20. Lecture 20

Ordinary Differential Equations (ODEs)

Look for $x(t) \in C^1(a,b)$ such that

$$\left\{ \; x'(t) = f(t, x(t)), \qquad t \in (a,b); x(t_0) = x_0 \right.$$

for some $t_0 \in (a,b)$, $x_0 \in \mathbb{R}$ and a function $f : (a,b) \times \mathbb{R} \to \mathbb{R}$.

We ask the following questions:

(1) Is there a function $x(t)$ satisfying the ODE and is it unique?
(2) How to approximate solutions?

Example 20.1 (Non-Existence)**.** *Consider the following ODE*

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

We check that $x(t) = \sec(t) = \frac{1}{\cos(t)}$ *is a solution. Indeed, we check*

$$x' = \frac{\sin(t)}{\cos^2(t)} = \frac{1}{\cos(t)}\tan(t) = x\tan(t)$$

and $x(0) = \frac{1}{\cos(0)} = 1$. *The function* $x(t) = \frac{1}{\cos(t)}$ *is pictured in Figure 9. This solution is only continuously defined on* $(-\pi/2, \pi/2)$ *and is unique in that interval. The blow up of the solution at* $t = \pm\pi/2$ *is caused by the blow up of* $\tan(t)$ *at* $t = \pm\pi/2$ *(see later).*

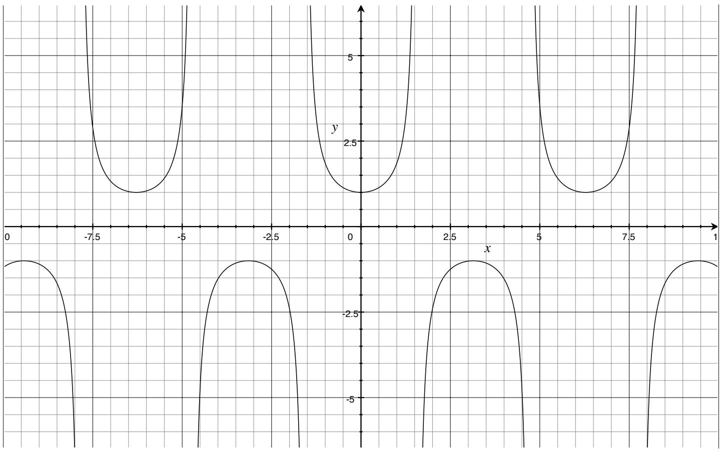

Figure 9. Function $x(t) = 1/\cos(t)$.

Example 20.2 (Non-Existence 2)**.** *Consider now the ODE*

$$x' = 1 + x^2, \qquad x(0) = 0.$$

In this case we find the solution separating the variables and taking anti-derivatives

$$\int \frac{dx}{1+x^2} = \int 1 \; dt + C.$$

This implies that

$$\arctan(x) = t + C$$

or

$$x = tan(t + C).$$

Using the initial condition, we find that $C = 0$, i.e.

$$x(t) = \tan(t).$$

Figure 10 depicts this function. Although the ODE looks harmless, the solution $x(t) = \tan(t)$ is continuously defined (and unique) on $(-\pi/2, \pi/2)$ and blows up at the endpoints.

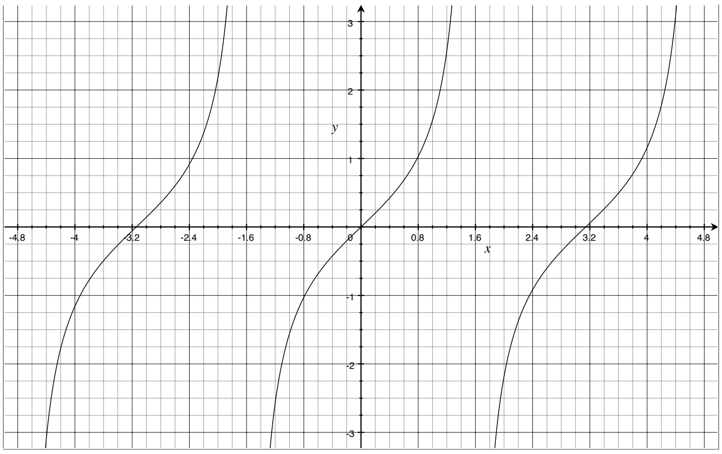

FIGURE 10. Function $x(t) = \tan(t)$.

Example 20.3 (Non-Uniqueness). *Consider now the ODE*

$$x' = x^{2/3}, \qquad x(0) = 0.$$

Clearly $x(t) = 0$ is a solution. Alternatively,

$$\int \frac{dx}{x^{2/3}} = \int 1 \; dt + C.$$

This implies that

$$3x^{1/3} = t + C$$

or

$$x(t) = \left(\frac{t + C}{3}\right)^3.$$

Using the initial condition, we find that $C = 0$, i.e.

$$x(t) = \left(\frac{t}{3}\right)^3.$$

We check that

$$x'(t) = \frac{3}{3}(t/3)^2 = \left((t/3)^6\right)^{1/3} = (x^2)^{1/3}.$$

Hence, we found 2 solutions (non-unique). There are, in fact, an infinite amount of solutions.

Systems of ODEs are treated similarly. We are given $F(t,x):(a,b)\times\mathbb{R}^n\to\mathbb{R}^n$,

$$x(t)=\begin{pmatrix}x_1(t)\\x_2(t)\\\vdots\\x_n(t)\end{pmatrix},\qquad x'(t)=\begin{pmatrix}x_1'(t)\\x_2'(t)\\\vdots\\x_n'(t)\end{pmatrix}.$$

We look for solutions $x_i(t)\in C^1(a,b)$ for $i=1,...,n$ satisfying

$$\begin{cases}x'(t)=F(t,x), & t\in(a,b)\\ x(0)=x_0,\end{cases}\tag{11}$$

where $x_0\in\mathbb{R}^n$ and $t_0\in(a,b)$ are given.

Definition 20.1 (Uniform Lipschitz)**.** *Assume that F is continuous on $[a,b]\times\mathbb{R}^n$. Then F satisfies a uniform Lipschitz condition if*

$$|F(t,x_1)-F(t,x_2)|\le L|x_1-x_1|$$

for all $t\in[a,b]$ and $x_1,x_2\in\mathbb{R}^n$.

