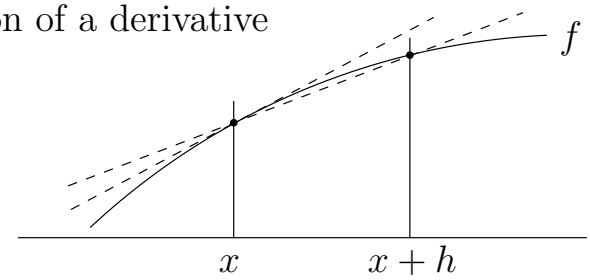
1
Thurs
1/11

warm-up : finite-difference approximation of a derivative



forward difference

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} = D_+ f(x)$$

question : how large is the error?

$$f(x+h) = f(x) + f'(x)h + \tfrac{1}{2}f''(x)h^2 + \cdots$$

$$\Rightarrow \frac{f(x+h) - f(x)}{h} = f'(x) + \underbrace{\frac{1}{2}f''(x)h + \cdots}$$

↑ approximation ; ↑ exact value ; ↑ truncation error

Hence the error is proportional to h; we write this as $D_+ f(x) = f'(x) + O(h)$.

For example, if $f(x) = e^x, x = 1$, then $f'(1) = e = 2.71828\ldots$ is the exact value.

h	$D_+ f(1)$	$\lvert D_+ f(1) - f'(1)\rvert$	$\lvert D_+ f(1) - f'(1)\rvert / h$
0.1	2.8588	0.1406	1.4056
0.05	2.7874	0.0691	1.3821
0.025	2.7525	0.0343	1.3705
↓	↓	↓	↓
0	e	0	$\frac{0}{0} = \frac{1}{2}f''(1) = \frac{e}{2}$

In general, if error $\approx ch^p$, then p is called the order of accuracy and $\log(\text{error}) \approx \log(ch^p) = \log c + p \log h$, so p is the slope of the data in a log-log plot. We see that $p = 1$ for large h (expected) and $p = -1$ for small h (unexpected).

question : why does the error increase for small h?

1. The computed value has two sources of error: truncation error is due to replacing the exact derivative $f'(x)$ by the finite-difference approximation $D_+ f(x)$, and roundoff error is due to using finite precision arithmetic.

2. The truncation error is $O(h)$ and the roundoff error is $O(\epsilon/h)$, where $\epsilon \approx 10^{-15}$ in Matlab.

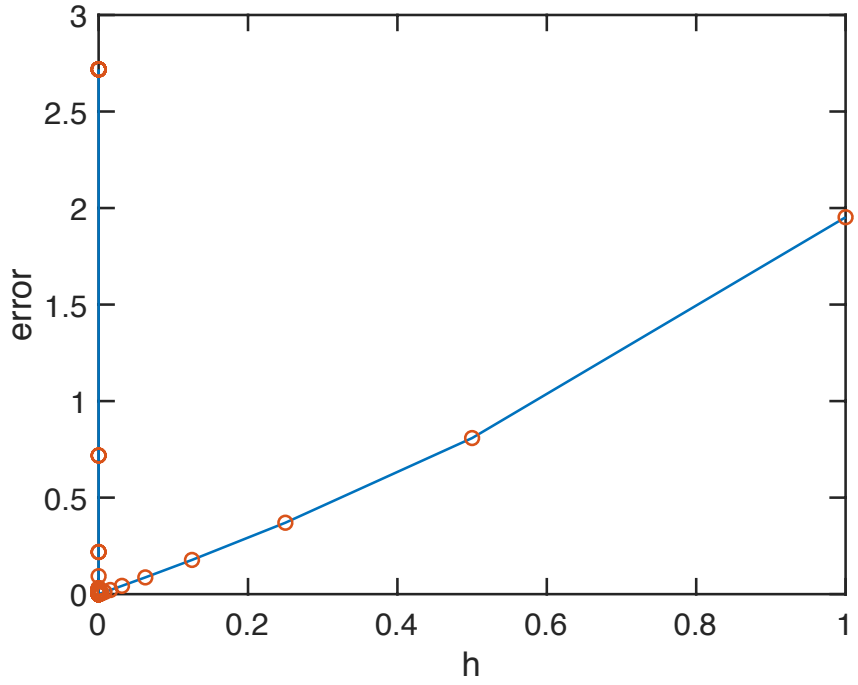
3. The total error is $O(h) + O(\epsilon/h)$, so for large h the truncation error dominates, but for small h the roundoff error dominates.

