tfrac{1}{6}(4 + 8) - \tfrac{1}{2}4 = 0, \\
C_4 &= \tfrac{1}{24}(\alpha_1 + 16\alpha_2) - \tfrac{1}{6}\beta_1 = \tfrac{1}{24}(4 + 16) - \tfrac{1}{6}4 = \tfrac{1}{6}.
\end{aligned} \tag{2}$$

$C_{p+1} = C_4 \neq 0$, the method is consistent of order $p = 3$, with an error constant $C_4 = 1/6$.

We use this method to solve an initial value problem,

$$\begin{cases} u' = -u, \\ u(0) = 1, \end{cases} \tag{3}$$

over $t \in [0, 1]$. The exact solution is $u(t) = \exp(-t)$.



Here, the value of u_{-1} was calculated using the exact solution.

Consider another initial value problem,

$$\begin{cases} u' = 0, \\ u(0) = 1, \end{cases} \tag{4}$$

over $t \in [0, 1]$. The exact solution is $u(t) = 1$. The numerical method reduces to a recurrent sequence,

$$u_{n+1} = -4u_n + 5u_{n-1}. \tag{5}$$

Suppose we begin our calculation with $u_0 = 1$ and $u_{-1} = 1 + \epsilon$, where, for example, $\epsilon = 0.001$. We obtain the following values: $u_1 = 1.005$, $u_2 = 0.98$, $u_3 = 1.105$, $u_4 = 0.48$, $u_5 = 3.605$, $u_6 = -12.02$, $u_7 = 66.105$, $u_8 = -324.52$, ...

The characteristic polynomial associated with (5) is

$$z^2 + 4z - 5 = (z - 1)(z + 5), \tag{6}$$

and we write the general term in the form

$$u_n = A(1)^n + B(-5)^n. \tag{7}$$

The coefficients A and B are determined by u_0 and u_{-1}. In our case, $A = 1 + 5\epsilon/6$ and $B = -5\epsilon/6$. We can clearly see that any small disturbance introduced at an instant t_m grows exponentially with $n \to \infty$. So in limit $h \to 0$ we have $u_N \to \infty$.

We can generalize this analysis to any multi-step method. We define the characteristic polynomial

$$\rho(z) = \sum_{j=0}^{s} \alpha_j z^j, \tag{8}$$

and we associate to the method a linear recurrence relation,

$$\sum_{j=0}^{s} \alpha_j v_{n+1+j-s} = 0. \tag{9}$$

Definition. A multi-step method is **zero-stable** if the solutions $\{v_n\}$ of the linear recurrent sequence (9) are bounded in limit $n \to \infty$.

Theorem. A multi-step method is zero-stable if the roots of $\rho(z)$ satisfy $|z| \leq 1$ and any root such that $|z| = 1$ is simple.

Proof. Let us investigate $\{v_n\}$ - the solutions of (9). Each solution is determined by its initial conditions, v_{1-s},...,v_0. Since (9) is linear, we find s independent solutions which form a basis of the solution space. All the solutions of the linear recurrence relation (9) are bounded in limit $n \to \infty$ iff each solution of the basis is bounded.
Let z be a solution of $\rho(z) = 0$. We see that $v_n = z^n$ is a solution of (9). These functions constitute a basis if $\rho(z)$ has only simple roots. They are bounded iff $|z|$ $\leq$1. In the case where $\rho(z)$ has a roots of multiplicity $m \geq 2$, the basis is z^n, nz^n, n^2z^n,..., $n^{m-1}z^n$. These functions are bounded iff $|z| < 1$. $\square$

Remark. One can take the conditions of this theorem as the definition of zero-stability.

Now we use this theorem to study the zero-stability of different classical schemes.

Leapfrog

$$u_{n+1} = u_{n-1} + 2hf_n \tag{10}$$

The characteristic polynomial is $\rho(z) = z^2 - 1$, its roots are simple, $\{1, -1\}$, the method is stable.

Adams–Bashforth and Adams–Moulton

$$u_{n+1} = u_n + h\sum_{j=0}^{s} \beta_j f_{n+1+j-s}. \tag{11}$$

In both cases, $\rho(z) = z^s - z^{s-1}$, the roots are $\{1, 0, ..., 0\}$ and the methods are stable.

BDF3

$$u_{n+1} = \frac{18}{11}u_n - \frac{9}{11}u_{n-1} + \frac{2}{11}u_{n-2} + \frac{6}{11}hf_{n+1} \tag{12}$$

The characteristic polynomial is $\rho(z) = 11z^3 - 18z^2 + 9z - 2$, its roots are $\{1, \frac{7-i\sqrt{39}}{22}, \frac{7+i\sqrt{39}}{22}\}$. Since $|\frac{7\pm i\sqrt{39}}{22}| = \frac{2}{11} < 1$, the method is stable.
It can be shown that an s-step backward differentiation method is stable if $1 \leq s \leq 6$ and unstable if $s \geq 7$.