Theorem 20.1 (Existence and Uniqueness)**.** *If F satisfies a uniform Lipschitz condition, the system of ODES* (11) *has a unique solution on $[a,b]$.*

Proof. (Sketch using Picard Iteration). We start by noting that integrating the differential equation gives

$$x(t)=x_0\int_{t_0}^t x'(s)\;ds=\int_{t_0}^t F(s,x(s))\;ds,$$

i.e.

$$x(t)=x_0+\int_{t_0}^t F(s,x(s))\;ds.\tag{12}$$

This is a fixed point equation for the vector valued function $x(t)$ on $B_\delta(t_0)=(t_0-\delta,t_0+\delta)$ for some $\delta>0$. Hence, we define a sequence of functions $x^j:B_\delta(t_0)\to\mathbb{R}^n$ by

$$x^0(t)=x_0,\qquad x^{j+1}(t)=x_0+\int_{t_0}^t F(s,x^j(s))\;ds.$$

Now, the proof follows the contraction mapping theorem (Theorem 18.1). For a vector valued function $u(t)$ defined on $B_\delta(t_0)$, we set

$$\|u\|_\infty=\max_{t\in B_\delta(t_0)}\|u(t)\|_\infty.$$

As

$$x^{j+1}(t)=x_0+\int_{t_0}^t F(s,x^j(s))\;ds,$$

we have that

$$\begin{aligned}\|x^{j+1}(t)-x^j(t)\|_\infty&=\|\int_{t_0}^t(F(s,x^j(s))-F(s,x^{j-1}(s)))\;ds\|_\infty\\&\le|t-t_0|\max_{s\text{ between }t_0\text{ and }t}\|F(s,x^j(s))-F(s,x^{j-1}(s))\|_\infty\\&\le\delta L\max_{s\text{ between }t_0\text{ and }t}\|x^j(s)-x^{j-1}(s)\|_\infty.\end{aligned}$$

Thus, $\|x^{j+1} - x^j\|_\infty \leq \delta L \|x^j - x^{j-1}\|_\infty$. This sequence converges in $\|.\|_\infty$ if $\rho := \delta L < 1$, which is guaranteed upon choosing δ sufficiently small. We denote by $x^* := \lim_{j\to\infty} x^j$. One then shows that each x^j is continuous and $\|x^j\|_\infty \leq C\|x_0\|_\infty$. As x^j converges uniformly on $B_\delta(t_0)$, its limit x^* is continuous and is the unique fixed point of (12). Differentiating the fixed point equation (12) shows that

$$(x^*)' = F(t, x^*).$$

□

21. Lecture 21

Numerical Ordinary Differential Equations (ODEs)

We discuss two techniques to approximate the ODE

$$\frac{d}{dt}x(t) = f(t, x(t)).$$

(1) Replace $\frac{d}{dt}x(t)$ by a finite difference approximation.
(2) Integrate:

$$x(t_{k+1}) - x(t_k) = \int_{t_k}^{t_{k+1}} \frac{d}{dt}x(t)\ dt = \int_{t_k}^{t_{k+1}} f(t, x(t))\ dt$$

or

$$x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} f(t, x(t))\ dt$$

and replace the integral by a quadrature.

Finite Difference Approximations. We saw that for $t \in [t_k, t_{k+1}]$

$$\frac{d}{dt}x(t) \approx \frac{x_{t_k+1} - x_{t_k}}{h_k},$$

where $h_k = t_{k+1} - t_k$. Using this in the ODE at $t = t_k$ yields the *Forward Euler* or forward difference relation

$$\frac{x(t_{k+1}) - x(t_k)}{h_k} = f(t_k, x(t_k)) + O(h_k).$$

This leads to the ODE scheme

$$\frac{x_{k+1} - x_k}{h_k} = f(t_k, x(t_k)),$$

where $x_k \approx x(t_k)$.

Similarly, but using $t = t_{k+1}$ in the ODE, yields the *Backward Euler* or backward difference relation

$$\frac{x(t_{k+1}) - x(t_k)}{h_k} = f(t_{k+1}, x(t_{k+1})) + O(h_k).$$

In turn, the ODE scheme becomes

$$\frac{x_{k+1} - x_k}{h_k} = f(t_{k+1}, x(t_{k+1})),$$

where again $x_k \approx x(t_k)$.

Remark 21.1 (Explicit Schemes). Given x_0, the Forward Euler determines x_k iteratively according to the relation

$$x_{k+1} = x_k + h_k f(t_k, x_k).$$

Given x_k, the computation of the right side of the above relation only involves addition, multiplication and the evaluation of $f(t_k, x_k)$. This is called an explicit method.

Remark 21.2 (Implicit Schemes). Given x_0, the Backward Euler determines x_k iteratively according to the relation

$$x_{k+1} - h_k f(t_k, x_{k+1}) = x_k.$$

Given x_k, the problem of determining x_{k+1} is generally a nonlinear. It can be written as a fixed point iteration

$$x = x_k + h_k f(t_{k+1}, x)$$

and one could use Picardo or Newton's method. ODE schemes that require the solution of (nonlinear) equations are called implicit. For certain types of problems, implicit methods have better stability properties.

Remark 21.3 (Systems of ODEs). We always derive ODE methods for a single ODE. As we shall see, schemes for systems of ODEs follow immediately from schemes for the single variable ODE.

ODE schemes can also be derived using a "Taylor series technique" as illustrated in the following examples.

Example 21.1 (Forward Euler)**.** *Set $h_j = t_{j+1} - t_j$ so that*

$$x(t_{j+1}) = x(t_j) + h_j x'(t_j) + O(h_j^2).$$

Replace $x(t_j)$ by an approximation x_j and throw away the $O(h_j^2)$ term:

$$x_{j+1} = x_j + h_j x'(t_j) = x_j + h_j f(t_j, x_j).$$

This is the forward Euler again.

Example 21.2 (Second Order)**.**

$$x(t_{j+1}) = x(t_j) + h_j x'(t_j) + \frac{h_j^2}{2} x''(t_j) + O(h_j^3).$$

As before,

$$x'(t_j) = f(t_j, x_j)$$

but also

$$x''(t) = \frac{d}{dt} x'(t) = \frac{d}{dt} f(t, x(t)) = \frac{\partial}{\partial t} f(t, x(t)) + \frac{\partial}{\partial y} f(t, x(t)) x'(t).$$

Throw away the $O(h_j^3)$ term and replace $x(t_j)$ by x_j to arrive at

$$x_{j+1} = x_j + h_j f(t_j, x_j) + \frac{h_j^2}{2} \left(f_t(t_j, x_j) + f_y(t_j, x_j) f(t_j, x_j) \right).$$

Example 21.3 (Third Order)**.**

$$x(t_{j+1}) = x(t_j) + h_j x'(t_j) + \frac{h_j^2}{2} x''(t_j) + \frac{h_j^3}{6} + O(h_j^4).$$

Here, we also use

$$x'''(t) = \frac{d}{dt} x''(t) = \frac{d}{dt}(f_t + f_y f) = f_{tt} + f_{ty} f + f(f_{yt} + f_{yy} f) + f_y(f_t + f_y f)$$

to deduce the scheme

$$x_{j+1} = x_j + h_j f(t_j, x_j) + \frac{h_j^2}{2} \left(f_t + f_y f \right) + \frac{h_j^3}{6} \left(f_{tt} + 2 f_{ty} f + f^2 f_{yy} + f_y f_t + f_y^2 f \right).$$

All the f's and derivatives are evaluated at (t_j, x_j).

The methods of Examples 21.2 and 21.3 are not used in practice. The reason is that you can derive higher order methods which only require the users to provide a routine for $f(t, x)$ (not its derivatives).

However, we shall use these methods to derive higher order methods only involving $f(t, x)$ for various values of t and x.

22. Lecture 22

Runge-Kutta Methods. These methods are based on Taylor series method.