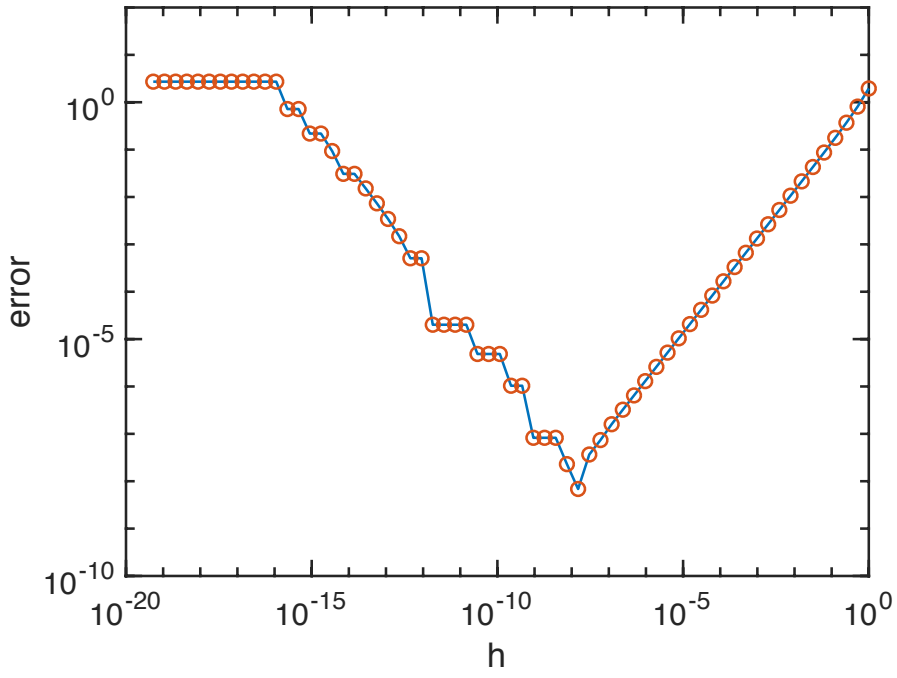
note : $D_- f(x) = \dfrac{f(x) - f(x-h)}{h}$: backward difference

$$D_0 f(x) = \frac{f(x+h) - f(x-h)}{2h} \text{ : central difference , hw}$$

finite-difference approximation of a derivative

```
% Matlab
   exact_value = exp(1);
   for j=1:65
     h(j) = 1/2^(j-1);
     computed_value = (exp(1+h(j)) - exp(1))/h(j);
     error(j) = abs(computed_value - exact_value);
   end
   plot(h,error,h,error,'o'); loglog(h,error,h,error,'o'); ...
```





1. IVP for ODEs : given $y' = f(y)$, $y(0) = y_0$, find $y(t)$

example : $y' = y$, $y(0) = y_0 \Rightarrow y(t) = y_0\, e^t$

$y' = \sin y$, $y(0) = y_0 \Rightarrow y(t) = ?$

definition : An IVP is well-posed if the following conditions are satisfied.