Remark. The coefficients α of a consistent scheme are such that $\sum_{j=0}^{s} \alpha_j = 0$. Consequently, $z = 1$ is a root of $\rho(z)$. In addition, the second characteristic polynomial can be defined as

$$\sigma(z) = \sum_{j=0}^{s} \beta_j z^j \tag{13}$$

and it is possible to show that $\rho'(1) = \sigma(1)$ is the second necessary condition for consistency.

Theorem. First Dahlquist barrier.
Order of consistency p of a stable s-step method is equal to

$$p \leq \begin{cases} s+2 & \text{if } s \text{ is even} \\ s+1 & \text{if } s \text{ is odd} \\ s & \text{if } \beta_s/\alpha_s \leq 0 \text{ (e.g. the method is explicit)} \end{cases} \tag{14}$$

We see that the Adams–Bashforth methods are optimal explicit methods (s steps, order $p = s$). Adams–Moulton are optimal if s is odd (order $p = s + 1$). An example of an optimal implicit method with an even number of steps s is the Simpson formula,

$$u_{n+1} = u_n + \frac{1}{3}h\left(f_{n-1} + 4f_n + f_{n+1}\right), \tag{15}$$

which is of order $p = s + 2$, where $s = 2$.

Convergence of multi-step methods.
We analyzed the consistency and the stability of the methods. It remains for us to show that schemes which possess these properties provide a good approximate solution to the initial-valued problem.
Consistency ensures that the error (due to truncation) is small when introduced. This error can grow as the calculation progresses. Growth can be exponential. The global error at $t = t_N$ is not simply the sum of local errors at previous steps but the superposition of what has become of those errors at t_N. The stability condition ensures that there is no error amplification. We will see that consistency and stability are necessary and sufficient for convergence.
We recall that the initial value problem has a unique solution $u(t)$ if the function $f(t, u)$ is continuous with respect to t and Lipschitz continuous with respect to u (see Cauchy theorem).

Let us consider initial values u_0, u_1, ..., u_{s-1}. We apply a multi-step method to calculate a sequence $\{u_n\}$. For all t and h given such that $t/h = n$ is an integer, we introduce the following notation:

$$u_h(t) = u_n \quad \text{if} \;\; t = nh. \tag{16}$$

Definition. A multi-step method is convergent, if for any initial value problem that satisfies the conditions of Cauchy theorem on the interval $[0, T]$, and for all initial values u_0, ..., u_{s-1} which satisfy the condition

$$u_h(jh) \to u(jh) \quad \text{if} \quad h \to 0, \quad j = 0, 1, ..., s-1, \tag{17}$$

the solution u_h satisfies

$$u_h \to u(t) \quad \text{if} \quad h \to 0 \tag{18}$$

for all $t \in [0, T]$, where $u(t)$ is the exact solution of the problem.

Definition. A multi-step method is convergent of order p if for any initial-valued problem with f sufficiently differentiable, there exist $h_0 > 0$ such that

$$||u_h(t) - u(t)|| \leq Ch^p \quad \text{where} \quad h \leq h_0 \tag{19}$$

if the initial values ave such that

$$||u_h(jh) - u(jh)|| \leq C_0 h^p \quad \text{where} \quad h \leq h_0, \quad j = 0, 1, ..., s-1. \tag{20}$$

With these definitions, we find the following principle:

$$convergence = stability + consistancy$$

In the context of multi-step methods, this principle becomes the Dahlquist theorem. In the context of PDEs, this is the Lax–Richtmyer theorem.

Theorem. (Dahlquist, 1956)

- If a multi-step method is convergent, it is necessarily stable and consistent (i.e. of order 1 at least).
- If a multi-step method is stable and consistent (of order p), it is convergent (of order p).

Sketch of the proof

1. *convergence* $\Rightarrow$ *stability*. In fact, we will prove $\neg$*stability* $\Rightarrow$ $\neg$*convergence*. Let us consider a special case: $u' = 0$, $u(0) = 0$. The exact solution is $u(t) = 0$. Suppose that the numerical method is unstable. Then it admits a solution $U_n = z^n$ with $|z|$ or $U_n = nz^n$ with $|z| = 1$. In both cases, let us use the initial values $u_h(nh) = \sqrt{k}U_n$, $0 \leq n \leq s-1$. The numerical solution will be $u_h = \sqrt{k}U_n$, $n \geq s$. In the limit of $h \to 0$, the first s values of u_h converge to zero, but $|u_h(nh)| \to \infty$ if $h \to 0$ and $nh = t = const$. The method is not convergent.