Example 22.1 (Second Order). *From Example 21.2:*
(13)

$$x(t_{k+1}) = x(t_k)+h_k f(t_k, x(t_k))+\frac{h_k^2}{2}\left(f_t(t_k, x(t_k)) + f_y(t_k, x(t_k))f(t_k, x(t_k))\right)+O(h_k^3).$$

We look for a method of the form

$$x(t_{k+1}) = x(t_k)+\omega_1 h_k f(t_k, x(t_k))+\omega_2 h_k f\left(t_k + \alpha h_k, x(t_k) + \alpha h_k f(t_k, x(t_k))\right)+O(h_k^3)$$

for some parameters ω_1, ω_2, α. In what follows, f, f_t, f_y and $\nabla f = (f_t, f_y)^t$ are evaluated at $(t_k, x(t_k))$. By Taylor theorem

$$\begin{aligned} f\left(t_k + \alpha h_k, x(t_k) + \alpha h_k f(t_k, x(t_k))\right) &= f + \nabla f \cdot (\alpha h_k, \alpha h_k f) + O(h_k^2) \\ &= f + \alpha h_k(f_t + f_y f) + O(h_k^2). \end{aligned}$$

Putting it together gives

$$x(t_{k+1}) = x(t_k) + h_k(\omega_1 + \omega_2)f + \alpha h_k^2 \omega_2(f_t + f_y f) + O(h_k^3).$$

Comparing this with (13)*, we see that we need to set*

$$\omega_1 + \omega_2 = 1, \qquad \omega_2 \alpha = \frac{1}{2}$$

(3 unknowns, 2 equations, multiple solutions). The numerical method reads

$$x_{k+1} = x_k + \omega_1 h_k f(t_k, x_k) + \omega_2 h_k f\left(t_k + \alpha h_k, x_k + \alpha h_k f(t_k, x_k)\right).$$

Example 22.2 (Heuns Method). *We set $\omega_1 = \omega_2 = \frac{1}{2}$ and $\alpha = 1$ to arrive at*

$$x_{k+1} = x_k + \frac{1}{2}h_k(F_1 + F_2),$$

where

$$\begin{aligned} F_1 &= f(t_k, x_k) \\ F_2 &= f(t_k + h_k, x_k + h_k F_1). \end{aligned}$$

Example 22.3 (Modified Euler). *We set $\omega_1 = 0$, $\omega_2 = 1$ and $\alpha = \frac{1}{2}$ to arrive at*

$$x_{k+1} = x_k + h_k F_2,$$

where

$$\begin{aligned} F_1 &= f(t_k, x_k) \\ F_2 &= f(t_k + \frac{1}{2}h_k, x_k + \frac{1}{2}h_k F_1). \end{aligned}$$

Remark 22.1 (Heun's method). Apply Heun's method to

$$x'(t) = f(t)$$

to get

$$x(t_{k+1}) = x(t_k) + \frac{h_k}{2}\left(f(t_k) + f(t_k + h_k)\right) + O(h_k^3).$$

Note that

$$\int_{t_k}^{t_{k+1}} x'(t)\ dt = x(t_{k+1}) - x(t_k) = \frac{h_k}{2}\left(f(t_k) + f(t_k + h_k)\right) + O(h_k^3).$$

Thus

$$\int_{t_k}^{t_{k+1}} f(t)\ dt = \int_{t_k}^{t_{k+1}} x'\ dt \approx \frac{h_k}{2}\left(f(t_k) + f(t_k + h_k)\right),$$

which is the Trapezoidal rule. The latter is locally 3rd order and globally 2nd order. Thus it is natural to define the order of the Heun's method to be 2.

Exercise 22.1 (Backward and Forward Euler)**.** *Use a similar reasoning to show that Backward and Forward Euler methods are first order.*

Remark 22.2 (Modified Euler). The modified Euler method applied to

$$x'(t) = f(t)$$

gives

$$x(t_{k+1}) = x(t_k) + h_k f(t_k + h_k/2).$$

This is the midpoint quadrature which is also globally second order, so the (global) order of the modified Euler is 2.

Heun's methods and modified Euler methods are called (explicit) Runge-Kutta methods. High order Runge-Kutta methods are tedious to derive but we have been extensively studied. One such method, which is widely used, is the Runge-Kutta 4th order method:

$$x_{k+1} = x_k + \frac{h_k}{6}\left(F_1 + 2F_2 + 2F_3 + F_4\right),$$

where

$$\begin{aligned} F_1 &= f(t_k, x_k), \\ F_2 &= f(t_k + \frac{h_k}{2}, x_k + \frac{h_k}{2}F_1), \\ F_3 &= f(t_k + \frac{h_k}{2}, x_k + \frac{h_k}{2}F_2), \\ F_4 &= f(t_k + h_k, x_k + h_k F_3). \end{aligned}$$

This method is 4th order.

Remark 22.3 (Simpson). Applying the Runge-Kutta 4th order method to

$$x'(t) = f(t)$$

gives

$$\begin{aligned} \int_{t_k}^{t_{k+1}} f(t)dt = \int_{t_k}^{t_{k+1}} x'(t)\ dt &\approx x_{k+1} - x_k \\ &= \frac{h_k}{6}\left(f(t_k) + 4f(t_k + h_k/2) + f(t_k + h_k)\right). \end{aligned}$$

This is Simpson's Rule, which is locally 5th order, globally 4th order.

23. Lecture 23

23.1. **Multistep Methods.** We start with the integral representation

$$x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} f(t, x(t))\ dt$$

and apply quadrature to the integral. Note that one only has approximation for $x(t)$ at t_j, $j = 0, ..., k$ so we have to use $t_0, ..., t_k$ and possibly t_{k+1} as quadrature nodes.

The *Adams Bashford* schemes use the $m+1$ nodes $t_k, ..., t_{k-m}$:

$$x(t_{k+1}) = x(t_k) + \sum_{j=0}^{m} w_j f(t_{k-j}, x(t_{k-j})) + O(h^{m+2})$$

when the weights w_j are taken so that

$$I(g) := \int_{t_k}^{t_{k+1}} g(t)\ dt = \sum_{j=0}^{m} w_j g(t_{k-j}) = I_m(g),$$

i.e. the quadrature is exact on $\mathbb{P}^m$. Note that Adams Bashford methods

$$x_{k+1} = x_k + \sum_{j=0}^{m} w_j f(t_{k-j}, x_{k-j})$$

are explicit and globally of order $m+1$.

The *Adams Moulton* schemes use instead the $m+1$ nodes $t_{k+1}, ..., t_{k-m+1}$:

$$x(t_{k+1}) = x(t_k) + \sum_{j=-1}^{m-1} w_j f(t_{k-j}, x(t_{k-j})) + O(h^{m+2}).$$

Adams Moulton methods

$$x_{k+1} = x_k + \sum_{j=-1}^{m-1} w_j f(t_{k-j}, x_{k-j})$$

are implicit since $f(t_{k+1}, x_{k+1})$ appears and globally of order $m+1$.

In general, the ODE schemes are

$$x_{k+1} = x_k + \sum_{j=\sigma}^{m+\sigma} w_j f(t_{k-j}, x_{k-j})$$

with $\sigma = 0$ for Adams Bashford and $\sigma = -1$ for Adams Moulton.

Remark 23.1 (Non Uniform Time-steps). If one only has

$$t_0 < t_1 < t_2 < ...$$

(i.e. no structure), then new quadrature weights need to be computed at each time step. Also, the scheme cannot be analyzed.

