Ordinary Differential Equations

The general problem

The Cauchy problem or initial value problem (IVP) is given by

$$\begin{aligned} x'(t) &= f(t, x(t)) \quad \forall (t, x) \in D \subset \mathbb{R}^2, \\ x(t_0) &= x_0. \end{aligned} \tag{1}$$

Phase Space [Arnold 1978]

The set of all possible states of a process is called its phase space.

For example, phase space of the above IVP is simply D.

- First of all, we need to know the existence and uniqueness of the solution of the above IVP.

Existence and Uniqueness Theorem (Local)

Let the function $f \in \mathcal{C}(D)$ ($D \subset \mathbb{R}^2$ is open) and f is local Lipschitz continuous in D with respect to x and uniformly in t:

$$|f(t, x_1 - f(t, x_2)| \leq L|x_1 - x_2|, \quad (t, x_1), (t, x_2) \in D,$$

where L is called the Lipschitz constant. Then, for every $(t_0, x_0) \in D$, the ordinary differential equation (1) has a unique solution $x(t) \in \mathcal{C}^1(I_\delta)$, $I_\delta = [t_0 - \delta, t_0 + \delta]$ on rectangle

$$R = \left\{(t, x) : |t - t_0| \leq a, |x - x_0| \leq b\right\},$$

where

$$\delta = \min\left\{a, \frac{b}{M}\right\}, \quad M = \sup\left\{|f(t, x)| : (t, x) \in R\right\}.$$

Proof ...

If f is continuous, but does not satisfy the Lipschitz condition

- The solution exists, but may not unique. See the Peano Existence Theorem [Walter 1998]. For example,

$$x'(t) = 2\sqrt{|x|}, \qquad x(0) = 0.$$

Exercise: Verify for any constant $c > 0$

$$x(t) = \begin{cases} (t-c)^2, & t \geq c \\ 0, & t < c \end{cases}$$

is a solution for the above IVP.

Loosely we characterize these cases as follow:

- The solution exists for all t.
- The solution blows up after finite time. For example,

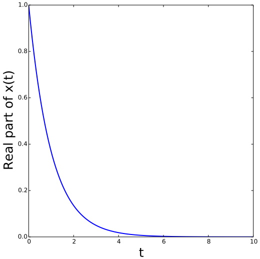
$$x'(t) = x^2, \qquad x(0) = 1.$$

  Analytically, for $-\infty < t < 1$ the IVP has the solution $x(t) = \frac{1}{1-t}$, which "blows up" when $t \to 1^-$.
- The solution collapses for some t. For example,

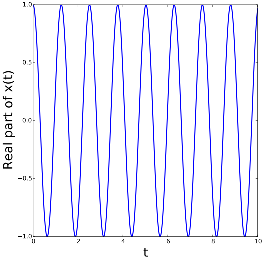
$$x'(t) = -x^{-1/2}, \qquad x(0) = 1.$$

  Analytically, for $-\infty < t < \frac{2}{3}$ the IVP has the solution $x(t) = (1 - \frac{3t}{2})^{\frac{2}{3}}$, which collapses at the singularity of $t = \frac{2}{3}$.

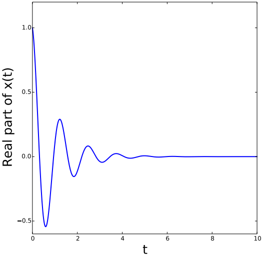
Example: Dahlquist's test equation



(a) $\lambda = -1.0$



(b) $\lambda = 5.0i$



(c) $\lambda = -1.0 + 5.0i$

- Example 1:

$$x'(t) = \lambda x(t) \tag{2}$$

- Here, can guess the general form of solution:

$$x(t) = A\exp(\lambda t) \tag{3}$$

- For *uniqueness*, need to specify an *initial value*, e.g. $x(0) = 1.0$.

- However, very often, **no** analytical solution available, hence need numerical approximations.
- How can we solve it **numerically**?
- Consider the integration form of the IVP

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds.$$

  - x_0 is known, but $\int_{t_0}^{t} f(s, x(s))ds$?

- For $t \in [t_0, t_0 + h]$, we can make the approximation

$$\int_{t_0}^{t} f(s, x(s))ds \approx (t - t_0)f(t_0, x(t_0)),$$

when h is sufficiently small. Hence,

$$x(t) = x_0 + \int_{t_0}^{t} \approx x_0 + (t - t_0)f(t_0, x_0). \tag{4}$$

- Give a sequence

$$t_0 = 0,\ t_1 = t_0 + h,\ t_2 = t_0 + 2h, \ldots, t_n = t_0 + n * h,$$

where $h > 0$ is called the time step, and we denote x_n be the numerical approximation of the exact solution $x(t_n)$.

- Motived by (4), we have

$$x_1 = x_0 + hf(t_0, x_0).$$

This procedure can be continued to produce approximations at t_2, t_3 and so on. In general, we obtain the recursive scheme

$$x_{n+1} = x_n + hf(t_n, x_n), \quad n = 0, 1, \ldots,$$

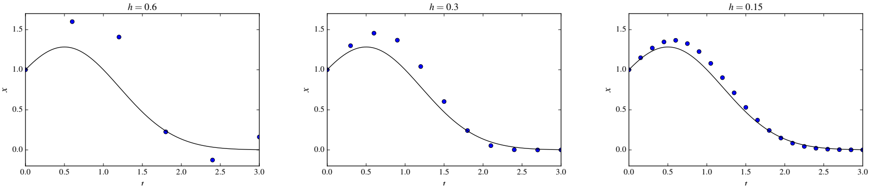
which is the celebrated **Euler method** (or forward Euler method).

- Graphic interpretation of Euler method.

- **Example**

$$x'(t) = (1 - 2t)x(t), \quad t \geq 0,$$
$$x(0) = 1$$

The exact solution is $x(t) = e^{t-t^2}$.



Note: there are errors between the numerical and the exact solutions!

- A realistic goal of numerical solution is not, however, to avoid errors.
- After all, we approximate since we do not know the exact solution in the first place.
- An error exists in every numerical method for ODEs.
- Our purpose is to understand the error and to ensure the error is under control (beyond a specified tolerance).

Global error (GE)

$$e_n := x(t_n) - x_n$$

- How does $e := \max_{n=0,\dots,N} |e_n|$ depend on h?
- It is easy to guess e_n is proportional to h

$$e_n \propto h.$$

- But we need to know more ...

Landau notation

Definition

A function $f(h)$ is said to be in $\mathcal{O}(h^p)$ as $h \to 0$ ("of order p") if there exists a $h_0 > 0$ and a $C > 0$ such that

$$|f(h)| \leq Ch^p \quad \text{for all} \quad 0 < h < h_0. \tag{5}$$

Examples:

- A function $f(h) = ah$ is in $\mathcal{O}(h)$, a function $f(h) = ah + b$ in $\mathcal{O}(1)$ as $h \to 0$.
- A polynomial $f(h) = \sum_{j=0}^{N} a_j h^j$ is of $\mathcal{O}(h^p)$ with p the largest index with $a_i = 0$ for $i < p$; e.g. $f(h) = a_2 h^2 + a_3 h^3 + \ldots$ is of order $\mathcal{O}(h^2)$ naturally.
- If $f(h)$ is in $\mathcal{O}(h^{p_1})$, g in $\mathcal{O}(h^{p_2})$ for $h \to 0$ and $p_1 < p_2$ then g decays faster than f to zero in the sense that there exists a h_0 such that

$$|g(h)| < |f(h)| \quad \text{for all} \quad 0 < h < h_0. \tag{6}$$

Convergence

Definition

A method is **convergent** at a point t_n if

$$|e_n| \to 0 \text{ as } h \to 0. \tag{7}$$

It is **convergent with order** p if

$$|e_n| = \mathcal{O}(h^p) \tag{8}$$

for some $p > 0$.

- The largest possible p is referred to as the **order** of a method
- A convergent method will eventually provide a good approximation of the analytical solution if h is made small enough
- The higher p, the quicker the error decays with h

Forward Euler - convergence

Theorem

When applied to an initial value problem

$$x'(t) = f(t, x(t)) = \lambda x(t) + g(t),\ 0 < t \leq T \tag{9a}$$
$$x(0) = 1 \tag{9b}$$

with $\lambda \in \mathbb{C}$ and g a continuously differentiable function, the forward Euler method converges and the GE at any $t \in [0, T]$ is $\mathcal{O}(h)$.

Gronwall's Lemma

Prove the Gronwall's Lemma: Let $A > 0,\ B \geq 0$. If

$$|e_{j+1}| \leq (1 + A)|e_j| + B,$$

then

$$|e_j| \leq |e_0|e^{jA} + \frac{B}{A}(e^{jA} - 1).$$

Hint: $e^x \geq 1 + x,\ x \geq 0.$

Proof 1 / 3

- Forward Euler for this particular IVP reads

$$x_{n+1} = x_n + h\lambda x_n + hg(t_n) = (1 + h\lambda)\, x_n + hg(t_n) \tag{10}$$

- As before, Taylor expansion of the exact solution gives us

$$x(t_{n+1}) = x(t_n) + h \underbrace{\left[\lambda x(t_n) + g(t_n)\right]}_{=f(t_n, x(t_n))} + R_1(t_n) \tag{11}$$

- Subtracting (10) from (11) gives

$$x(t_{n+1}) - x_{n+1} = e_{n+1} = \underbrace{x(t_n) - x_n}_{=e_n} + h\lambda\left(x(t_n) - x_n\right) + R_1(t_n) \tag{12a}$$

$$= (1 + h\lambda)\, e_n + \underbrace{R_1(t_n)}_{=:T_{n+1}} \tag{12b}$$

  Because $x_0 = x(0)$, it is $e_0 = 0$.