2. *convergence* $\Rightarrow$ *consistancy*. For the consistancy, we must ensure

$$\begin{aligned} &\alpha_0 + ... + \alpha_s = 0, \\ &(\alpha_1 + 2\alpha_2 + ... + s\alpha_s) - (\beta_0 + ... + \beta_s) = 0. \end{aligned} \tag{21}$$

Let us consider the problem $u' = 0$, $u(0) = 1$. Its solution is $u(t) = 1$. The initial values $u_h(nh) = 1$, $0 \leq n \leq s-1$ will be taken from the exact solution. The multi-step method is written in the form

$$\alpha_s u_h(t + sh) + \alpha_{s-1} u_h(t + (s-1)h) + ... + \alpha_0 u_h(t) = 0.$$

In the limit of $h \to 0$ we obtain the first equation (21). To obtain the second one, we consider the problem $u' = 1$, $u(0) = 0$. Its exact solution is $u(t) = t$. Note that the numerical solution is given by $u_n = nhK$ ($u_h = tK$), where $K = (\alpha_1 + 2\alpha_2 + ... + s\alpha_s)/(\beta_0 + ... + \beta_s)$. $K = 1$ is necessary for convergence.

3. *stability + consistancy* $\Rightarrow$ *convergence*. We rewrite the multi-step method as a multi-stage method. Let $\psi(t_i, u_i, ..., u_{i+s-1}, h)$ be a function defined implicitly by

$$\psi = \sum_{j=0}^{s-1} \beta'_j f(t_i + jh, u_{i+j}) + \beta'_s f(t_i + sh, h\psi - \sum_{j=0}^{s-1} \alpha'_j u_{i+j}), \tag{22}$$

where $\alpha'_j = \alpha_j/\alpha_k$, $\beta'_j = \beta_j/\beta_k$. The multi-step method is written in a matrix form

$$U_{i+1} = AU_i + h\Phi(t_i, U_i, h), \tag{23}$$

where $U_i = (u_{i+s-1}, ..., u_i)^T$, $i \geq 0$ and $\Phi(t_i, U_i, h) = (\psi(t_i, U_i, h), 0, ..., 0)^T$,

$$A = \begin{pmatrix} -\alpha'_{s-1} & -\alpha'_{s-2} & ... & -\alpha'_1 & -\alpha'_0 \\ 1 & 0 & ... & 0 & 0 \\ ... & & & & \\ 0 & 0 & ... & 1 & 0 \end{pmatrix}. \tag{24}$$

Since $f(t,u)$ is Lipschitz continuous with respect to u, $\Phi(t,U,h)$ is Lipschitz continuous with respect to U (with a constant Λ). By using the stability we can show that there is a norm such that $||A|| \leq 1$. We then obtain

$$||AU + h\Phi(t,U,h) - AV - h\Phi(t,V,h)|| \leq (1 + h\Lambda)||U - V||. \tag{25}$$

Using the consistency, we get

$$||\hat{U}_{i+1} - U(t_{i+1})|| \leq Mh^{p+1}, \tag{26}$$

where $M = const$ and

$$\hat{U}_{i+1} = AU(t_i) + h\Phi(t_i, U(t_i), h). \tag{27}$$

The end of the proof is similar to that for the Runge–Kutta methods.

Let us discuss more about the *stability* of time-stepping methods. We will consider only the methods that converge to the exact solution of the initial value problem in the limit $h \to 0$. In practice, we evaluate the numerical solution with a finite time step h. In addition, we want to maximize h for a given error $||u_N - u(t_N)||$. Therefore, it is important to study the behavior of the scheme for finite h and large t.

Example. Let us use the explicit Euler method to solve the following problem:

$$u'(t) = -8u(t) + 40(3\exp(-t/8) + 1), \quad u(0) = 100. \tag{28}$$

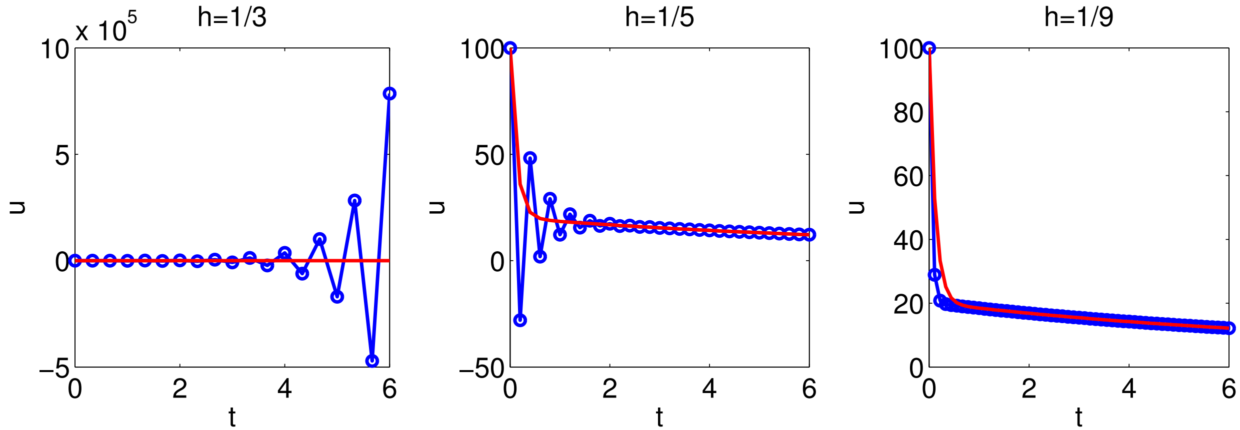
The exact solution is

$$u(t) = \frac{1675}{21}\exp(-8t) + \frac{320}{21}\exp(-t/8) + 5. \tag{29}$$

The explicit Euler method applied to (28) gives

$$u_{n+1} = (1 - 8h)u_n + h\left(120\exp(-t_n/8) + 40\right), \quad n = 0, 1, 2, \ldots \tag{30}$$

with $t_n = nh$, $x_0 = 100$.

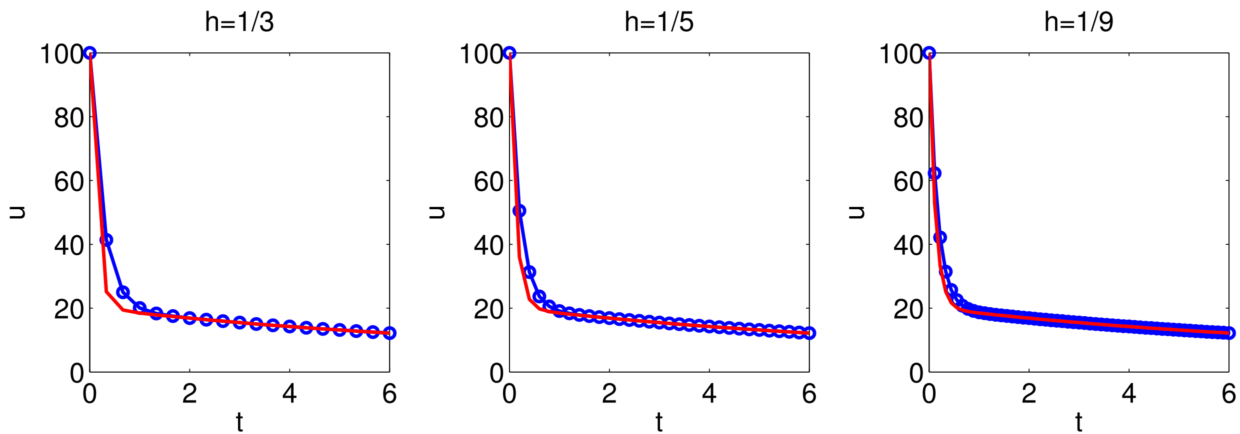


Let us now apply the implicit (backward) Euler method to the same problem:

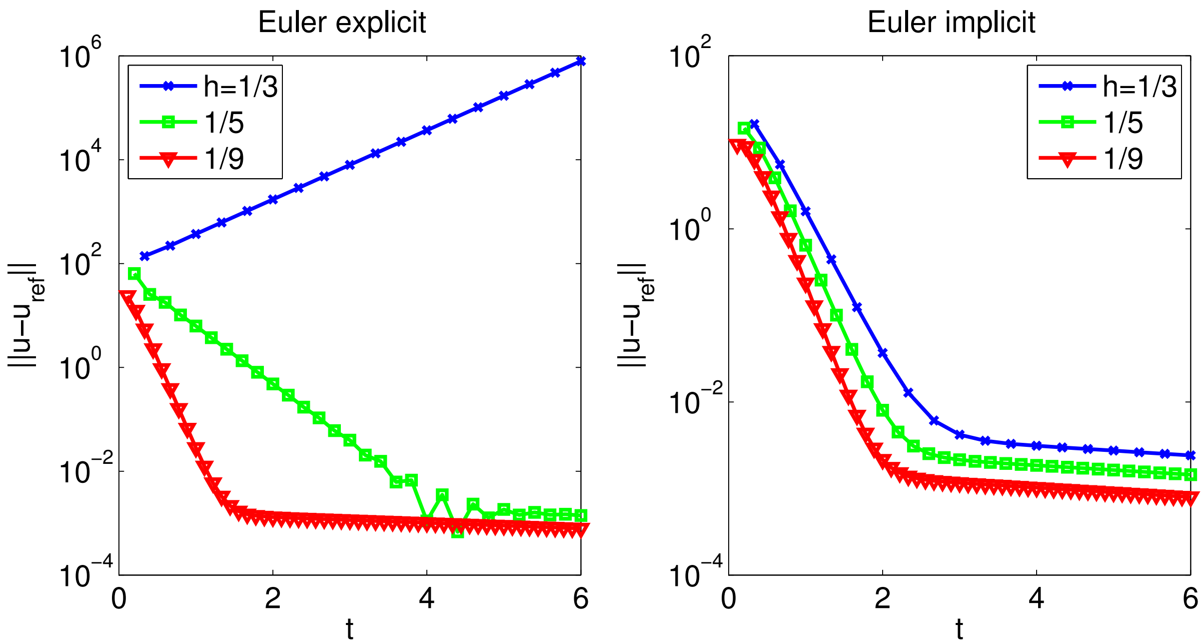
$$u_{n+1} = u_n - 8hu_{n+1} + h\left(120\exp(-t_{n+1}/8) + 40\right), \quad n = 0, 1, \ldots \tag{31}$$

with $t_{n+1} = (n+1)h$, $x_0 = 100$. Solving with respect to u_{n+1}, we find

$$u_{n+1} = \frac{1}{1+8h}\left[u_n + h\left(120\exp(-t_{n+1}/8) + 40\right)\right]. \tag{32}$$



Global error: comparison between the explicit and the implicit Euler methods.



To study the zero-stability, test problem was

$$u'(t) = 0, \quad u(0) = u_0. \tag{33}$$

To study the **absolute stability**, we consider a model problem

$$\begin{aligned} &u'(t) = \lambda u(t), \quad \lambda \in \mathbb{C}, \quad t > 0 \\ &u(0) = 1 \end{aligned} \tag{34}$$

The exact solution of this model problem is $u(t) = \exp(\lambda t)$.
Note that $\lim_{t\to\infty} |u(t)| = 0$ if $\Re(\lambda) < 0$.
A numerical method for (34) is *absolutely stable* with λh if

$$|u_n| \to 0 \quad \text{as} \quad n \to \infty. \tag{35}$$

The multi-step method for this equation becomes

$$\sum_{j=0}^{s} \alpha_j u_{n+1+j-s} = \hat{h} \sum_{j=0}^{s} \beta_j u_{n+1+j-s}, \tag{36}$$

where $\hat{h} = \lambda h$.

Definition. A numerical method for the initial value problem (34) is *absolutely stable* with $\hat{h} = \lambda h$ if

$$|u_n| \to 0 \quad \text{as} \quad n \to \infty. \tag{37}$$

For multi-step methods, we define the stability polynomial,

$$\pi_{\hat{h}}(z) = \rho(z) - \hat{h}\sigma(z) = \sum_{j=0}^{s} \left(\alpha_j - \hat{h}\beta_j\right) z^j. \tag{38}$$

Theorem (root condition). A multi-step method is absolutely stable with a value of $\hat{h} = \lambda h$ iff the roots of $\pi_{\hat{h}}(z)$ satisfy $|z| \leq 1$ and all roots such that $|z| = 1$, are simple.

The **region of absolute stability** of a time-stepping scheme is a set of complex numbers $\mathcal{A} = \{\lambda h$ for which the numerical solution decays to zero$\}$.

Examples of the regions of absolute stability.

- Forward Euler method applied to the model problem (34) yields

$$u_{n+1} = u_n + \lambda h u_n.$$

The stability polynomial is $\pi_{\hat{h}}(z) = z - 1 - \hat{h}$. Its only root is $z = 1 + \hat{h}$ and we look for $\hat{h}$ such that $|1 + \hat{h}| \leq 1$. The region of absolute stability is the interior of a unit disk centered at $(-1, 0)$. For example, if we solve $u' = -8u$ using the Euler method, we have $\hat{h} = -8h$ and $h < 1/4$ for stability.

- Backward Euler method.

$$u_{n+1} = u_n + \lambda h u_{n+1}.$$

The stability polynomial is: $\pi_{\hat{h}}(z) = z - 1 - z\hat{h} = z(1 - \hat{h}) - 1$. The absolute stability region consists of points $\hat{h}$ such that $|1/(1 - \hat{h})| \leq 1$, i.e., in the exterior of a unit disk with center at $(1, 0)$.