Remark 23.2 (Starting Values). When $\sigma = 0$, one needs the starting values

$$x_0, x_1, ..., x_m$$

to compute $x_{m+1}, x_{m+2}, ...$. The values for $x_0, ..., x_m$ need to be computed by some other method.

When $\sigma = -1$, one needs the starting values

$$x_0, x_1, ..., x_{m-1}$$

to compute $x_m, x_{m+1}, ...$ The values for $x_1, ..., x_{m-1}$ need to be computed by some other method.

Remark 23.3 (More General). The methods can be made still more general, e.g.

$$\sum_{i=0}^{m} \alpha_i x_{j-i} = \sum_{i=0}^{m} \beta_i f(t_{j-i}, x_{j-i})$$

with $\alpha_0 = 1$. In this case, the method is explicit if $\beta_0 = 0$, otherwise, it is implicit.

From here onward, we assume *a uniform time step*, i.e.

$$t_j = t_0 + jh$$

for some fixed $h > 0$.

We return to Adams Bashford and compute the weights for the quadrature problem

$$\widehat{I}(g) := \int_m^{m+1} g(x)\ dx \approx \sum_{j=0}^{m} \widehat{w}_j g(m-j) =: \widehat{I}_m(g)$$

so that $\widehat{I}_m$ is exact on $\mathbb{P}^m$. Now for $t_j = t_0 + jh$, we get the quadrature

$$I(g) := \int_{t_k}^{t_{k+1}} g(t)\ dt \approx h \sum_{j=0}^{m} \widehat{w}_j g(t_{k-j}) = I_m(g)$$

by translating $\widehat{I}_m(g)$.

Example 23.1 (3rd order). *We set $m = 2$:*

$$\widehat{I}(g) = \int_2^3 g(x)\ dx \approx \sum_{j=0}^{2} \widehat{w}_j g(2-j) =: \widehat{I}_2(g).$$

To make it exact for $\mathbb{P}^2$, we require

$$\begin{aligned}
\widehat{I}(1) &= \int_2^3 1\ dx = 1 = \widehat{w}_0 + \widehat{w}_1 + \widehat{w}_2 = \widehat{I}_2(1), \\
\widehat{I}(x) &= \int_2^3 x\ dx = \left.\frac{x^2}{2}\right|_2^3 = \frac{5}{2} = 2\widehat{w}_0 + \widehat{w}_1 = \widehat{I}_2(x), \\
\widehat{I}(x^2) &= \int_2^3 x^2\ dx = \left.\frac{x^3}{3}\right|_2^3 = \frac{19}{3} = 4\widehat{w}_0 + \widehat{w}_1 = \widehat{I}_2(x^2).
\end{aligned}$$

The last two conditions imply

$$2\widehat{w}_0 = \frac{19}{3} - \frac{5}{2} = \frac{23}{6}$$

and so

$$\widehat{w}_0 = \frac{23}{12}.$$

Using this in the second constraint yields

$$\frac{23}{6} + \widehat{w}_1 = \frac{5}{2}$$

and so

$$\widehat{w}_1 = \frac{5}{2} - \frac{23}{6} = -\frac{8}{6} = -\frac{4}{3}.$$

It remains to use the first constraint to get

$$\widehat{w}_2 = 1 - \frac{23}{12} + \frac{4}{3} = \frac{12 - 23 + 16}{12} = \frac{5}{12}.$$

The resulting scheme reads

$$x_{k+1} = x_k + \frac{h}{12}\left\{23f(t_k, x_k) - 16f(t_{k-1}, x_{k-1}) + 5f(t_{k-2}, x_{k-2})\right\}.$$

24. Lecture 24

Similarly as for Adams-Bashford, we now derive the Adams-Moulton (for equally spaced nodes, i.e. $x_i = x_0 + hi$), we solve the interpolation problem

$$\widehat{I}(g) := \int_{m-1}^{m} g(s)\ ds \approx \sum_{j=0}^{m} \widehat{w}_j g(m-j) =: \widehat{I}_m(g),$$

which is exact on $\mathbb{P}^m$. The resulting ODE scheme is

$$x_k = x_{k-1} + h \sum_{j=0}^{m} \widehat{w}_j f(t_{k-j}, x_{k-j}).$$

Exercise 24.1 (Undetermined Coefficients)**.** *Compute the coefficients* $\{\widehat{w}_j\}_{j=0}^{2}$ *for the third order method using the undetermined coefficients.*

Backwards Differences. Consider approximating the derivative

$$f'(x) \approx \alpha_0 f(x) + \alpha_1 f(x-h) + ... + \alpha_m f(x-hm).$$

Recall that to find such approximation, we first find the polynomial $p_m \in \mathbb{P}^m$ interpolating

$$p_m(t) = f(t) \tag{14}$$

for $t = x, x-h, ..., x-hm$ and we then set

$$f'(x) \approx p'_m(x).$$

Applying Newton form to solve the interpolation problem (14), we get

$$p_m(t) = \sum_{i=0}^{m} c_i q_i(t),$$

where

$$q_0(t) \equiv 1, \qquad q_j(t) = \prod_{l=0}^{j-1}(t - x + lh).$$

As we shall see in the Appendix at the end of this lecture (the audio refers to the wrong pages),

$$c_i = \frac{1}{i!} D_h^i f(x). \tag{15}$$

Here

$$\begin{aligned} D_h^0 f(x) &= f(x), \\ D_h^1 f(x) &= \frac{f(x) - f(x-h)}{h}, \\ D_h^{j+1} f(x) &= D_h^1[D_h^j f(x)]. \end{aligned}$$

Example 24.1 (D_h^2)**.**

$$D_h^2 f(x) = D_h^1\left[\frac{f(x) - f(x-h)}{h}\right] = \frac{\frac{f(x)-f(x-h)}{h} - \frac{f(x-h)-f(x-2h)}{h}}{h} = \frac{f(x) - 2f(x-h) + f(x-2h)}{h^2}.$$

Example 24.2 (D_h^3)**.**

$$D_h^3 f(x) = \frac{f(x) - 3f(x-h) + 3f(x-2h) - f(x-3)}{h^2}.$$

Notice that the coefficients $1, -3, 3, -1$ *in front of the functions are the binomial coefficients.*

Returning to the Newton's form for p_m, we find that

$$\begin{aligned} p_m(t) = f(x_0) &+ D_h^1 f(x)(t-x) + \frac{1}{2!} D_h^2 f(x)(t-x)(t-x+h) \\ &+ \frac{1}{3!} D_h^3 f(x)(t-x)(t-x+h)(t-x+2h) \\ &+ ... + \frac{1}{m!} D_h^m f(x)(t-x)(t-x+h) \cdot ... \cdot (t-x+(m-1)h). \end{aligned}$$

Differentiating the above expression and taking $t = x$, we arrive at

$$\begin{aligned} p_m'(x) = 0 &+ D_h^1 f(x) + \frac{1}{2!} D_h^2 f(x) h \\ &+ \frac{1}{3!} D_h^3 f(x) h \cdot 2h \\ &+ ... + \frac{1}{m!} D_h^m f(x) h (2h) \cdot ... \cdot (m-1)h, \end{aligned}$$

i.e.

$$p_m'(x) = \sum_{i=1}^{m} \frac{h^{i-1}}{i} D_h^i f(x)$$

and so

$$f'(x) \approx \sum_{i=1}^{m} \frac{h^{i-1}}{i} D_h^i f(x).$$

The corresponding ODE schemes are

$$\sum_{i=1}^{m} \frac{h^{i-1}}{i} D_h^i x_i = f(t_m, x_m).$$

They are all implicit.