- Now, we have a recursion formula for the global error e_n instead of x_n

Proof 2 / 3

- Unrolling the recursion formula gives

$$
\begin{aligned}
e_1 &= T_1 && \text{(13a)}\\
e_2 &= (1+\lambda h)\, e_1 + T_2 = (1+\lambda h)\, T_1 + T_2 && \text{(13b)}\\
e_3 &= (1+\lambda h)\, e_2 + T_3 = (1+\lambda h)^2\, T_1 + (1+h\lambda)\, T_2 + T_3 && \text{(13c)}\\
&\vdots && \text{(13d)}
\end{aligned}
$$

- In closed form

$$e_n = \sum_{j=1}^{n} (1+h\lambda)^{n-j}\, T_j \tag{14}$$

(can by shown rigorously e.g. by induction)

- Next step: Find a bound for the right hand side.
- Note how the global error at t_n depends on the errors on all the previous steps!

Proof 3 / 3

- It is $\exp(x) \geq 1 + x$ for all $x > 0$ (without proof here)
- Let $x = h|\lambda| > 0$ so that

$$|1 + h\lambda| \leq 1 + h|\lambda| \leq \exp(h|\lambda|) \tag{15}$$

- Now the absolute value of the terms in the sum can be estimated by

$$|1 + h\lambda|^{n-j} \leq \exp((n-j)h|\lambda|) = \exp(|\lambda| t_{n-j}) \leq \exp(|\lambda| T) \tag{16}$$

  using $(n-j)h = t_{n-j} \leq T$.
- Because $|T_j| \leq Ch^2$ for a constant C independent of j and h, we get

$$|e_n| \leq \sum_{j=1}^{n} \exp(|\lambda| T) Ch^2 \leq nh^2 C \exp(|\lambda| T) = C' Th \tag{17}$$

  using $T = nh$.
- In summary, $|e_n| = \mathcal{O}(h)$.

Backward Euler Method

- The Euler method sometimes is referred as forward Euler method. This means we have
- Backward Euler method:
  - In the forward Euler method, we make the approximation

$$\int_{t_n}^{t_{n+1}} f(s, x(s))ds = hf(t_n, x(t_n)).$$

  - Instead of this, we can also make an approximation as

$$\int_{t_n}^{t_{n+1}} f(s, x(s))ds = hf(t_{n+1}, x(t_{n+1})).$$

  - Then we have the backward Euler method

$$x_{n+1} = x_n + hf(t_{n+1}, x_{n+1}).$$

Forward Euler vs. Backward Euler

- What is the difference between these two Euler methods?
  - Forward Euler (explicit):

$$x_{n+1} = x_n + hf(t_n, x_n).$$

  - Backward Euler (implicit):

$$x_{n+1} = x_n + hf(t_{n+1}, x_{n+1}).$$

- In the backward Euler method, we need to solve a equation (usually nonlinear) to obtain x_{n+1}. Such as Newton's method.
- Stability, we will discuss this in future, but you will have a first impression about this concept in the exercise.

Higher order methods

- We showed the Euler is convergent with order $\mathcal{O}(h)$. Unless we have a constant $C \ll 1$, reaching an error of e.g. 10^{-6} will require at least a million time steps (if $h = 0.1$)!
- Method of order $p > 1$ can be much more efficient here: For $p = 2$, would need only a thousand steps, e.g. For $p = 6$ only ten! (In an ideal world, at least...)
- One approach is computing additional *intermediate steps* in a time step (so-called *stages*). This leads to the class of *Runge-Kutta methods* (RKM). Such as

$$k_1 = x_n + hf(t_n, x_n) \tag{18a}$$

$$x_{n+1} = x_n + \frac{h}{2}\left(f(t_n, x_n) + f(t_{n+1}, k_1)\right) \tag{18b}$$

- Another approach is using a number of *old* values $x_{n-1}, x_{n-2}, \ldots$. This leads to *linear multi-step methods* (LLM).

Trapezoidal Rule

- Start again with Taylor expansion of the solution

$$x(t+h) = x(t) + hx'(t) + \frac{1}{2}h^2x''(t) + \mathcal{O}(h^3) \tag{19}$$

- In addition, compute expansion of the derivative of the solution

$$x'(t+h) = x'(t) + hx''(t) + \mathcal{O}(h^2) \tag{20}$$

so that

$$hx''(t) = x'(t+h) - x'(t) + \mathcal{O}(h^2) \tag{21}$$

- Combining both gives

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) + \frac{1}{2}h\left[x'(t+h) - x'(t) + \mathcal{O}(h^2)\right] + \mathcal{O}(h^3) && \text{(22a)} \\ &= x(t) + \frac{1}{2}h\left[x'(t+h) + x'(t)\right] + \mathcal{O}(h^3) && \text{(22b)} \end{aligned}$$

- Inserting $t = t_n$ and $x_n \approx x(t_n)$ while ignoring the remainder term gives

$$x_{n+1} = x_n + \frac{1}{2}h\left[f(t_{n+1}, x_{n+1}) + f(t_n, x_n)\right] \tag{23}$$

Two step Adams-Bashforth method AB(2)

- As for trapezoidal rule, start from

$$x(t+h) = x(t) + hx'(t) + \frac{1}{2}h^2 x''(t) + \mathcal{O}(h^3) \tag{24}$$

- But now expand

$$x'(t-h) = x'(t) - hx''(t) + \mathcal{O}(h^2) \Rightarrow hx''(t) = x'(t) - x'(t-h) + \mathcal{O}(h^2) \tag{25}$$

- Combination

$$x(t+h) = x(t) + hx'(t) + \frac{1}{2}h\left[x'(t) - x'(t-h) + \mathcal{O}(h^2)\right] + \mathcal{O}(h^3) \tag{26a}$$

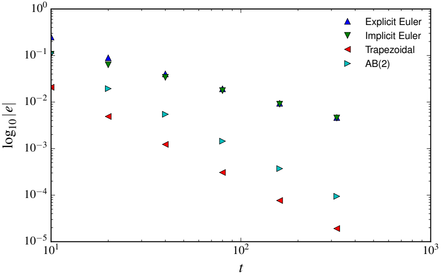
$$= x(t) + \frac{1}{2}h\left[3x'(t) - x'(t-h)\right] + \mathcal{O}(h^3) \tag{26b}$$

- Gives rise to the method

$$x_{n+1} = x_n + \frac{1}{2}h\left(3f_n - f_{n-1}\right) \tag{27}$$

- What is difference between trapezoidal rule and AB(2)?

- Consider IVP
$$x'(t) = -2x(t), \quad t \in [0,4], \quad x(0) = 1.$$
- For $N = 10, 20, 40, 80, 160, 320$, recall $e = \max_{1 \le n \le N} |e_n|$.



- What is the difference about the slopes?

Examples of LMMs

k	p	Method	Name
1	1	$x_{n+1} = x_n + hf_n$	forward Euler
1	1	$x_{n+1} = x_n + hf_{n+1}$	backward Euler
1	2	$x_{n+1} = x_n + \frac{1}{2}h\,(f_n + f_{n+1})$	trapezoidal rule
2	2	$x_{n+1} = x_n + \frac{1}{2}h\,(3f_n - f_{n-1})$	Adams-Bashforth-2
2	2	$x_{n+1} = x_n + \frac{1}{12}h\,(5f_{n+1} + 8f_n - f_{n-1})$	Adams-Moulton-2
2	4	$x_{n+1} = x_{n-1} + \frac{1}{3}h\,(f_{n+1} + 4f_n + f_{n-1})$	Simpson's rule
2	3	$x_{n+1} = -4x_n + 5x_{n-1} + h\,(4f_n + 2f_{n-1})$	Dahlquist

Table: Examples of LMMs with step number k and order p. Cf. Griffiths, Higham, p. 48.

General form of two-step methods

- Consider the following general form of expansions resulting in two-step methods

$$\begin{aligned} &x(t+2h) + \alpha_1 x(t+h) + \alpha_0 x(t) \\ &\quad = h\left[\beta_2 x'(t+2h) + \beta_1 x'(t+h) + \beta_0 x'(t)\right] + \mathcal{O}(h^{p+1}) \end{aligned}$$

- The corresponding two-step method then is

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right] \tag{28}$$

- **Note:** For $\beta_2 = 0$ we get an *explicit* method, for $\beta_2 \neq 0$ an *implicit* method.
- Parameters $(-1, 0, 1, 0, 0)$ or $(-1, 0, 0, 1, 0)$ lead to

$$x_{n+2} - x_{n+1} = hf_{n+2} \quad \text{and} \quad x_{n+2} - x_{n+1} = hf_{n+1}, \tag{29}$$

that is forward and backward Euler, with index shifted by one.