- Trapezoidal rule.

$$u_{n+1} = u_n + \frac{\lambda h}{2}(u_{n+1} + u_n). \tag{39}$$

Stability polynomial: $\pi_{\hat{h}}(z) = (z-1) - \frac{1}{2}\hat{h}(z+1) = (1 - \frac{1}{2}\hat{h})z - (1 + \frac{1}{2}\hat{h})$. The absolute stability region consists of points $\hat{h}$ such that $|2+\hat{h}| \le |2-\hat{h}|$. This is a half-plane $\Re\hat{h} \le 0$.

- Leapfrog method.

$$u_{n+1} = u_{n-1} + 2\lambda h u_n \tag{40}$$

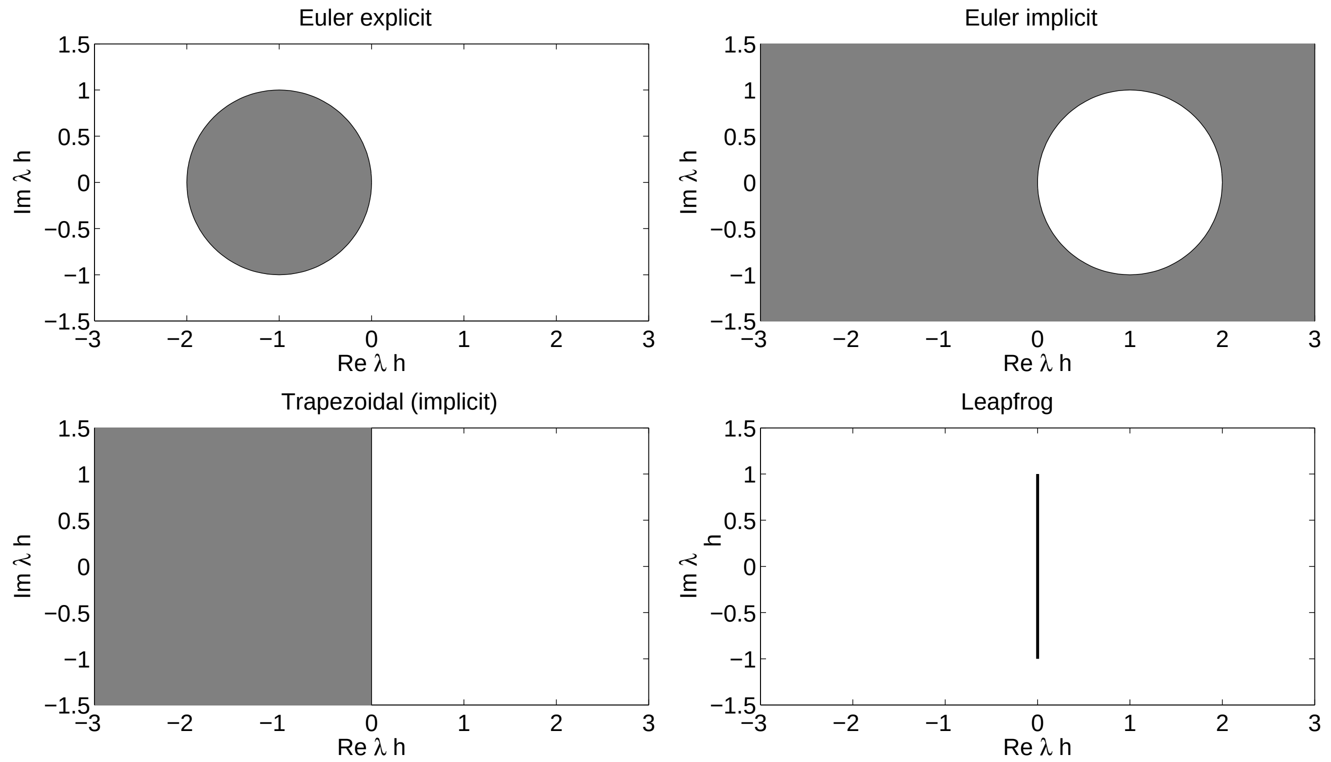
Stability polynomial: $\pi_{\hat{h}}(z) = z^2 - 2\hat{h}z - 1$. We have $z_1 z_2 = -1$. The conditions $|z_1| \le 1$ and $|z_2| \le 1$ yield $|z_1| = |z_2| = 1$. In addition, $z_1 + z_2 = 2\hat{h}$. Therefore, $z_1 = (z_1)^{-1} + 2\hat{h}$. Since $\Re z = \Re z^{-1}$ and $\Im z = -\Im z^{-1}$, we obtain $\Re\hat{h} = 0$ et $\Im\hat{h} = \Im z_1 \le 1$.