Example 24.3 (1st order method)**.**

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}.$$

The corresponding ODE scheme is the Backward Euler scheme

$$\frac{x_j - x_{j-1}}{h} = f(t_j, x_j).$$

Example 24.4 (2nd order method)**.**

$$f'(x) \approx \frac{f(x) - f(x-h)}{h} + \frac{f(x) - 2f(x-h) + f(x=2h)}{2h} = \frac{3/2 f(x) - 2f(x-h) + 1/2 f(x-2h)}{h}.$$

The corresponding ODE scheme is

$$\frac{3/2 x_j - 2x_{j-1} + 1/2 x_{j-2}}{h} = f(t_j, x_j).$$

Example 24.5 (3rd order method)**.**

$$f'(x) \approx \frac{3/2f(x) - 2f(x-h) + 1/2f(x-2h)}{h} + \frac{f(x) - 3f(x-h) + 3f(x-2h) - f(x-3h)}{3h}$$
$$= \frac{1}{6h}\left(11f(x) - 18f(x-h) + 9f(x-2h) - 2f(x-3h)\right).$$

The corresponding ODE scheme is

$$\frac{11x_j - 18x_{j-1} + 9x_{j-2} - 2x_{j-3}}{6h} = f(t_j, x_j).$$

Appendix of Lecture 24. We now return to the justification of (15) and refer to Section 6.2 of the book. In the latter, the coefficients

$$c_i = f[x, x-h, ..., x-hi],$$

the divided difference notation. We want to show that

$$f[x, x-h, ..., x-h(i+i)] = \frac{1}{i!}D^i f(x).$$

To prove this, we proceed by induction. Clearly, we have

$$c_0 = f[x] := f(x).$$

Suppose that

$$f[x, x-h, ..., x-hl] = \frac{1}{l!}D^l f(x).$$

This implies that

$$f[x-h, x-2h, ..., x-h(l+1)] = \frac{1}{l!}D^l f(x-h).$$

We now invoke Theorem 1 of Section 6.2, which guarantees that

$$f[x_0, x_1, ..., x_n] = \frac{f[x_1, ..., x_n] - f[x_0, ..., x_{n-1}]}{x_n - x_0}.$$

Therefore,

$$f[x, x-h, ..., x-(l+1)h] = \frac{f[x-h, ..., x-(l+1)h] - f[x, ..., x-lh]}{x-(l+1)h-x}$$
$$= \frac{D_h^l f(x) - D^l f(x-h)}{h(l+1)}\frac{1}{l!} = \frac{1}{(l+1)!}D_h^{l+1}f(x).$$

This ends the induction proof.

25. Lecture 25

25.1. **Errors in ODE schemes.** Stability and local truncation error are two central notions are when studying numerical methods for ODE schemes. We first provide an intuitive description.

(1) *Local truncation error:* The numerical schemes we have seen we obtained by dropping error terms. The truncation error characterized this error made at each step.
(2) *Stability:* If you make an error computing x_j to approximation $x(t_j)$, that error propagates through all remaining steps (this is unavoidable). A stable scheme prevents this error from increasing.

We will only consider the Adams-Bashford (AB) and Adams-Moulon (AM) schemes for the analysis.

Regarding the *Stability*, the error at any time propagates but do not grow too much. A necessary condition for stability, which we will take as the definition of stability in this class, is the following.

Definition 25.1 (Stability)**.** *We say that a scheme is* stable *if when applied to*

$$x'(t) = 0, \qquad x(t_0) = \varepsilon,$$

the approximate solutions remains bounded at any fixed time independently of the mesh size h.

In the case $t_0 = 0$ and $h = 1/m$, $x_m^h \approx x(1)$ then stability holds at time $t = 1$ if

$$|x_m^h| \leq C \qquad \text{as} \qquad h = 1/m \to 0.$$

Example 25.1. *AB/AM The AB or AM schemes applied to the above ODE read*

$$x_{k+1}^h = x_k^h, \qquad x_0^h = \varepsilon \qquad \Longrightarrow \qquad x_k^h = \varepsilon$$

and are therefore stable.

We now discuss the local truncation error and start with a definition.

Definition 25.2 (Local truncation error)**.** *The* local truncation error *is the error made when one substitutes the solution into the numerical ODE scheme.*

For example,

$$\frac{x(t_{k+1}) - x(t_k)}{h} - \sum_{j=\sigma}^{m+\sigma} \widehat{w}_j f(t_{k-j}, x(t_j - k)) =: \rho_h(k+1)$$

is the local truncation for the numerical scheme

$$\frac{x_{k+1} - x_k}{h} = \sum_{j=\sigma}^{m+\sigma} \widehat{w}_j f(t_{k-j}, x_{j-k}).$$

Note that $\rho_h(k+1) = O(h^{m+1})$ for AB/AM.

The error between $x(t_k)$ and x_k given by the AB scheme is described in the following theorem.

Theorem 25.1 (Error Estimate for Adams Bashford)**.** *Assume that f satisfies a uniform Lipschitz condition, i.e.*

$$|f(t,x) - f(t,y)| \leq M(x-y), \qquad x, y \in \mathbb{R}, \qquad t \in [0,T].$$

Let $h > 0$, $t_k := kh$ and $x_0, x_1, ..$ be the sequence of the $m+1$ points Adams-Brashford approximates. Set $\rho_h := \max_{m<k\leq T/h} |\rho_h(k)|$, $\overline{w} := \max_{j=0,\ldots,m} \widehat{w}_j$ and $\varepsilon_k := \max_{j=0,\ldots,k} |x(t_j) - x_j|$. Then for $m < k \leq T/h$, there holds

$$\varepsilon_k \leq e^{\delta t_k}[\varepsilon_m + \rho_h/\delta],$$

where $\delta := (m+1)\overline{w}M$.

Proof. We start by noting that the Lipschitz assumption on f guarantees a unique solution of the ODE for $t \geq t_0 = 0$. The AB approximate solutions are given by

$$x_{k+1} = x_k + h\sum_{j=0}^{m} \widehat{w}_j f(t_{k-j}, x_{k-j}).$$

Subtracting this from the local truncation relation

$$\frac{x(t_{k+1}) - x(t_k)}{h} - \sum_{j=0}^{m} \widehat{w}_j f(t_{k-j}, x(t_{k-j})) =: \rho_h(k+1)$$

and setting $e_k := x(t_k) - x_k$ gives

$$e_{k+1} = e_k + h\sum_{j=0}^{m} \widehat{w}_j[f(t_{k-j}, x(t_{k-j})) - f(t_{k-j}, x_{k-j})] + h\rho_h(k+1).$$