1. A solution exists.
2. The solution is unique.
3. The solution depends continuously on the data (i.e. y_0 , f).

definition : f satisfies a Lipschitz condition on a domain D if there exists a constant $L \geq 0$ such that $|f(y) - f(u)| \leq L\,|y - u|$ for all $y\,, u \in D$; we typically assume that $f(y)$ is smooth, so $L = \max |f_y|$. The Lipschitz constant L controls how quickly nearby solutions can diverge from each other; a related concept is the Lyapunov exponent (Math 558).

theorem : If f satisfies a Lipschitz condition, then the IVP is well-posed.

example

$y' = ay + b$, $y(0) = y_0 \Rightarrow L = |a|$, $y(t) = y_0\, e^{at} + \dfrac{b}{a}(e^{at} - 1)$

$y' = y^{1/2}$, $y(0) = 0 \Rightarrow L = \infty$ on $\{y \geq 0\}$, $y(t) = 0\,, \dfrac{t^2}{4}$: non-unique solution

$y' = y^2$, $y(0) = 1 \Rightarrow L = \infty$ on $\{y \geq 1\}$, $y(t) = (t-1)^{-1} \Rightarrow \lim\limits_{t \to 1} y(t) = \infty$

Euler's method

2
Tues
1/16

$h = \Delta t$: time step , $t_n = nh$, $n = 0, 1, 2, 3, \ldots$

$y_n = y(t_n)$: exact solution , $y' = f(y)$, $y(0) = y_0$

u_n : approximation , $\dfrac{u_{n+1} - u_n}{h} = f(u_n) \Rightarrow u_{n+1} = u_n + hf(u_n)$, $u_0 = y_0$

example : $y' = y$, $y_0 = 1 \Rightarrow u_{n+1} = u_n + hu_n = (1+h)u_n$

$u_0 = 1$

$u_1 = 1 + h$

$u_2 = (1+h)^2$, $\ldots$, $u_n = (1+h)^n$

consider $t_n = 1 \Rightarrow nh = 1$, $y_n = y(1) = e = 2.7182$

n	h	u_n	$\lvert y_n - u_n \rvert$	$\lvert y_n - u_n \rvert / h$
1	1	2.0000	0.7183	0.7183
2	1/2	2.2500	0.4683	0.9366
4	1/4	2.4414	0.2769	1.1075
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
∞	0	e	0	? : hw