General form of two-step methods

- For parameters $(-1, 0, \frac{1}{2}, \frac{1}{2}, 0)$ we get

$$x(t+2h) - x(t+h) = h\left[\frac{1}{2}x'(t+2h) + \frac{1}{2}x'(t+h)\right] \tag{30}$$

  or

$$x_{n+2} - x_{n+1} = \frac{1}{2}h\left[f_{n+2} + f_{n+1}\right] \tag{31}$$

  which is trapezoidal rule shifted by one index.

- For parameters $(-1, 0, 0, \frac{3}{2}, -\frac{1}{2})$ we get

$$x(t+2h) - x(t+h) = h\left[\frac{3}{2}x'(t+h) - \frac{1}{2}x'(t)\right] \tag{32}$$

  or

$$x_{n+2} - x_{n+1} = \frac{1}{2}h\left[3f_{n+1} - f_n\right] \tag{33}$$

  which is Adams-Bashforth-2 with index shifted by one.

- How can we find the parameters? E.g. find coefficients $(\alpha_1, \alpha_0, \beta_2, \beta_1, \beta_0)$ that maximize p?

Consistency

Definition

The *linear difference operator* $\mathcal{L}_h$ associated with

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right] \tag{34}$$

is defined for some arbitrary continuously differentiable function $z(t)$ by

$$\begin{aligned}\mathcal{L}_h(z(t)) :=& z(t+2h) + \alpha_1 z(t+h) + \alpha_0 z(t) \\ & - h\left[\beta_2 z'(t+2h) + \beta_1 z'(t+h) + \beta_0 z'(t)\right]\end{aligned}$$

Note:

- $\mathcal{L}_h$ is a linear operator: $\mathcal{L}_h(az(t) + bw(t)) = a\mathcal{L}_h(z(t)) + b\mathcal{L}_h(w(t))$
- $\mathcal{L}_h$ is essentially equal to the remainder term in the Taylor series expansion

Consistency

Definition

A linear difference operator $\mathcal{L}_h$ is said to be *consistent of order p* if

$$\mathcal{L}_h(z(t)) = \mathcal{O}(h^{p+1}) \tag{35}$$

with $p > 0$ for *every* smooth function z.

- A LMM whose difference operator is consistent of order p for some $p > 0$ is said to be *consistent*. Otherwise, the LMM is called *inconsistent*.
- Note the $p+1$ in the definition above: A method that is consistent of order p has a LTE of order $p+1$ and can give rise to a method that is convergent of order p (if it is stable).

Consistency of forward Euler

- The linear difference operator for forward Euler is

$$\mathcal{L}_h(z(t)) = z(t+h) - z(t) - hz'(t) \tag{36}$$

- By Taylor expansion, for any smooth function $z(t)$, it is

$$\mathcal{L}_h(z(t)) = \frac{1}{2}h^2 z''(t) + \mathcal{O}(h^3) \tag{37}$$

so that $\mathcal{L}_h(z(t)) = \mathcal{O}(h^2)$ and the difference operator is consistent with order $p = 1$.

Construction of LMMs – I

- For the general two-step LLM

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right] \tag{38}$$

the associated difference operator is

$$\begin{aligned}\mathcal{L}_h z(t) = {} & z(t+2h) + \alpha_1 z(t+h) + \alpha_0 z(t) - \\ & h\left[\beta_2 z'(t+2h) + \beta_1 z'(t+h) + \beta_0 z'(t)\right]\end{aligned}$$

- Consider the expansions

$$z(t+2h) = z(t) + 2hz'(t) + 2h^2 z''(t) + \dots \tag{39}$$

$$z(t+h) = z(t) + hz'(t) + \frac{1}{2}h^2 z''(t) + \dots \tag{40}$$

$$z'(t+2h) = z'(t) + 2hz''(t) + 2h^2 z'''(t) + \dots \tag{41}$$

$$z'(t+h) = z'(t) + hz''(t) + \frac{1}{2}h^2 z'''(t) + \dots \tag{42}$$

that allow to express $\mathcal{L}_h$ using only $z(t)$, $z'(t)$, etc.

Construction of LMMs – II

- Construct LLM which is consistent of at least order $p = 1$
- Appropriate collection of terms gives

$$\mathcal{L}_h(z(t)) = (1 + \alpha_1 + \alpha_0)\, z(t) + h\left[2 + \alpha_1 - (\beta_2 + \beta_1 + \beta_0)\right] z'(t) + \mathcal{O}(h^2)$$

- Consistency of order $p = 1$ requires

$$1 + \alpha_1 + \alpha_0 = 0 \tag{43a}$$

$$2 + \alpha_1 = \beta_2 + \beta_1 + \beta_0 \tag{43b}$$

- Had so far $(-1, 0, 1, 0, 0)$ and $(-1, 0, 0, 1, 0)$ (Euler methods): Check.
- Also $(-1, 0, 0, \frac{3}{2}, -\frac{1}{2})$ (AB-2):

$$1 - 1 + 0 == 0 \tag{44a}$$

$$2 - 1 == \frac{3}{2} - \frac{1}{2} + 0 \tag{44b}$$

Construction of LMMs – III

Definition

The *first* and *second characteristic polynomial* of the LMM

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right] \tag{45}$$

are defined to be

$$\rho(r) = r^2 + \alpha_1 r + \alpha_0, \quad \sigma(r) = \beta_2 r^2 + \beta_1 r + \beta_0 \tag{46}$$

- In the following lectures, will will connect properties of these polynomials to properties of the associated LMM
- The concept can of course be extended to LMMs with more than two steps

Consistency and characteristic polynomials

Theorem

The two-step LMM

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right] \tag{47}$$

is consistent with the ODE $x'(t) = f(t, x(t))$ if, and only if,

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1). \tag{48}$$

Proof.

Need to show that $\mathcal{L}_h(z(t)) = \mathcal{O}(h^{p+1})$ for some $p > 0$. Appropriate collection of higher order terms (as indicated before) gives

$$\mathcal{L}_h(z(t)) = C_0 z(t) + C_1 h z'(t) + \ldots + C_p h^p z^{(p)}(t) + \mathcal{O}(h^{p+1}) \tag{49}$$

with $C_0 = 1 + \alpha_1 + \alpha_0 = \rho(1)$ *and*
$C_1 = 2 + \alpha_1 - (\beta_1 + \beta_2 + \beta_3) = \rho'(1) - \sigma(1)$. □

Convergence

Theorem

A convergent LMM is consistent.

- Suppose that the LMM

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right] \tag{50}$$

  is convergent to the exact solution $x(t)$.
- Convergence means the global error vanishes as $h \to 0$, which implies that

$$x_{n+2} \to x(t^* + 2h), \quad x_{n+1} \to x(t^* + h), \quad x_n \to x(t^*) \tag{51}$$

  as $h \to 0$ when $t_n = t^*$.
- Because also $t_{n+2}, t_{n+1} \to t^*$, taking the limit of both sides of (87) leads to

$$\rho(1)x(t^*) = 0. \tag{52}$$

- In general $x(t^*) \neq 0$, and so $\rho(1) = 0$ and the first consistency condition is met.

Convergence - II

- Now to the second condition... consider

$$\frac{x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n}{h} = \beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n. \tag{53}$$

- The right hand sides converges to $\sigma(1) f(t^*, x(t^*))$.
- For the left hand side, by l'Hospital's rule, we get

$$\lim_{h \to 0} \frac{x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n}{h} = (2 + \alpha_1)\, x'(t^*) \tag{54}$$

  using $\partial_h x_{n+2} = 2x'(t^* + 2h)$, $\partial_h x_{n+1} = x'(t^* + h)$ and $\partial_h x_n = 0$.
- Thus, the function $x(t)$ satisfies at $t = t^*$

$$\rho'(1) x'(t^*) = \sigma(1) f(t^*, x(t^*)) \tag{55}$$

  which, unless $\rho'(1) = \sigma(1)$, is not the correct ODE.
- **Key point here:** A non-consistent LMM cannot be convergent!

General k-step methods - I

- The general form of a k-step LMM is

$$x_{n+k} + \alpha_{k-1}x_{n+k-1} + \ldots + \alpha_0 x_n = \tag{56a}$$

$$h\left[\beta_k f_{n+k} + \beta_{k-1} f_{n+k-1} + \ldots + \beta_0 f_n\right] \tag{56b}$$

  It is implicit unless $\beta_k = 0$.