Now apply the Lipschitz condition and the triangle inequality

$$\begin{aligned}|e_{k+1}| &\leq |e_k| + h\sum_{j=0}^{m} |\widehat{w}_j||f(t_{k-j}, x(t_{k-j})) - f(t_{k-j}, x_{k-j})| + h|\rho_h(k+1)| \\ &|e_k| + hM\sum_{j=0}^{m} |\widehat{w}_j||x(t_{k-j}) - x_{k-j})| + h|\rho_h(k+1)| \\ &|e_k| + hM\overline{w}\sum_{j=0}^{m} |e_{k-j}| + h|.\end{aligned}$$

Using the definition of ρ_h, ε_k and δ, we get

$$|e_{k+1}| \leq |e_k| + \delta h\varepsilon_k + h\rho_h,$$

i.e.

$$\varepsilon_{k+1} \leq (1+\delta h)\varepsilon_k + h\rho_h.$$

Repeating the application of the above estimate gives

$$\begin{aligned}\varepsilon_k &\leq (1+\delta h)\varepsilon_{k-1} + h\rho_h \\ &\leq (1+\delta h)^2\varepsilon_{k-2} + h\rho_h(1 + (1+\delta h)) \\ &\vdots \\ &\leq (1+\delta h)^{k-m}\varepsilon_m + h\rho_h \sum_{j=0}^{k-m-1} (1+\delta h)^j.\end{aligned}$$

Now $k \leq T/h$, so

$$(1+\delta h)^{k-m} \leq (e^{\delta h})^{k-m} \leq e^{\delta kh} \leq e^{\delta T}.$$

Also

$$\sum_{j=0}^{k-m-1}(1+\delta h)^j = \frac{(1+\delta h)^{k-m}-1}{(1+\delta h)-1} \leq \frac{e^{\delta T}}{\delta h}.$$

Thus,

$$\varepsilon_k \leq e^{\delta T}(\varepsilon_m + \rho_h/\delta),$$

which is the desired result. □

Before providing a similar result for the AM scheme, we first show that the scheme is well defined. Recall the AM is implicit and x_k is the solution (if exists) to

$$x_{k+1} = x_k + h\sum_{j=-1}^{m-1}\widehat{w}_j f(t_{k-j}, x_{k-j}).$$

Lemma 25.1 (AM - Well defined)**.** *Assume that f satisfies a uniform Lipschitz condition with constant M. Given $x_0, ..., x_k$. Then the $k+1$ iterate of the $m+1$ points AM is well defined provided $h < \frac{1}{|\widehat{w}_{-1}|M}$.*

Proof. We consider the fixed point iteration for x_{k+1}, namely

$$y_0 = x_k$$

$$y_{l+1} = x_k + h\widehat{w}_{-1}f(t_{k+1}, y_l) + h\sum_{j=0}^{m-1}\widehat{w}_j f(t_{k-j}, x_{k-j}) := G(y_l)$$

i.e. $y_{l+1} = G(y_l)$. Note that

$$G(y) - G(z) - h\widehat{w}_{-1}[f(t_{k+1}, y) - f(t_{k+1}, z)]$$

so that

$$|G(y) - G(z)| \leq h|\widehat{w}_{-1}|M|y - z|.$$

This means that under the assumption $h|\widehat{w}_{-1}|M < 1$, G is a contraction mapping so the sequence y_l converges to a unique fixed point, x_{k+1}. □

26. Lecture 26

In the previous lecture, we provided an error analysis for the Adams-Bashford schemes (see Theorem 25.1) and showed that Adams-Moulton schemes were well defined provided that $h \leq h_0$ (Lemma 25.1). Here h_0 depends on $|\widehat{w}_{-1}|$ and the uniform Lipschitz constant M of f. To analyze Adams-Moulton, we need the following lemma.

Lemma 26.1. *Let $\alpha, \beta, h > 0$ with $\alpha h \leq 1/2$. Then,*

$$(1-\alpha h)^{-1} \leq (1+2\alpha h),$$

$$(1-\alpha h)^{-1}(1+\beta h) \leq e^{\gamma h},$$

where $\gamma := \beta + 2\alpha$ and

$$\sum_{j=0}^{n-1} e^{j\gamma h} \leq \frac{e^{n\gamma h}}{\gamma h}.$$

Proof. The assumption $\alpha h \leq 1/2$ implies that $-1/2 \leq -\alpha h$, $1/2 \leq 1-\alpha h$ or

$$(1-\alpha h)^{-1} \leq 2.$$

Thus

$$(1-\alpha h)^{-1} = 1 + \alpha h + (\alpha h)^2 + (\alpha h)^3 + ... = 1 + \alpha h + \alpha^2 h^2 (1 + \alpha h + (\alpha h)^2 + ...)$$

$$1 + \alpha h + \alpha^2 h^2 (1-\alpha h)^{-1} = 1 + \alpha h + \alpha h \underbrace{(\alpha h (1-\alpha h)^{-1})}_{\leq 1}$$

$$\leq 1 + 2\alpha h.$$

Also,

$$1 + \beta h \leq 1 + \beta h + \frac{(\beta h)^2}{2!} + ... = e^{\beta h}$$

so

$$(1-\alpha h)^{-1}(1+\beta h) \leq e^{2\alpha h} e^{\beta h} = e^{\gamma h}.$$

Finally, the geometric serie relation

$$\sum_{j=0}^{n-1} e^{j\gamma h} = \sum_{j=0}^{n-1} (e^{\gamma h})^j = \frac{e^{n\gamma h} - 1}{e^{\gamma h} - 1}.$$

□

We can now provide the error analysis for Adams-Moulton.

Theorem 26.1 (Error Estimate for Adams Bashford)**.** *Assume that f satisfies a uniform Lipschitz condition, i.e.*

$$|f(t,x) - f(t,y)| \leq M(x-y), \qquad x, y \in \mathbb{R}, \qquad t \in [0,T].$$

Let $h > 0$, $t_k := kh$ and $x_0, x_1, ..$ be the sequence of the $m+1$ points Adams-Moulton approximates. Set $\rho_h := \max_{m-1<k\leq T/h} |\rho_h(k)|$, $\overline{w} := \max_{j=0,...,m-1} |\widehat{w}_j|$ and $\varepsilon_k := \max_{j=0,...,k} |x(t_j) - x_j|$. Assume that $h \leq h_0 = \frac{1}{2|\widehat{w}_{-1}|M}$, then for $m \leq k \leq T/h$, there holds

$$\varepsilon_k \leq e^{\delta t_k}[\varepsilon_{m-1} + 2\rho_h/\delta],$$

where $\delta := M(m\overline{w} + 2|\widehat{w}_{-1}|)$.