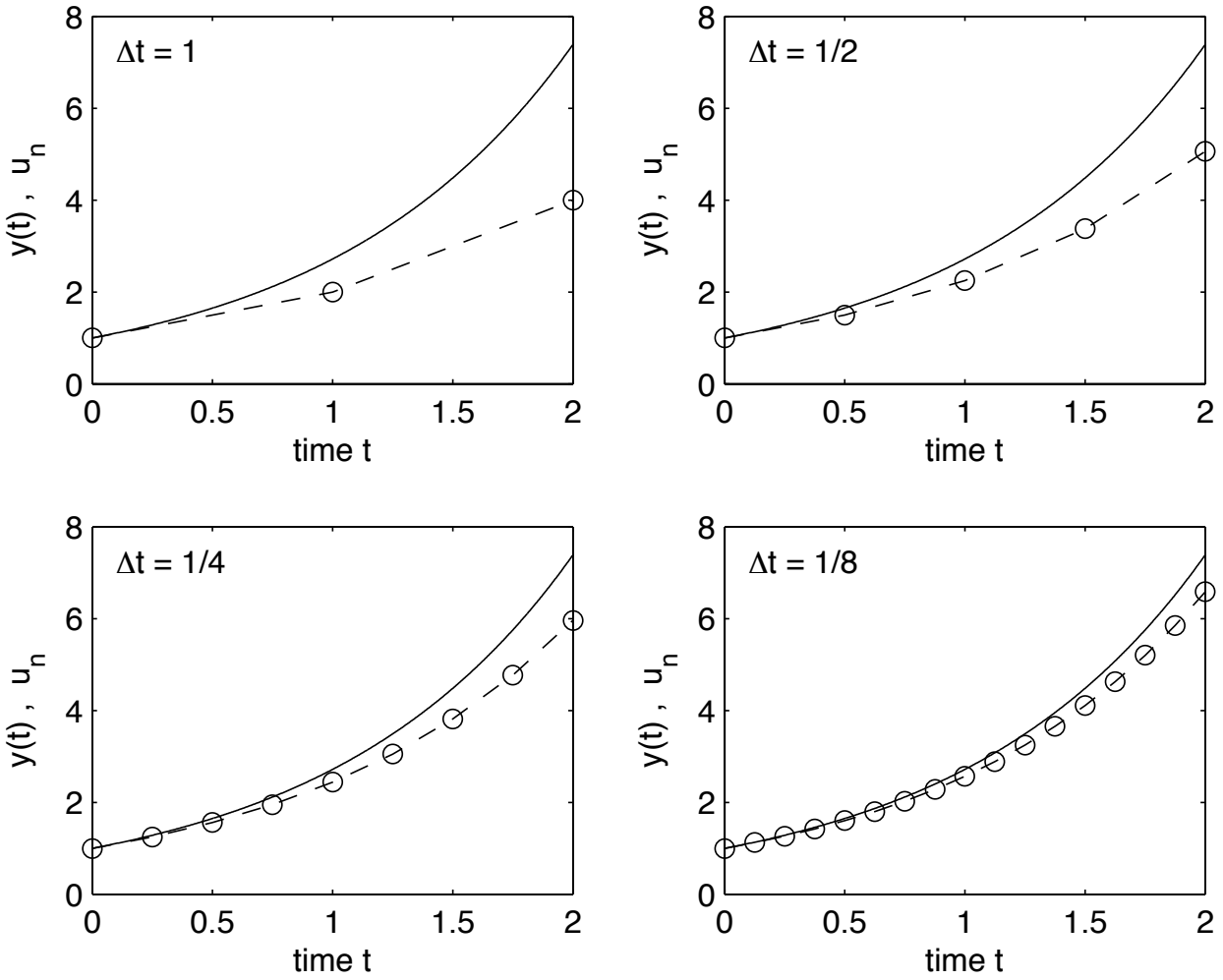
Euler's method

IVP : $y' = y \ , \ y(0) = 1$

finite-difference scheme : $u_{n+1} = u_n + hu_n \ , \ u_0 = 1$

The solid line is the exact solution $y(t)$.

The circles are the numerical solution u_n.



1. For a fixed time t, the error decreases as the time step $h \to 0$.
2. For a fixed time step h, the error increases as time $t \to \infty$.

convergence proof (special case)

$y' = y \ , \ y_0 = 1 \ \Rightarrow \ y(t) = e^t$

$u_{n+1} = u_n + hu_n \ , \ u_0 = 1 \ \Rightarrow \ u_n = (1+h)^n$

consider $t = 1 \ , \ h = 1/n$

$$\lim_{h\to 0} u_n = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = L$$

$$\ln L = \ln \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = \lim_{n\to\infty} \ln \left(1 + \frac{1}{n}\right)^n = \lim_{n\to\infty} n \ln \left(1 + \frac{1}{n}\right) = \infty \cdot 0$$

$$= \lim_{h\to 0} \frac{\ln(1+h)}{h} = \frac{0}{0} = \lim_{h\to 0} \frac{1}{1+h} = 1 \Rightarrow \ln L = 1 \Rightarrow L = e$$

$\Rightarrow \lim_{h\to 0} u_n = e = y(1) \ \Rightarrow$ Euler's method converges

hw : $y' = ay + b \ , \ y(0) = y_0$

---

convergence proof (general case)

$y' = f(y) \ , \ y(0) = y_0$

$u_{n+1} = u_n + hf(u_n) \ , \ u_0$: given

fix $t > 0$, set $h = t/n \ \Rightarrow \ t = nh = t_n \ , \ y_n = y(t_n) = y(t)$

goal : $\lim_{h\to 0} u_n = y(t)$

step 1 : substitute y_n into Euler's method

$y_{n+1} = y_n + hf(y_n) + \tau_n$, τ_n : local truncation error

$y_{n+1} = y(t_{n+1}) = y(t_n + h) = y(t_n) + hy'(t_n) + \frac{1}{2}h^2 y''(\tilde{t}\,)$