- The characteristic polynomials read

$$\rho(r) = r^k + \alpha_{k-1} r^{k-1} + \ldots + \alpha_0 \tag{57}$$

  and

$$\sigma(r) = \beta_k r^k + \beta_{k-1} r^{k-1} + \ldots + \beta_0 \tag{58}$$

- The associated linear difference operator is

$$\mathcal{L}_h(z(t)) = \sum_{j=0}^{k} \alpha_j z(t+jh) - h\beta_j z'(t+jh) \tag{59}$$

General k-step methods - II

- The difference operator can be expanded as

$$\mathcal{L}_h(z(t)) = C_0 z(t) + C_1 h z'(t) + \dots \tag{60}$$
$$+ C_p h^p z^{(p)}(t) + C_{p+1} h^{p+1} z^{(p+1)}(t) + \mathcal{O}(h^{p+2}) \tag{61}$$

  with $C_0 = \rho(1)$ and $C_1 = \rho'(1) - \sigma(1)$.
- Our consistency conditions derived for 2-step methods apply here, too!
- The method has order p if $C_0 = C_1 = \dots = C_p = 0$; the first non-zero coefficient C_{p+1} is called the *error constant*.
- We have $2k+1$ arbitrary coefficients α_k and β_k to set for an implicit method and $2k$ for an explicit method, ideally, we could eliminate the same number of coefficients and generate an order $2k$ implicit or $2k-1$ explicit method.
- However, because of stability, convergent methods do, in general, not achieve such high orders!

Convergence

Definition

The LMM

$$\sum_{j=0}^{k-1} \alpha_j x_{n+j} + x_{n+k} = h \sum_{j=0}^{k} \beta_k f_{n+k}$$

with starting values satisfying

$$\lim_{h\to 0} x_j = \eta, \quad j = 0, \ldots, k-1.$$

is said to be convergent, if for all initial value problems $x'(t) = f(t, x(t))$, $x(0) = \eta$ with a unique solution on $[0, T]$,

$$\lim_{h\to 0, nh=t^*} x_n = x(t^*) \tag{62}$$

holds for all $t^* \in [0, T]$.

- A convergent LMM is consistent.
- Recall that consistency implies $\rho(1) = 0$ and $\rho'(1) = \sigma(1)$:

$$\sum_{j=0}^{k} \alpha_j = 0 \quad \sum_{j=0}^{k} j\alpha_k = \sum_{j=0}^{k} \beta_k \tag{63}$$

with $\alpha_k = 1$.

- Does a consistent LMM also convergent?

A consistent yet useless LMM

- Consider the LMM (Dahlquist)

$$x_{n+2} + 4x_{n+1} - 5x_n = h\left[4f_{n+1} + 2f_n\right] \tag{64}$$

It has characteristic polynomials

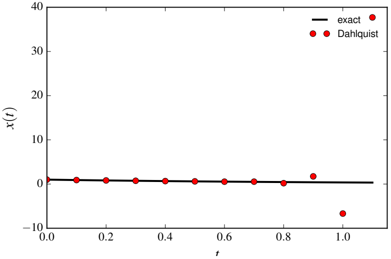
$$\rho(r) = r^2 + 4r - 5 \quad \text{and} \quad \sigma(r) = 4r + 2. \tag{65}$$

- The method satisfies $\rho(1) = 0$ and $\rho'(1) = \sigma(1)$ and is thus consistent.
- Apply it to the IVP

$$x'(t) = -x(t) \tag{66}$$

$$x(0) = 1 \tag{67}$$

- For $h = 0.1$,



- Why does the Dahlquist method not stable (or fail to converge)?

Reason of failure

- Apply it to the trivial IVP $x'(t) = 0$ with $x(0) = 1$ with starting values $x_0 = 1$ and $x_1 = 1 + h$ (slightly perturbed). Note that $x_1 \to x(t_1) = 1$ as $h \to 0$!
- Leads to recursion

$$x_{n+2} + 4x_{n+1} - 5x_n = 0 \tag{68}$$

with auxiliary equation (cf. Appendix of Griffiths book)

$$r^2 + 4r - 5 = (r-1)(r+5)\,. \tag{69}$$

- The general solution of the recursion is (again, trust the book)

$$x_n = c_1 + c_2\,(-5)^n \tag{70}$$

and from the starting values we get $A + B = 1$ and $1 + h = A + B(-5)$ so that $A = 1 + h/6$ and $B = -h/6$ and

$$x_n = 1 - \frac{1}{6}h\,[1 + (-5)^n] \tag{71}$$

Reason of failure

- In the closed form of the recursion

$$x_n = 1 - \frac{1}{6} h \left[1 + (-5)^n\right] \tag{72}$$

it is clearly the term $(-5)^n$ that leads to disaster: Suppose e.g. that $t = nh = 1$, i.e.$h = 1/n$ so that

$$h(-5)^n = \frac{5^n}{n} \to \infty \text{ as } h \to 0 \text{ and } n \to \infty. \tag{73}$$

- The auxiliary polynomial is the first characteristic polynomial, which satisfies $\rho(1) = 1$, i.e. $r = 1$ is one root for all consistent methods
- A two step LMM can factorize

$$\rho(r) = (r-1)\,(r-a)\,.$$

- If $a \neq 1$, applying the method corresponding to this polynomial to $x'(t) = 0$ gives

$$x_n = c_1 + c_2 a^n$$

  Should look for methods with $|a| \leq 1$ *to avoid blow-up as* $n \to \infty$
- If $a \neq 1$,

$$x_n = c_1 + c_2 n$$

  Still blow-up as $n \to \infty$. The problem comes from the double root of $r = 1$.
- In general, for k-step LMMs ($k \geq 3$), double root of $r = -1$ also cases problem. For example

$$x_{n+3} + x_{n+2} - x_{n+1} - x_n = 4hf_n$$

Root condition

Definition

A polynomial ρ is said to satisfy the *root condition*, if all its roots r lie within or on the unit circle (i.e. $|r| \leq 1$) and all roots on the boundary (i.e. with $|r| = 1$) are simple roots.

A polynomial satisfies the *strict root condition* if all roots lie inside the unit circle, i.e. $|r| < 1$.

- Note: Simple root means that $\lambda - r$ is a factor of $\rho(r)$ but $(\lambda - r)^2$ is not.
- For example, $\rho(r) = r^2 - 1 = (r + 1)(r - 1)$ satisfies the root condition, but $\rho(r) = (r - 1)^2$ does not.

Zero-stability

Definition

A LMM is said to be *zero-stable* if its first characteristic polynomial $\rho(r)$ satisfies the root condition.

- Note: All consistent one-step methods have $\rho(r) = r - 1$ and automatically satisfy the root condition. Therefore, it did not pop up when studying Euler methods.

The end of it all

Theorem (Dahlquist Equivalence Theorem (1956))

For LMMs applied to the IVP $x'(t) = f(t, x(t))$,

$$\textit{consistency} + \textit{zero-stability} \Leftrightarrow \textit{convergence}$$

Theorem (First Dahlquist Barrier (1959))

A zero-stable k-step LMM can not attain an order of convergence greater than $k+1$ if k is odd and greater than $k+2$ if k is even. If the method is also explicit, then it can not attain an order greater than k.

- Our previous method

$$x_{n+2} + 4x_{n+1} - 5x_n = h(4f_{n+1} + 2f_n) \tag{74}$$

can now immediately be dismissed: It has $k = 2$, $p = 3$ and is explicit and therefore cannot be stable.

Famous families of LMM

- Adams-Bashforth (1883):
  1. Characteristic polynomials $\rho(r) = r^k - r^{k-1} = r^{k-1}(r-1)$
  2. Explicit: $x_{n+k} - x_{n+k-1} = h\sum_{j=0}^{k-1} \beta_j f_{n+j}$ with $\beta_k = 0$
  3. Coefficients β_j chosen such that $C_0 = C_1 = \ldots = C_{k-1} = 0$ in $\mathcal{L}_h$, giving order $p = k$.
  4. Important members: Forward Euler ($k = 1$), AB(2) ($k = 2$) and AB(3) ($k = 3$), reading

$$x_{n+3} - x_{n+2} = \frac{h}{12}\left(23f_{n+2} - 48f_{n+1} + 5f_n\right) \tag{75}$$

- Adams-Moulton (1926):
  1. Implicit version of Adams-Bashforth, i.e. use $\beta_k \neq 0$.
  2. Order $p = k + 1$
  3. Examples: Trapezoidal rule ($k = 1$), AM(2) and AM(3):

$$x_{n+3} - x_{n+2} = \frac{h}{24}\left(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n\right) \tag{76}$$

Famous families of LMM

- Nyström methods (1925):
  1. Explicit methods with $k \leq 2$
  2. $\rho(r) = r^k - r^{k-2}$, so general form $x_{n+k} - x_{n+k-2} = h\sum_{j=0}^{k-1} \beta_j f_{n+j}$
  3. Again choose β_j to achieve $k = p$, as for Adams-Bashforth.
  4. Examples: Midpoint rule ($k = 1$)
- Milne-Simpson (1926):
  1. Implicit analogues of Nyström methods, i.e. $\beta_k \neq 0$.
  2. Example: Simpson rule with $k = 2$ and $p = 4$ (maximum order)
- Backward differentiation formulas (BDF, 1952):
  1. Generalization of backward Euler
  2. Simplest possible second characteristic polynomial for an implicit method: $\sigma(r) = \beta_k r^k$, thus general form $\sum_{j=0}^{k} \alpha_k x_{n+k} = h\beta_k f_{n+k}$
  3. $k+1$ free coefficients chosen for order $p = k$ (not the optimal $k+2$, alas)
  4. Important family, because of compensating strengths, cf. Chapter 6.