Proof. The AM scheme read

$$x_{k+1} - h\widehat{w}_{-1} f(t_{k+1}, x_{k+1}) = x_k + h \sum_{j=0}^{m-1} \widehat{w}_j f(t_{k-j}, x_{k-j})$$

so that the local truncation error satisfies

$$x(t_{k+1}) - h\widehat{w}_{-1} f(t_{k+1}, x(t_{k+1})) = x(t_k) + h \sum_{j=0}^{m-1} \widehat{w}_j f(t_{k-j}, x_{k-j}) + h\rho_h(k+1).$$

Subtracting the two equations, taking absolute values on both sides and proceeding as in the proof of Theorem 25.1, we get

$$\begin{aligned}|e_{k+1} - h\widehat{w}_{-1}[f(t_{k+1}, x(t_{k+1})) - f(t_{k+1}, x_{k+1})]| &\le |e_k| + h \sum_{j=0}^{m-1} |\widehat{w}_j| M |e_{k-j}| + h\rho_h \\ &\le \varepsilon_k(1 + mM\overline{w}h) + h\rho_h.\end{aligned}$$

where $e_k := x(t_k) - x_k$. Also,

$$\begin{aligned}&|e_{k+1} - h\widehat{w}_{-1}[f(t_{k+1}, x(t_{k+1})) - f(t_{k+1}, x_{k+1})]| \\ &\qquad \ge |e_{k+1}| - h|\widehat{w}_{-1}||f(t_{k+1}, x(t_{k+1})) - f(t_{k+1}, x_{k+1})| \\ &\qquad \ge |e_{k+1}| - h|\widehat{w}_{-1}|M|e_{k+1}|.\end{aligned}$$

Combining the above two estimates we get

$$\varepsilon_{k+1}(1 - h|\widehat{w}_{-1}|M) \le \varepsilon_k(1 + mM\overline{w}h) + h\rho_h$$

or

$$\varepsilon_{k+1} \le (1 - h|\widehat{w}_{-1}|M)^{-1}(1 + mM\overline{w}h)\varepsilon_k + h\rho_h(1 - h|\widehat{w}_{-1}|M)^{-1}.$$

Lemma 26.1 gives (since $h|\widehat{w}_{-1}|M \le 1/2$)

$$\varepsilon_{k+1} \le e^{\gamma h}\varepsilon_k + 2h\rho_h,$$

where

$$\gamma = \delta = M(m\overline{w} + 2|\widehat{w}_{-1}|).$$

Applying repetitively (as in Theorem 25.1) gives

$$\begin{aligned}\varepsilon_k &\le e^{(k-m+1)\delta h}\varepsilon_{m-1} + 2h\rho_h \sum_{j=0}^{k-m} e^{j\delta h} \\ &\le e^{\delta T}(\varepsilon_{m-1} + 2\rho_h/\delta),\end{aligned}$$

where we used Lemma 26.1 for the second inequality. This ends the proof. □

Stiff ODE's (systems)

We start with an example.

Example 26.1 (Backward Euler)**.** *Consider $x(t) \in \mathbb{R}^2$ satisfying*

$$x'(t) = -\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix}}_{=:A} x(t), \qquad x(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Here M is large and positive. These equations decouple:

$$x_1'(t) = -x_1, \qquad x_2'(t) = -Mx_2.$$

The solution is

$$x_1(t) = e^{-t}, \qquad x_2(t) = e^{-Mt},$$

i.e.

$$x(t) = \begin{pmatrix} e^{-t} \\ e^{-Mt} \end{pmatrix}.$$

Now consider the Backward-Euler (fixed step size h) approximation applied to the system is

$$\frac{1}{h}(x_{j+1}(t) - x_j(t) = -Ax_{j+1}.$$

(Warning: here the subindices in x_j denote the approximation at $t_j = jh + t_0$, not the components of x.) Thus

$$(I + hA)x_{j+1} = x_j$$

or

$$x_{j+1} = (I + hA)^{-1}x_j = \begin{pmatrix} (1+h)^{-1} & 0 \\ 0 & (1+Mh)^{-1} \end{pmatrix} x_j.$$

This leads to

$$x_{j+1} = \begin{pmatrix} (1+h)^{-j} & 0 \\ 0 & (1+Mh)^{-j} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The approximate solutions remain bounded for any j and $h > 0$. By the way, $(1+\alpha)^{-j} \approx e^{-\alpha j}$ and so $(1+Mh)^{-j} \approx e^{-Mt_j}$.

Example 26.2 (Forward Euler)**.** *Consider again the system*

$$x'(t) = -\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix}}_{=:A} x(t), \qquad x(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The Forward-Euler scheme reads

$$\frac{1}{h}(x_{j+1}(t) - x_j(t) = -Ax_j.$$

or

$$(I + hA)x_{j+1} = x_j$$

or

$$x_{j+1} = x_j - hAx_j = (I - hA)x_j = \begin{pmatrix} (1-h) & 0 \\ 0 & (1-Mh) \end{pmatrix} x_j$$

so

$$x_{j+1} = \begin{pmatrix} (1-h)^j & 0 \\ 0 & (1-Mh)^j \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

In this case, the solutions remain bounded only if

$$|1 - Mh| \leq 1 \qquad \textit{i.e.} \qquad Mh \leq 2.$$

When $Mh > 2$, the solutions (at least the second component) blow up exponentially (unstable).

Remark 26.1. In the previous examples $x_2(t) = e^{-Mt}$ becomes small fast so a successful scheme should approximate the first component accurately while making the second small.

27. Lecture 27

We start this lecture with another example of stiff system.

Example 27.1 (Parabolic problem). *We consider the parabolic partial differential boundary value problem: find $u : [0,1] \times [0,T] \to \mathbb{R}$ such that*

$$\begin{cases} u_t(x,t) - u_{xx}(x,t) & = 0, & x \in (0,1), \quad t \in (0,T], \\ u(0,t) = u(1,t) & = 0, & t \in [0,T] \\ u(x,0) & = u_0(x), & x \in (0,1) \end{cases}$$

where u_0 is a given initial condition.

A finite difference approximation to this system can be constructed by introducing a grid $x_j = jh$, $h = 1/N$ and a approximation

$$w(x_i,t) \approx u(x_i,t), \qquad i = 0,...,N, \quad t \in [0,T].$$

Using the finite difference approximation to $-u_{xx}$,

$$-u_{xx}(x_i,t) = \frac{2u(x_i,t) - u(x_{i-1},t) - u(x_{i+1},t)}{h^2} + O(h^2)$$

we replace $u(x_i,t)$ by $w(x_i,t)$ and drop the $O(h^2)$ term to arrive at

$$w_t(x_i,t) + \frac{2w(x_i,t) - w(x_{i-1},t) - w(x_{i+1},t)}{h^2} = 0$$

together with $w(x_i,0) = u_0(x_i)$ and $w(x_0,t) = w(x_N,t) = 0$. Now we introduce the vector $v(t)$ defined component wise by

$$v_j(t) = w(x_j,t), \qquad j = 1,...,N-1$$

which therefore satisfies

$$v' = -\frac{1}{h^2}Av,$$

where

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & \mathit{0} & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & \mathit{0} & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

The eigenvectors ψ_j, $j = 1,...,N-1$, of A are known and given by

$$(\psi_j)_k = \sin\left(\frac{\pi jk}{N}\right), \qquad k = 1,...,N-1.$$

The associated eigenvalues are

$$\lambda_j = 2 - 2\cos\left(\frac{\pi j}{N}\right).$$

This means that

$$M^{-1}AM = D$$

where the jth column of M is ψ_j and D is the diagonal matrix with entries $D_{jj} = \lambda_j$. Multiplying the ODE system by M^{-1}, we find

$$M^{-1}v'(t) = -\frac{1}{h^2}M^{-1}AMM^{-1}v$$

or using the notation $\tilde{v}(t) = M^{-1}v(t)$

$$\tilde{v}' = -\frac{1}{h^2}D\tilde{v},$$

i.e.

$$\tilde{v}_j' = -\frac{1}{h^2}\lambda_j\tilde{v}_j.$$

The solution is

$$\tilde{v}_j = e^{-\frac{\lambda_j}{h^2}t}\tilde{v}_j(0).$$

Note that

$$\lambda_1 = 2 - 2\cos\left(\frac{i\pi}{N}\right) \approx 2 - 2\left(1 - \frac{\pi^2}{2N^2}\right) = \frac{\pi^2}{N^2} = \pi^2h^2.$$

Hence $\frac{\lambda_1}{h^2} \approx \pi^2$*. Similarly,*

$$\lambda_{N-1} = 2 - 2\cos\left(\frac{N-1\pi}{N}\right) \approx 4$$

so

$$\frac{\lambda_{N-1}}{h^2} \approx \frac{4}{h^2} = 4N^2.$$

Therefore, we realize that this is a stiff ODE but you do not know the components without knowing the eigenvectors. For more general problems, the eigenvectors are unknown but the eigenvalues behaves the same way.