Definition. A numerical method is A-*stable* if its absolute stability region $\mathcal{A}$ contains the half-plane $\Re(\hat{h}) < 0$.

Theorem (second Dahlquist barrier).

1. There is no explicit A-stable multistep method.
2. The order p of the A-stable implicit methods satisfies $p \leq 2$.

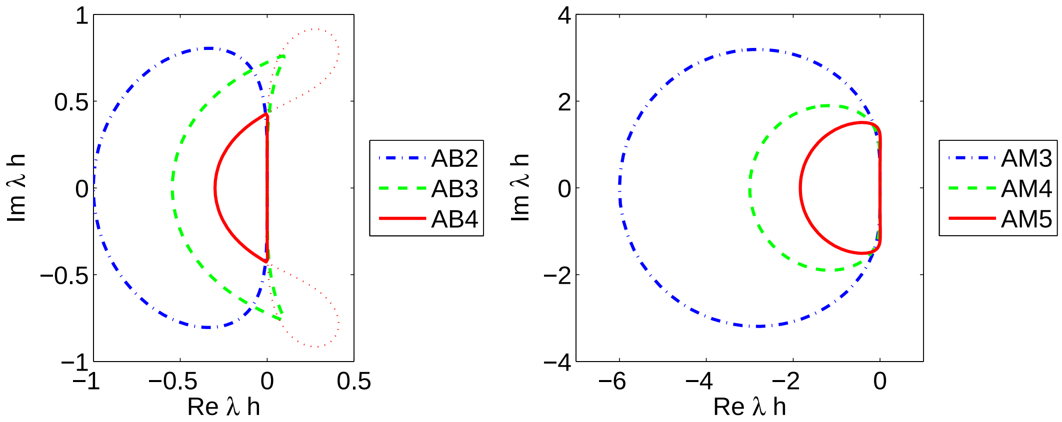
Definition. A numerical method is A_0-*stable* if its absolute stability region $\mathcal{A}$ contains the negative real semi-axis, $\Re(\hat{h}) < 0$, $\Im(\hat{h}) = 0$.


Euler explicit
Euler implicit
Trapezoidal (implicit)
Leapfrog
Im λ h
Re λ h
1.5
1
0.5
0
−0.5
−1
−1.5
−3
−2
−1
0
1
2
3

To find the boundary of a stability region one can use the following method. Solve $\pi_{\hat{h}}(z) = 0$ with respect to $\hat{h}$: $\hat{h} = \rho(z)/\sigma(z)$. Then, on parameterize $z = \exp(\mathrm{i}\theta)$ and plot $\Im\hat{h}(\theta)$ vs. $\Re\hat{h}(\theta)$. For the Adams schemes,

$$\hat{h}(\theta) = \frac{\exp(\mathrm{i}s\theta)\,(1 - \exp(-\mathrm{i}\theta))}{\sum_{j=0}^{s} \beta_j \exp(\mathrm{i}j\theta)}. \tag{41}$$

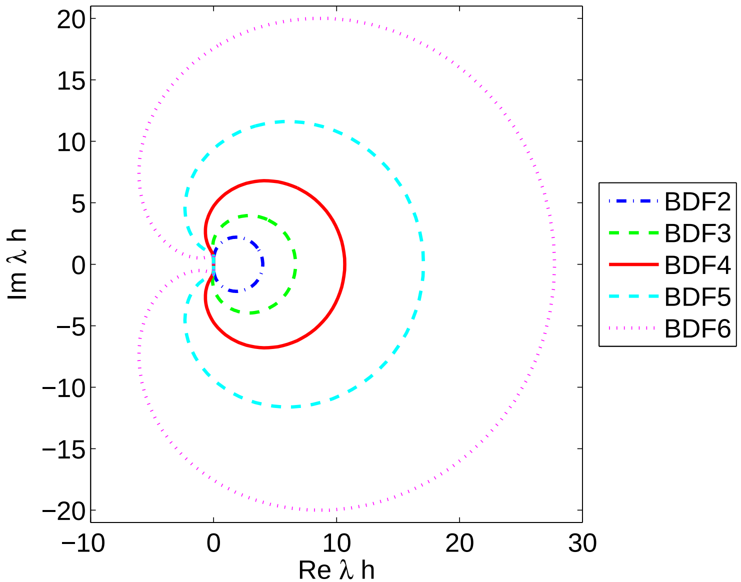
The regions of absolute stability are points in the interior of closed contours.



For backward differentiation formulas (BDF) we find

$$\hat{h}(\theta) = \frac{\sum_{j=0}^{s} \alpha_j \exp(\mathrm{i}j\theta)}{\beta_s \exp(\mathrm{i}s\theta)}. \tag{42}$$

The regions of absolute stability are all points outside of closed contours.