Convergence Check

- Let the global error $e = \max_{1 \le j \le n} |e_j|$
- For a order p LMM with step size h, we have the consistency:

$$\mathcal{L}(z(t)) = C_{p+1} h^{p+1} + \mathcal{O}(h^{p+2})$$

  and convergence:

$$e_h = Ch^p + \mathcal{O}(h^{p+1})$$

- If we half the step size, $h/2$, then

$$e_{h/2} = C \left(\frac{h}{2}\right)^p + \mathcal{O}(\left(\frac{h}{2}\right)^{p+1}).$$

- Thus,

$$\frac{e_h}{e_{h/2}} = \frac{Ch^p + \mathcal{O}(h^{p+1})}{C\left(\frac{h}{2}\right)^p + \mathcal{O}(\left(\frac{h}{2}\right)^{p+1})} \to \frac{C}{C\left(\frac{h}{2}\right)^p} = 2^p \qquad \text{as } h \to 0$$

- Numerically, $p = \log_2\left(\frac{e_h}{e_{h/2}}\right)$, see check_conv.py

Recurrence relations and auxiliary polynomials

- Consider the linear constant-coefficient recursion

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = 0$$

- Suppose $u_n = r^n$ and plug into the scheme:

$$\sum_{j=0}^{k} \alpha_j r^{n+j} = 0 \Rightarrow \sum_{j=0}^{k} \alpha_j r^j = 0$$

- Thus, r must be a root of the auxiliary polynomial

$$\rho(r) = \sum_{j=0}^{k} \alpha_j r^j = \alpha_k (r - r_1)(r - r_2) \cdots (r - r_k)$$

- Any linear combination is also a solution, so if all r_j distinct:

$$u_n = c_1 r_1^n + c_2 r_2^n + \cdots + c_k r_k^n$$

Recurrence relations and auxiliary polynomials

- The coefficients c_i can be determined from the initial condition $u_0, u_1, \ldots, u_{r-1}$:

$$\begin{aligned} c_1 + c_2 + \cdots + c_k &= u_0 \\ c_1 r_1 + c_2 r_2 + \cdots + c_k r_k &= u_1 \\ &\cdots \\ c_1 r_1^{k-1} + c_2 r_2^{k-1} + \cdots + c_k r_k^{k-1} &= u_{k-1} \end{aligned}$$

- Example: $u_{n+2} + 4u_{n+1} - 5u_n = 0$

$$\rho(r) = (r+5)(r-1) \Rightarrow u_n = c_1 + c_2(-5)^n$$

- If a root is repeated, for example $r_1 = r_2 = \cdots = r_d$, the solution is

$$u_n = (\sum_{i=1}^{d} c_i n^{i-1}) r_1^n + c_{d+1} r_{d+1}^n + \cdots c_k r_k^n.$$

- Example: $u_{n+3} - 2u_{n+2} + \frac{5}{4} u_{n+1} - \frac{1}{4} u_n = 0$

$$\rho(r) = (r-1)(r-0.5)^2 \Rightarrow u_n = (c_1 + c_2 n)0.5^n + c_3$$

Recurrence relations and polynomials

- Consider e.g. the recursion

$$x_{n+2} + ax_{n+1} + bx_n = 0.$$

Its so-called auxiliary equation reads

$$r^2 + ar + b = \rho(r) = 0. \tag{77}$$

- If the auxiliary equation has roots r_1 and r_2, the general solution reads

$$x_n = \begin{cases} c_1 r_1^n + c_2 r_2^n & : r_1 \neq r_2 \\ (c_1 + c_2 n) r_1^n & : r_1 = r_2 \end{cases}$$

for some constants r_1, r_2.

- Again, long-term behavior of the solution is governed by the roots of the auxiliary equation / the characteristic polynomial

Recall

Before, we analyzed convergence of Euler's method for the IVP

$$x'(t) = \lambda x(t) + g(t), \quad 0 < t \leq T \tag{78a}$$

$$x(0) = 1 \tag{78b}$$

- Now: Extend to $k = 2$
- Key concepts carry over to general LMM.

Local truncation error

- As shown before, LMMs are constructed from Taylor expansions such that

$$\mathcal{L}_h(z(t)) = C_{p+1}h^{p+1}z^{(p+1)} + \dots \tag{79}$$

  where z is some arbitrary, continuously differentiable function.
- The LTE, denoted as T_{n+2}, is defined as the difference operator applied to the exact solution x of the IVP at $t = t_{n+2}$, i.e.

$$T_{n+2} = \mathcal{L}_h(x(t_n)) \tag{80}$$

- **If the solution is $(p+1)$-times continuously differentiable**, we have

$$T_{n+2} = C_{p+1}h^{p+1}x^{(p+1)}(t) + \dots \tag{81}$$

  so that

$$T_{n+2} = \mathcal{O}(h^{p+1}) \tag{82}$$

Local Truncation Error

- Denote as $y_n = x(t_n)$ the exact solution of the IVP at a grid point t_n. The general two-step LMM applied to

$$x'(t) = \lambda x(t) + g(t), \quad 0 < t \leq T \tag{83a}$$
$$x(0) = 1 \tag{83b}$$

results in

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_2 x_n = \\ h\lambda\left(\beta_2 x_{n+2} + \beta_1 x_{n+1} + \beta_0 x_n\right) + h\left(\beta_2 g(t_{n+2}) + \beta_1 g(t_{n+1}) + \beta_0 g(t_n)\right) \tag{84}$$

- The exact solution y_n satisfies

$$y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n =$$
$$h\lambda\left(\beta_2 y_{n+2} + \beta_1 y_{n+1} + \beta_0 y_n\right) \tag{85}$$
$$+ h\left(\beta_2 y_{n+2} + \beta_1 y_{n+1} + \beta_0 y_n\right) + T_{n+2} \tag{86}$$

- This is analogously to what we did for the Euler method!

Global error

- Again, subtracting the recursion for the approximate solution x_{n+1} and the recursion for the exact solution y_{n+1} gives a recursion for the global error:

$$(1 - h\lambda\beta_2)\, e_{n+2} + (\alpha_1 - h\lambda\beta_1)\, e_{n+1} + (\alpha_0 - h\lambda\beta_0)\, e_n = T_{n+2} \tag{87}$$

with starting values $e_0 = 0$ and $e_1 = x(t_1) - \eta_1 = \mathcal{O}(h)$.

- Eq. (87) governs how *local* errors T_{n+2} accumulate into a *global* error e_n.

- Now, for simplicity, assume that $T_{n+2} = T$ for all n. Then, we can again derive a solution of the recursion in closed form by superimposing a particular solution with the general homogeneous solution.

Global error

- Particular solution with $e_n = P$ constant

$$P = \frac{T}{h\lambda\sigma(1)}. \tag{88}$$

So for $T = \mathcal{O}(h^{p+1})$ it follows that $P = \mathcal{O}(h^p)$.

- General homogeneous solution: If the auxiliary equation

$$(1 - h\lambda\beta_2)r^2 + (\alpha_1 - h\lambda\beta_1)\, r + (\alpha_0 - h\lambda\beta_0) = 0 \tag{89}$$

has distinct roots $r_1 \neq r_2$, the contribution reads

$$Ar_1^n + Br_2^n \tag{90}$$

- The general solution then is

$$e_n = Ar_1^n + Br_2^n + P \tag{91}$$

with constants A and B depending on the starting values.

Discussion of global error

- As $h \to 0$, the roots r_1, r_2 of

$$(1 - h\lambda\beta_2)r^2 + (\alpha_1 - h\lambda\beta_1)\, r + (\alpha_0 - h\lambda\beta_0) = 0 \tag{92}$$

  tend to the roots of the first characteristic polynomial $\rho(r)$: If the root condition is violated, they will lead to divergence.

- Note that the LTE contributes through the term $P = \mathcal{O}(h^p)$; so consistency $(p > 0)$ ensures $P \to 0$ as $h \to 0$.

- The equation very nicely illustrates the interplay between LTE and accumulation:
  1. Consistency ensures that P is small
  2. Zero-stability ensures that local errors are not amplified through the first two terms

Interpretation of the LTE

- Suppose we have exact values y_{n+1}, y_n out of which we compute an approximate value $\tilde{x}_{n+2}$:

$$\tilde{x}_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = \\ h\lambda\left(\beta_2 \tilde{x}_{n+2} + \beta_1 y_{n+1} + \beta_0 y_n\right) + h\left(\beta_2 g(t_{n+2}) + \beta_1 g(t_{n+1}) + \beta_0 g(t_n)\right) \tag{93}$$

- Again, subtract this from the recursion for the exact solution:

$$\left(1 - h\lambda\beta_2\right)\left(y_{n+2} - \tilde{x}_{n+2}\right) = T_{n+2} \tag{94}$$

- For an explicit method ($\beta_2 \neq 0$), this reduces to

$$y_{n+2} - \tilde{x}_{n+2} = T_{n+2}, \tag{95}$$

i.e. *the LTE is the error committed in one step if the back values are exact* ("localizing assumption")

Interpretation of the LTE

- Have

$$(1 - h\lambda\beta_2)\,(y_{n+2} - \tilde{x}_{n+2}) = T_{n+2} \tag{96}$$

- For an implicit method ($\beta_2 \neq 0$), expand

$$(1 - h\lambda\beta_2)^{-1} = 1 + h\lambda\beta_2 + \mathcal{O}(h^2) = 1 + \mathcal{O}(h) \tag{97}$$

  so that

$$(1 + \mathcal{O}(h))\,T_{n+2} = y_{n+2} - \tilde{x}_{n+2} \tag{98}$$

  and because $T_{n+2} = \mathcal{O}(h^{p+1})$,

$$T_{n+2} = y_{n+2} - \tilde{x}_{n+2} + \mathcal{O}(h^{p+2}) \tag{99}$$

- Here, *the leading term in the LTE is the error committed in one step if the back values are exact.*

Absolute Stability

- Convergence involves the limit $x_n \to x(t^*)$ for $n \to \infty$ and $h \to 0$ in such a way that $t_n = t_0 + nh = t^*$ is fixed.
- Thus, convergent methods generate solutions arbitrarily close to the exact solution, *if h is made sufficiently small.*
- In the following, we investigate performance of methods when h is finite, e.g. not arbitrarily small.