$\Rightarrow \ y_{n+1} = y_n + hf(y_n) + \frac{1}{2}h^2 y''(\tilde{t}\,) \ \Rightarrow \ \tau_n = \frac{1}{2}h^2 y''(\tilde{t}\,) = O(h^2)$

step 2 : analyze error , $e_n = y_n - u_n$

$e_{n+1} = y_{n+1} - u_{n+1} = y_n + hf(y_n) + \tau_n - (u_n + hf(u_n))$

$= (y_n - u_n) + h(f(y_n) - f(u_n)) + \tau_n$

$|e_{n+1}| \le |e_n| + h|f(y_n) - f(u_n)| + |\tau_n| \ \le \ |e_n| + hL|y_n - u_n| + \tau \ , \ \tau = \max|\tau_n|$

$|e_{n+1}| \le (1 + hL)|e_n| + \tau$

$|e_1| \le (1 + hL)|e_0| + \tau$

$|e_2| \le (1 + hL)|e_1| + \tau \le (1 + hL)((1 + hL)|e_0| + \tau) + \tau$

$= (1 + hL)^2|e_0| + (1 + (1 + hL))\tau$

$|e_n| \le (1+hL)^n|e_0| + \sum_{i=0}^{n-1}(1+hL)^i\tau$

$1+x \le e^x \;\Rightarrow\; (1+x)^n \le e^{nx} \;\Rightarrow\; (1+hL)^n \le e^{nhL} = e^{Lt}$

$\sum_{i=0}^{n-1}(1+hL)^i = \dfrac{(1+hL)^n - 1}{(1+hL)-1} \le \dfrac{e^{Lt}-1}{hL}$

recall : $\sum_{i=0}^{n} r^i = \dfrac{r^{n+1}-1}{r-1}$ for $r \ne 1$

$\Rightarrow |e_n| \;\le\; e^{Lt}|e_0| + \dfrac{e^{Lt}-1}{hL}\,\tau \;\;,\;\; \tau = \max|\tau_n| = \frac{1}{2}h^2M \;\;,\;\; M = \max|y''(t)|$

$|e_n| \;\le\; e^{Lt}|e_0| + \dfrac{e^{Lt}-1}{L}\cdot\dfrac{Mh}{2}$

if $\lim_{h\to 0}|e_0| = 0$, then $\lim_{h\to 0}|e_n| = 0 \;\Rightarrow\; \lim_{h\to 0} u_n = y(t)$ ok

---

roundoff error

$u_{n+1} = u_n + hf(u_n)$: exact arithmetic

$v_{n+1} = v_n + hf(v_n) + \epsilon_n$: finite precision arithmetic

$y_{n+1} = y_n + hf(y_n) + \tau_n$

let $e_n = y_n - v_n$

$e_{n+1} \;=\; e_n + h(f(y_n) - f(v_n)) + \tau_n - \epsilon_n$

$|e_{n+1}| \;\le\; (1+hL)|e_n| + \tau + \epsilon \;,\; \tau = \max|\tau_n| \;,\; \epsilon = \max|\epsilon_n|$

$|e_n| \;\le\; e^{Lt}|e_0| + \dfrac{e^{Lt}-1}{L}\left(\dfrac{Mh}{2} + \dfrac{\epsilon}{h}\right) \;,\; \dfrac{\epsilon}{h} \to \infty \;\text{ as }\; h \to 0$

---

note : $|y_n - u_n| \;\le\; \dfrac{e^{Lt}-1}{L}\cdot\dfrac{Mh}{2}$: error bound

We can write this as $y_n - u_n = O(h)$; this means there exists a constant C such that $\dfrac{|y_n - u_n|}{h} \le C$ for h sufficiently small; we can take $C = \dfrac{e^{Lt}-1}{L}\cdot\dfrac{M}{2}$.