Absolute stability of Runge–Kutta methods

The notion of absolute stability also applies to the Runge–Kutta method.
Let us consider second-order methods.

$$\begin{cases} k_1 = f(t_n, u_n) \\ k_2 = f(t_n + h/(2\theta), u_n + k_1 h/(2\theta)) \\ u_{n+1} = u_n + h((1-\theta)k_1 + \theta k_2) \end{cases}$$

$$\begin{array}{l|ll} 0 & 0 & 0 \\ 1/(2\theta) & 1/(2\theta) & 0 \\ \hline & 1-\theta & \theta \end{array}$$

For our model problem with $f = \lambda u$ we obtain

$$\begin{aligned} k_1 &= \lambda u_n, \\ k_2 &= \lambda(1 + h\lambda/(2\theta))u_n, \\ u_{n+1} &= u_n + (1-\theta)hk_1 + \theta h k_2 = (1 + \hat{h}(1 + \hat{h}/2))u_n = M(\hat{h})u_n, \end{aligned}$$

where $M(\hat{h}) = 1 + \hat{h} + \hat{h}^2/2$ is a *stability function*. Note that it does not depend on θ.

More generally, an explicit Runge–Kutta method with s stages and of order s has the stability function

$$M(\hat{h}) = 1 + \hat{h} + \frac{1}{2!}\hat{h}^2 + ... + \frac{1}{s!}\hat{h}^s. \tag{43}$$

The region of absolute stability is given by

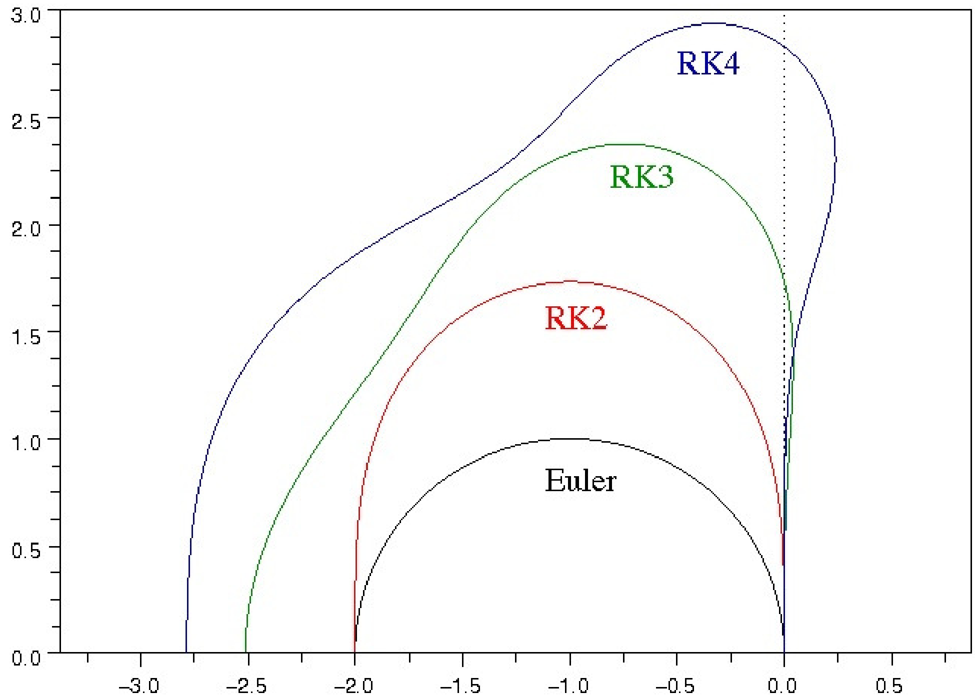
$$|M(\hat{h})| \leq 1. \tag{44}$$

Note that

$$|M(\hat{h})| = \exp(\hat{h}) + \mathcal{O}(\hat{h}^{s+1}) \tag{45}$$

therefore, $|M(\hat{h})| \to \infty$ at $\hat{h} \to -\infty$. There is no explicit A-stable or A_0-stable RK method.

For Runge–Kutta (RK) schemes, the regions of absolute stability are points in the interior of closed contours.



Stiff problems have the following symptoms:

- The problem has two or more significantly different time scales;
- The stability constraint on the time step is more severe that the precision constraint;
- Implicit methods solve the problem more effectively than explicit methods.

In stiff problems, all eigenvalues have negative real part, but the ratio

$$\frac{\max_j(-\Re(\lambda_j))}{\min_j(-\Re(\lambda_j))} \tag{46}$$

can be very large.