Example

- Consider the IVP

$$x'(t) = -8x(t) - 40\left(3\exp\left(-\frac{t}{8} - 1\right)\right), \quad x(0) = 100 \tag{100}$$

- Analytic solution

$$x(t) = \frac{1675}{21}\exp(-8t) + \frac{320}{21}\exp\left(-\frac{t}{8}\right) + 5 \tag{101}$$

- Typical behavior: For problems with exponentially decaying solution, forward Euler is unstable unless h is very small.
- This points into the direction of a much more general topic: So-called *stiff problems* which can be characterized loosely as problems for which explicit methods require a very small time step and are generally not effective.
- To avoid instability, we are forced to use a very small time step and produce a solution that is probably much more accurate than required

Example

- Now consider IVP

$$x'(t) = -\frac{1}{8}\left(x(t) - 5 - 5025\exp\left(-8t\right)\right), \quad x(0) = 100 \tag{102}$$

- Analytic solution is again

$$x(t) = \frac{1675}{21}\exp\left(-8t\right) + \frac{320}{21}\exp\left(-\frac{t}{8}\right) + 5 \tag{103}$$

- Apparently, the stability problems stem from the exponentially decaying term in the ODE, not the source term: Therefore, analyze the homogeneous problem in the following.

Absolute stability

- Examine what happens if we apply a convergent LMM to the scalar model problem
$$x'(t) = \lambda x(t), \quad \lambda \in \mathbb{C} \text{ with } \mathrm{Re}(\lambda) < 0. \tag{104}$$
- The exact solution is
$$x(t) = c\exp(\lambda t) \tag{105}$$
and $x(t) \to 0$ as $t \to \infty$.
- Look for LMM that have
$$x_n \to 0 \text{ as } n \to \infty \tag{106}$$
for a *fixed* step size h (note how this is different from convergence).

Absolute Stability

Definition

A LMM is said to be *absolutely stable*, if, when applied to the scalar test problem with $\mathrm{Re}(\lambda) < 0$, and a given fixed value $\hat{h} = h\lambda$, its solutions tend to zero as $n \to \infty$ for any choice of starting values.

- Consider e.g. implicit Euler

$$x_{n+1} = \left(\frac{1}{1 - \hat{h}} \right)^{n+1} x_0 \to 0, \text{ because } \hat{h} > 0. \tag{107}$$

  This holds for any x_0, so implicit Euler is absolutely stable.

- More generally, consider the general two-step LMM

$$\left(1 - \hat{h}\beta_2\right) x_{n+2} + \left(\alpha_1 - \hat{h}\beta_1\right) x_{n+1} + \left(\alpha_0 - \hat{h}\beta_0\right) x_n = 0 \tag{108}$$

Absolute Stability

- The auxiliary equation reads

$$\left(1 - \hat{h}\beta_2\right) r^2 + \left(\alpha_1 - \hat{h}\beta_1\right) r + \left(\alpha_0 - \hat{h}\beta_0\right) = 0 \tag{109}$$

- Denote the polynomial as p and note that $p(r) = \rho(r) - \hat{h}\sigma(r)$.
- Polynomial p is called the *stability polynomial* of the LMM. It has two roots r_1 and r_2, so the general solution of the recursion is

$$x_n = ar_1^n + br_2^n \tag{110}$$

  assuming $r_1 \neq r_2$.
- The LMM is hence absolutely stable, if and only if the stability polynomial satisfies the *strict root condition*: That is, $|r_1| < 1$, $|r_2| < 1$.

- For general k step LMM,

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}, \qquad \alpha_k = 1$$

- The stability polynomial

$$p(r) = \sum_{j=0}^{k} (\alpha_j - \hat{h}\beta_j) r^j$$

- The LMM is hence absolutely stable, if and only if the stability polynomial satisfies the strict root condition: $|r_j| < 1$, $j = 1, \ldots, k$

Region of absolute stability

Definition

The set of values $\mathcal{R} := \{z \in \mathbb{C} : \text{LMM is absolutely stable}\} \subset \mathbb{C}$ is called the *region of absolute stability*.

Definition

The largest interval $\mathcal{R}_0 = \left(\hat{h}_0, 0\right) \subset \mathbb{R}$ with $\hat{h}_0 < 0$ for which the LMM is absolutely stable for all values $\hat{h} \in \mathcal{R}_0$ is called *interval of absolute stability.*

- As shown before, for the implicit Euler it is $\mathcal{R} = \mathbb{C}^-$ (left complex half-plane) and $\mathcal{R}_0 = \mathbb{R}^-$ (negative real axis).

Forward Euler: Region of absolute stability

- Forward Euler applied to the model problem gives

$$x_{n+1} = \left(1 + \hat{h}\right) x_n \tag{111}$$

- The stability polynomial is

$$p(r) = r - 1 - \hat{h} \tag{112}$$

  with single root $r_1 = 1 + \hat{h}$.
- The region where $|r_1| < 1$ is thus a circle with radius 1 around $\hat{h} = -1$:

$$\mathcal{R} = \{z \in \mathbb{C} : |z + 1| < 1\}, \quad \mathcal{R}_0 = (-2, 0). \tag{113}$$

- This is again the bound

$$h < \frac{2}{|\lambda|} \tag{114}$$

- For $x'(t) = -8x(t)$, $\hat{h} = -8h$ and absolute stability requires $h < 1/4$; for $x'(t) = -80x(t)$, however, $h < 1/40$ is required!

- How to find easy-to-check reformulation of criterion for absolute stability?
- How to find the region of absolute stability?
- How to find the interval of absolute stability?

- It is complicated, since the absolute stability is related to the roots of stability polynomial

$$\sum_{j=0}^{k}(\alpha_j - \hat{h}\beta_j)r^j = 0$$

- Let us consider the interval of absolute stability of the two step method.

Jury conditions

Lemma

A quadratic polynomial $q(r) = r^2 + ar + b$ with $a, b \in \mathbb{R}$, satisfies the strict root condition if and only if the three following conditions are fulfilled:

$(i)\ q(0) = b < 1, \quad (ii)\ q(1) = 1 + a + b > 0, \quad$ *and* $\quad (iii)\ q(-1) = 1 - a + b > 0$

- The proof is left for exercise.
- The roots of the polynomial are

$$r_1, r_2 = \frac{1}{2}\left(-a \pm \sqrt{a^2 - 4b}\right). \tag{115}$$

Find out when the strict root condition holds $|r_1| < 1$ and $|r_2| < 1$.

- When $a^2 < 4b$,
- When $a^2 \geq 4b$,

Jury conditions

- The stability polynomial of the general two-step LMM is

$$\left(1-\hat{h}\beta_2\right) r^2+\left(\alpha_1-\hat{h}\beta_1\right) r+\left(\alpha_0-\hat{h}\beta_0\right)=0$$

- Normalize to

$$r^2+\frac{\alpha_1-\hat{h}\beta_1}{1-\hat{h}\beta_2} r+\frac{\alpha_0-\hat{h}\beta_0}{1-\hat{h}\beta_2}=0$$

- Because the denominators are always positive for $\hat{h}>0$, $q(\pm 1)>0$ can be replaced by $p(\pm 1)>0$.
- Also, for explicit methods we have $\beta_2=0$ and $q(r)$ coincides with the stability polynomial

Example

- Find the interval of absolute stability of the two step Adams Bashforth method
$$x_{n+2} - x_{n+1} = h(\frac{3}{2}f_{n+1} - \frac{1}{2}f_n) \tag{116}$$
- The stability polynomial is
$$p(r) = r^2 - (1 + \frac{3}{2}\hat{h})r + \frac{1}{2}\hat{h} \tag{117}$$
If λ in $\hat{h} = h\lambda$ is real, the coefficients are real and we can use the Lemma.
- It is $q(r) \equiv p(r)$, so we can check for $p(\pm 1) > 0$ and $p(0) < 1$:
$$p(0) < 1 \Leftrightarrow \frac{1}{2}\hat{h} < 1 \Leftrightarrow \hat{h} < 2 \tag{118a}$$
$$p(1) > 0 \Leftrightarrow -\hat{h} > 0 \Leftrightarrow \hat{h} < 0 \tag{118b}$$
$$p(-1) > 0 \Leftrightarrow 2 + 2\hat{h} > 0 \Leftrightarrow \hat{h} > -1. \tag{118c}$$
- To satisfy all three, we need $-1 < \hat{h} < 0$, so the interval of absolute stability is $\mathcal{R}_0 = (-1, 0)$.
- Exercise: find the interval of absolute stability of the two step Adams-Moulton method.