To characterize/understand the stability of stiff ODEs, we consider the single variable equation for given $\lambda \in \mathbb{C}$:

$$u' = \lambda u, \qquad t > 0$$

supplemented with the initial condition $u(0) = u_0$. Its solution is

$$u(t) = e^{\lambda t}u_0 = e^{(\mathfrak{Re}(\lambda)+i\mathfrak{Im}(\lambda))t}u_0 = e^{\mathfrak{Re}\lambda t}\left(\cos(\mathfrak{Im}(\lambda)t + i\sin(\mathfrak{Im}(\lambda)t)\right)$$

using the Euler's formula. Hence,

$$|u(t)| = |u_0||e^{\lambda t}| = e^{\mathfrak{Re}(\lambda)t}|u_0|,$$

which proves that the solution remains bounded as t gets large if an only if $\mathfrak{Re}(\lambda) \leq 0$. This leads to the following definition.

Definition 27.1 (A-stable)**.** *An ODE method is A-stable if when applied to the ODE*

$$u'(t) = \lambda u(t), \quad t > 0; \qquad u(0) = u_0$$

with fixed timestep h*, the resulting approximation remains bounded for all* h *and all* λ *with* $\mathfrak{Re}(\lambda) \leq 0$*.*

$A-$stable methods are good schemes for stiff ODEs.

Example 27.2 (Backward Euler)**.** *The Backward Euler scheme applied to* $u' = \lambda u$*, with* $u(0) = v$ *given, reads*

$$\frac{u_j - u_{j-1}}{h} = \lambda u_j, \qquad u_0 = v$$

or

$$u_j = u_{j-1} + h\lambda u_j$$

i.e.

$$u_j = (1 - h\lambda)^{-1}u_{j-1} = (1 - h\lambda)^{-j}u_0.$$

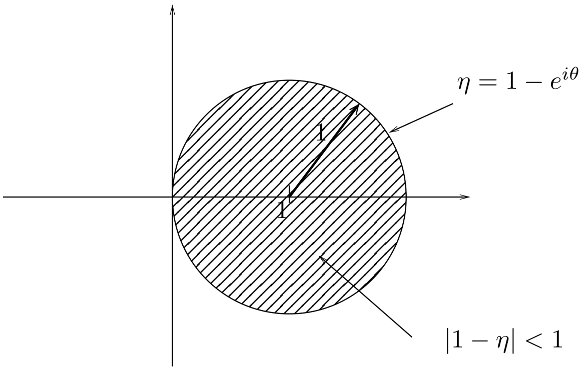

FIGURE 11. Absolute convergence region for Backward Euler.

Hence,

$$|u_j| = |1 - h\lambda|^{-j}|u_0|$$

but

$$|1 - h\lambda|^2 = |1 - h\mathfrak{Re}(\lambda) - ih\mathfrak{Im}(\lambda)|^2 = (1 - h\mathfrak{Re}(\lambda))^2 + h^2\mathfrak{Im}(\lambda)^2 \geq (1 - h\mathfrak{Re}(\lambda))^2$$

and if $\mathfrak{Re}(\lambda) \leq 0$, $1 - h\mathfrak{Re}(\lambda) \geq 1$. *This implies that*

$$(1 - h\mathfrak{Re}(\lambda))^2 \geq 1$$

or

$$\frac{1}{|1 - h\lambda|^j} \leq \frac{1}{|1 - h\lambda|} \leq 1.$$

In turn, we obtain

$$|u_j| \leq |u_0| = |v|,$$

which shows that the approximation remains uniformly in h and therefore that Backward Euler is A-stable.

The previous example also indicates that the Backward Euler scheme is stable if and only if

$$\frac{1}{|1 - h\lambda|} \leq 1$$

or

$$|1 - h\lambda| \geq 1.$$

Note that $1 - \lambda h = 1 - \eta$ (define $\eta = \lambda h$) is a continuous function of η. Moreover,

$$|1 - \eta| = 1 \qquad \Longrightarrow \qquad 1 - \eta = e^{i\theta} \qquad \text{for some} \quad \theta \in [0, 2\pi],$$

i.e.

$$\eta = 1 - e^{i\theta}.$$

Thus the scheme diverges if and only if

$$\eta \in \{z \in \mathbb{C} \ : \ |1 - z| < 1\} =: B_1(1).$$

This region is depicted in Figure 11.

Definition 27.2 (Absolute Stability)**.** *The region of absolute stability of a scheme is the set of* $\eta = h\lambda$ *such that the approximate solutions remain bounded.*

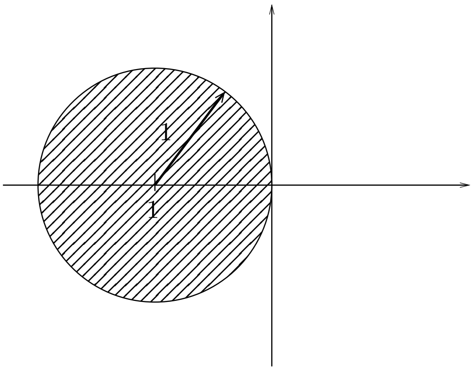

FIGURE 12. Absolute convergence region for Forward Euler (dashed region).

Example 27.3 (Backward Euler). *The region of absolute stability for Backward Euler is*

$$\mathbb{C} \setminus B_1(1) = \{z \in \mathbb{C} \ : \ |1 - z| \geq 1\}\,.$$

Example 27.4 (Forward Euler). *The Forward Euler scheme applied to $u' = \lambda u$ with $u(0) = u_0$ is*

$$u_j = (1 + h\lambda)u_{j-1}$$

or

$$u_j = (1 + h\lambda)^j u_0.$$

The solutions remain bounded if and only if

$$|1 + h\lambda| \leq 1.$$

Note that $|1 + h\lambda| = 1$ if and only if

$$1 + h\lambda = e^{i\theta}, \qquad \theta \in [0, 2\pi].$$

If $\eta := h\lambda$ then $1 + \eta = e^{i\theta}$ or $\eta = (e^{i\theta} - 1)$, see Figure 12. Forward Euler is not $A-$stable.

Remark 27.1 (Aboslute and A-stability). A scheme is $A-$stable if and only if the region of absolute stability contains

$$\{z \in \mathbb{C} \ : \ \mathfrak{Re}(z) \leq 0\}.$$