Example

- Find the interval of absolute stability for the two-step mid-point rule

$$x_{n+2} - x_n = 2hf_{n+1} \tag{119}$$

- Excursus: Consistency and zero-stability? Please **do** try this at home.
- The stability polynomial is

$$p(r) = r^2 - 2\hat{h}r - 1 \tag{120}$$

  with roots $r_+ = \hat{h} + \sqrt{1 + \hat{h}^2}$ and $r_- = \hat{h} - \sqrt{1 + \hat{h}^2}$.
- It is $\sqrt{1 + \hat{h}^2} > 1$ and thus, for $\hat{h} < 0$, $r_- < -1$ and $|r_-| > 1$: The method can never be absolutely stable!
- Effect: see `oct_midpoint.py`...method applied to $x'(t) = -8x(t)$, $x(0) = 1$.

Example

- Solutions produced by midpoint rule for this specific problem read

$$x_n = Ar_+^n + Br_-^n$$

with

$$r_+ = \exp(\hat{h}) + \mathcal{O}(\hat{h}^3) \quad \text{and} \quad r_- = -\exp(-\hat{h}) + \mathcal{O}(\hat{h}^3)$$

(Taylor expansion).

- Powers of both roots at $t^* = nh$ read

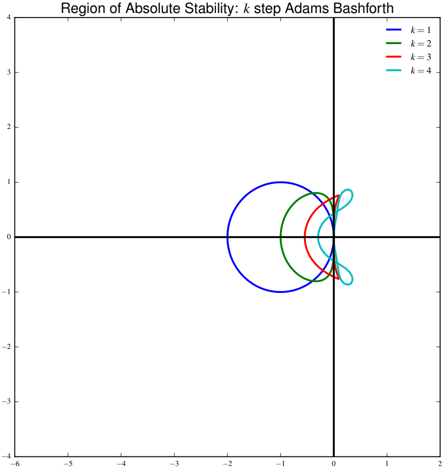
$$r_+^n = \exp(\lambda t^*) + \mathcal{O}(h^2) \quad \text{and} \quad r_-^n = (-1)^n \exp(-\lambda t^*) + \mathcal{O}(h^2)$$

- The first root is the approximation of the exact solution; the second is a *spurious root*: It is a purely numerical artefact.
- The spurious root leads to a term

$$Br_-^n = -\frac{1}{12}h^3(-1)^n \exp(-\lambda t^*) + \mathcal{O}(h^4) \tag{121}$$

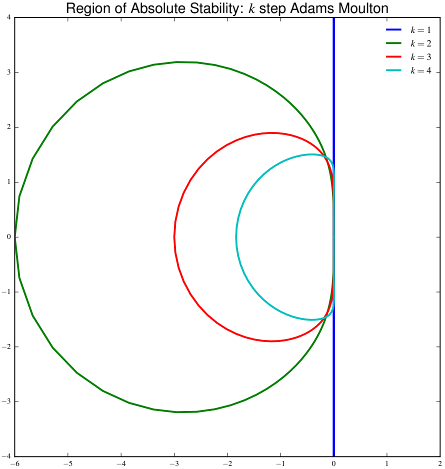
which causes the late-time oscillations.

- Adams-Bashforth method, $x_{n+k} - x_{n+k-1} = h \sum_{j=0}^{k-1} \beta_j f_{n+j}$ with $\beta_k = 0$



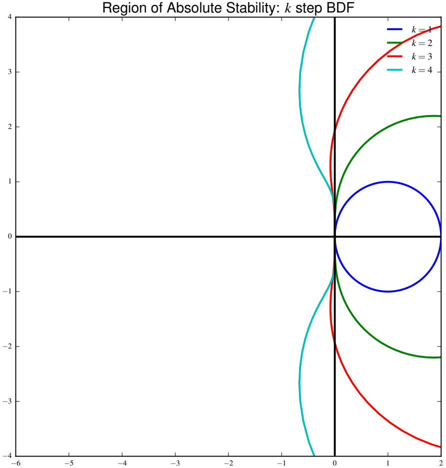
- What happens when k increases?

- Adams-Moulton method, $x_{n+k} - x_{n+k-1} = h \sum_{j=0}^{k-1} \beta_j f_{n+j}$ with $\beta_k \neq 0$



- What is the difference between explicit and implicit Adams methods?

- BDF method, $\sum_{j=0}^{k} \alpha_j x_{n+j} = h\beta_k f_{n+k}$



A-stability

Definition

A numerical method is said to be *A-stable* if its region of absolute stability $\mathcal{R}$ includes the entire left complex half plane $\{z \in \mathbb{C} : \mathrm{Real}(z) < 0\}$.

Definition

A numerical method is said to be *A_0-stable* if its interval of absolute stability $\mathcal{R}_0$ includes the entire negative real axis.

Theorem

Dahlquist's Second Barrier Theorem

1. *There is no A-stable explicit LMM*
2. *An A-stable (implicit) LMM cannot have order $p > 2$*
3. *The order-two A-stable LMM with scaled error constant $C_{p+1}/\sigma(1)$ of smallest magnitude is the trapezoidal rule.*

Announcements

Announcements:

- Exam on Wednesday, October 12.
- On Monday, October 10. Discussion of the assignments (TA).

Example

- An example for a system of two ODEs is

$$u'(t) = -tu(t)v(t) \tag{122}$$
$$v'(t) = -u(t)^2 \tag{123}$$

  with initial values $u(0) = 1$ and $v(0) = 2$ for example.
- Can collect $\mathbf{u}(t) = (u(t), v(t)) \in \mathbb{R}^2$ and write as a vector-valued ODE

$$\mathbf{u}'(t) = \mathbf{F}(\mathbf{u}, t) \tag{124}$$

  with $\mathbf{u}(0) = (1, 2)$ and

$$\mathbf{F}(\mathbf{u}, t) = \mathbf{F}\left(\begin{pmatrix} u \\ v \end{pmatrix}, t\right) = \begin{pmatrix} -tu(t)v(t) \\ -u(t)^2 \end{pmatrix} \tag{125}$$

Diagonalization of linear systems of ODEs

- For $A \in \mathbb{R}^{m,m}$, $\mathbf{u}(t) \in \mathbb{R}^m$,

$$\mathbf{u}'(t) = A\mathbf{u}(t) \tag{126}$$

  is a linear system of m ODEs. Here, the right hand side function is given by

$$\mathbf{F}(\mathbf{u}, t) = A\mathbf{u}(t). \tag{127}$$

- Now assume that A is diagonalizable and has m linearly independent eigenvectors $\mathbf{v}_j$ with corresponding eigenvalues λ_j, i.e.

$$A\mathbf{v}_j = \lambda_j \mathbf{v}_j \tag{128}$$

- Then (remember linear algebra...) there exists a possibly complex decomposition of A such that

$$V^{-1}AV = \Lambda \tag{129}$$

  where Λ is a diagonal matrix with entries λ_j on the diagonal and

$$V = \left[\mathbf{v}_1 \mathbf{v}_2 \ldots \mathbf{v}_m\right] \in \mathbb{C}^{m,m} \tag{130}$$

Diagonalization of linear systems of ODEs

- Given a linear ODE system
$$\mathbf{u}'(t) = A\mathbf{u}(t) \tag{131}$$
with a decomposition $A = V\Lambda V^{-1}$, that is $V^{-1}AV = \Lambda$.
- Define $\mathbf{x}(t) := V^{-1}\mathbf{u}(t)$. Then, $\mathbf{x}(t)$ solves to system of ODEs
$$\mathbf{x}'(t) = \Lambda\mathbf{x}(t) \tag{132}$$
which consists of m independent scalar ODEs
$$x_j'(t) = \lambda_j x_j(t) \tag{133}$$
and solution $x_j(t) = x_j(0)\exp(\lambda_j t)$.
- The eigenvalues of A determine the solution!

Example

- Linear system of ODEs $\mathbf{u}'(t) = A\mathbf{u}(t)$ with

$$A = \begin{bmatrix} 1 & 3 \\ -2 & -4 \end{bmatrix} \tag{134}$$

- The eigenvalue decomposition of A is

$$V^{-1}AV = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \tag{135}$$

with

$$V = \begin{bmatrix} 3 & -1 \\ -2 & 1 \end{bmatrix} \tag{136}$$

- Check:

$$\begin{bmatrix} 1 & 3 \\ -2 & -4 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} = \begin{bmatrix} -3 \\ 2 \end{bmatrix} = (-1) \begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{137}$$

and

$$\begin{bmatrix} 1 & 3 \\ -2 & -4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} = (-2) \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{138}$$

Example

- The transformed variable now solves

$$\mathbf{x}'(t) = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \mathbf{x}(t) \tag{139}$$

and thus

$$\mathbf{x}(t) = \begin{bmatrix} A\exp(-t) \\ B\exp(-2t) \end{bmatrix} \tag{140}$$

- Now transform back

$$\mathbf{u}(t) = V\mathbf{x}(t) = \begin{bmatrix} 3 & -1 \\ -2 & 1 \end{bmatrix} \mathbf{x} = A\exp(-t) \begin{bmatrix} 3 \\ -2 \end{bmatrix} + B\exp(-2t) \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{141}$$

Note how long-term behavior is solely governed by the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -2$.

- Obviously, we have $\mathbf{u}(t) \to 0$ as $t \to \infty$: Can therefore generalize absolute stability to linear systems of ODEs.

Theorem

Theorem

If A is a diagonalizable matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$, then the solutions of $\mathbf{u}'(t) = A\mathbf{u}(t)$ tend to zero as $t \to \infty$ for all choices of initial values if, and only if, $Real(\lambda_j) < 0$ for each $j = 1, \ldots, m$.

Diagonalizing LMMs

- Applying the general two-step LMM

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\left(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\right) \tag{142}$$

to the linear system $\mathbf{u}'(t) = A\mathbf{u}(t)$ gives

$$\mathbf{u}_{n+2} + \alpha_1 \mathbf{u}_{n+1} + \alpha_0 \mathbf{u}_n = hA\left(\beta_2 \mathbf{u}_{n+2} + \beta_1 \mathbf{u}_{n+1} + \beta_0 \mathbf{u}_n\right) \tag{143}$$

- Now set $\mathbf{u}_{n+j} = V^{-1}\mathbf{x}_{n+j}$ for $j = 0, 1, 2$ and multiply (143) with V^{-1} to get

$$\begin{aligned}
&V^{-1}\mathbf{u}_{n+2} + \alpha_1 V^{-1}\mathbf{u}_{n+1} + \alpha_0 V^{-1}\mathbf{u}_n && (144)\\
&\quad = hV^{-1}A\left(\beta_2 \mathbf{u}_{n+2} + \beta_1 \mathbf{u}_{n+1} + \beta_0 \mathbf{u}_n\right) && (145)\\
&\quad = hV^{-1}AV\left(\beta_2 V^{-1}\mathbf{u}_{n+2} + \beta_1 V^{-1}\mathbf{u}_{n+1} + \beta_0 V^{-1}\mathbf{u}_n\right) && (146)
\end{aligned}$$

which simplifies to

$$\mathbf{x}_{n+2} + \alpha_1 \mathbf{x}_{n+1} + \alpha_0 \mathbf{x}_n = h\Lambda\left(\beta_2 \mathbf{x}_{n+2} + \beta_1 \mathbf{x}_{n+1} + \beta_0 \mathbf{x}_n\right) \tag{147}$$

- Hence: Diagonalizing the LMM is equivalent to applying the LMM to the diagonalized system!

Absolute stability for systems

Definition

A LMM is absolutely stable for the diagonalizable system $\mathbf{u}'(t) = A\mathbf{u}(t)$ if $h\lambda_j \in \mathcal{R}$ (the region of absolute stability) for every eigenvalue λ_j of A.

- The definition ensures that if $\mathbf{u}(t) \to 0$ for every choice of initial conditions and the LMM is absolute stable for a given h, the approximate solution $\mathbf{u}_0, \mathbf{u}_1, \ldots$ also tends to zero.

Example

- Consider forward Euler applied to the system

$$u'(t) = -11u(t) + 100v(t) \tag{148}$$
$$v'(t) = u(t) - 11v(t). \tag{149}$$

- In matrix form, the system reads

$$\begin{bmatrix} u(t) \\ v(t) \end{bmatrix}' = \begin{bmatrix} -11 & 100 \\ 1 & -11 \end{bmatrix} \begin{bmatrix} u(t) \\ v(t) \end{bmatrix} \tag{150}$$

  with eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -21$. Both are real, so for forward Euler to be absolutely stable, $h\lambda_j$ must be in $\mathcal{R}_0 = (-2, 0)$ for $j = 1, 2$.

- For the method to be absolutely stable, the time-step h must satisfy

$$-2 < -h < 0 \quad \text{and} \quad -2 < -21h < 0 \tag{151}$$

  i.e. $0 < h < \frac{2}{21} \approx 0.0952$.

- Note: The most rapidly decaying component sets the maximum allowed time-step!

Stiff systems

- Consider a system where the ratio

$$\frac{\max_{j=1,\dots,M} Real(\lambda_j)}{\min_{j=1,\dots,M} Real(\lambda_j)} \tag{152}$$

  is very large.

- The maximum will require a very small time-step h for the method to be stable.

- If (most) other components have much smaller eigenvalues, these are way over-resolved: We will end up with a solution that is much more accurate than is probably needed.

- Such systems are typically called *stiff*.

- A-stable methods are particularly important for this kind of problems: They allow to choose h solely on grounds of accuracy.

Introduction

- For a k-step LMM, $\mathbf{x}_{n+k}$ is computed from

$$\mathbf{x}_{n+k} + \alpha_{k-1}\mathbf{x}_{n+k-1} + \ldots + \alpha_0\mathbf{x}_n = h\left(\beta_k\mathbf{f}_{n+k} + \ldots \beta_0\mathbf{f}_n\right) \tag{153}$$

  with $\mathbf{f}_{n+k} = \mathbf{f}\left(t_{n+k}, \mathbf{x}_{n+k}\right)$.

- Collect all terms without $\mathbf{x}_{n+k}$ which are known from previous time steps

$$\mathbf{g}_n := h\left(\beta_{k-1}\mathbf{f}_{n+k-1} + \ldots \beta_0\mathbf{f}_n\right) - \alpha_{k-1}\mathbf{x}_{n+k-1} - \ldots - \alpha_0\mathbf{x}_n. \tag{154}$$

  Then, $\mathbf{x}_{n+k}$ is the solution $\mathbf{u}$ of the nonlinear equation

$$\mathbf{u} = h\beta_k\mathbf{f}(t_{n+k}, \mathbf{u}) + \mathbf{g}_n. \tag{155}$$

- For $h = 0$ or $\beta_k = 0$ (explicit LMM), there is always the solution $\mathbf{u} = \mathbf{g}_n$.
- Otherwise, since $\mathbf{f}$ is in general nonlinear, the equation may have one, no or multiple solutions.

Fixed point iteration

- Note that

$$\mathbf{u} = h\beta_k \mathbf{f}(t_{n+k}, \mathbf{u}) + \mathbf{g}_n. \tag{156}$$

is also a fixed point equation, i.e. find $\mathbf{u}$ such that

$$\mathbf{u} = \Phi(\mathbf{u}) \tag{157}$$

with $\Phi(\mathbf{u}) := h\beta_k \mathbf{f}(t_{n+k}, \mathbf{u}) + \mathbf{g}_n$.

- Can try to solve with fixed point or Picard iteration

$$\mathbf{u}^{[l+1]} = \Phi(\mathbf{u}^{[l]}) = h\beta_k \mathbf{f}(t_{n+k}, \mathbf{u}^{[l]}) + \mathbf{g}_n \tag{158}$$

- Important: How can we choose the starting value $\mathbf{u}^{[0]}$? Use e.g. $\mathbf{x}_{n+k-1}$ or some form of predictor.

Fixed point iteration

- When does Picard iteration converge?
- Suppose

$$\mathbf{u}^{[l]} = \mathbf{x}_{n+k} + \mathbf{E}^{[l]}. \tag{159}$$

Using a vector-valued Taylor expansion, we find

$$\mathbf{f}(t_{n+k}, \mathbf{u}^{[l]}) = \mathbf{f}(t_{n+k}, \mathbf{x}_{n+k} + \mathbf{E}^{[l]}) \approx \mathbf{f}(t_{n+k}, \mathbf{x}_{n+k}) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(t_{n+k}, \mathbf{x}_{n+k})\mathbf{E}^{[l]} \tag{160}$$

- This gives for the error $\mathbf{E}^{[l]}$ the iteration

$$\mathbf{E}^{[l+1]} \approx h\beta_k B \mathbf{E}^{[l]} \tag{161}$$

with

$$B = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(t_{n+k}, \mathbf{x}_{n+k}) \tag{162}$$

the Jacobi matrix at $(t_{n+k}, \mathbf{x}_{n+k})$.

Fixed point iteration

- Now let λ_B be an eigenvalue of B with eigenvector $\mathbf{v}$. Then, assuming equality and that $\mathbf{E}^{[0]} = \mathbf{v}$,

$$\mathbf{E}^{[1]} = h\beta_k B\mathbf{v} = (h\beta_k\lambda_B)\,\mathbf{v} \tag{163}$$

  so that

$$\mathbf{E}^{[l]} = (h\beta_k\lambda_B)^l\,\mathbf{v} \tag{164}$$

  Note: Can be generalized by using a projection of $\mathbf{E}^{[0]}$ to the eigenbasis.

- It follows that $\mathbf{E}^{[l]}$ cannot go to zero as $l \to \infty$ unless

$$h\,|\beta_k\lambda_B| < 1 \tag{165}$$

  for all eigenvalues of B.

- Note that this gives a restriction on h not dissimilar to that required for absolute stability! Particularly for stiff problems, fixed point iteration can therefore be expected to give bad results.