example : $y' = y$, $y(0) = 1$, $L = 1$, $t = 1$, $M = e \;\Rightarrow\; C = \dfrac{e-1}{1}\cdot\dfrac{e}{2} = 2.3354$

However, we saw that $\dfrac{|y_n - u_n|}{h} = 1.2200$ for $h = 1/8$, so we want to understand more precisely how the error depends on h.

claim : $u_n = y_n + hE_n + O(h^2)$: asymptotic expansion

$y_n = y(t_n) = y(t)$, $E_n = E(t_n) = E(t)$: principal error function

$E' = f_y(y)E - \frac{1}{2}y''$, $E(0) = 0$, $E(t)$

This means there exists a constant C such that $\dfrac{|u_n - (y_n + hE_n)|}{h^2} \le C$ for h sufficiently small.

example : $y' = y$, $y(0) = 1$

$E' = E - \frac{1}{2}e^t \Rightarrow E(t) = -\frac{1}{2}te^t$, $E(1) = -\frac{1}{2}e = -1.3591$

$\Rightarrow u_n = y_n - 1.3591h + O(h^2)$

---

proof : define $d_n = u_n - (y_n + hE_n)$

3
Thurs
1/18

we already know that $d_n = O(h)$, we must show that $d_n = O(h^2)$

$d_{n+1} = u_{n+1} - (y_{n+1} + hE_{n+1})$

$= \cancel{u_n} + \cancel{hf(u_n)} - (\cancel{y_n} + \cancel{hy'_n} + \frac{1}{2}h^2y''_n + O(h^3) + h(\cancel{E_n} + hE'_n + O(h^2)))$

$= u_n - (y_n + hE_n) + h(f(u_n) - y'_n) - h^2(\frac{1}{2}y''_n + E'_n) + O(h^3)$

$f(u_n) = f(y_n + hE_n + d_n) = f(y_n) + f_y(y_n)(hE_n + d_n) + O(h^2)$

$f(u_n) - y'_n = f_y(y_n)(hE_n + d_n) + O(h^2)$

$d_{n+1} = d_n + hf_y(y_n)(hE_n + d_n) + O(h^3) - h^2(\frac{1}{2}y''_n + E'_n) + O(h^3)$

$= d_n + hf_y(y_n)d_n + h^2\underbrace{(f_y(y_n)E_n - \frac{1}{2}y''_n - E'_n)} + O(h^3)$ $= 0$ by definition of E

$d_{n+1} = d_n + hf_y(y_n)d_n + O(h^3)$

$|d_{n+1}| \le (1 + hL)|d_n| + \kappa$, $\kappa = O(h^3)$, assume $d_0 = 0$

$|d_n| \le \dfrac{e^{Lt} - 1}{hL}\kappa = O(h^2)$ ok

---

summary

$u_n = y_n + O(h)$

$u_n = y_n + hE_n + O(h^2)$

$u_n = y_n + hE_n + h^2D_n + O(h^3)$, where $D_n = D(t_n)$

hw : find the equation satisfied by $D(t)$

application : Richardson extrapolation

$u_n = y_n + hE_n + h^2 D_n + \cdots$

$A_{00} = u^h = y + hE + h^2 D + O(h^3)$

$A_{10} = u^{h/2} = y + \frac{h}{2}E + \frac{h^2}{4}D + O(h^3)$

$A_{20} = u^{h/4} = y + \frac{h}{4}E + \frac{h^2}{16}D + O(h^3)$

eliminate $O(h)$ term

$2A_{10} - A_{00} = A_{11} = y - \frac{h^2}{2}D + O(h^3)$

$2A_{20} - A_{10} = A_{21} = y - \frac{h^2}{8}D + O(h^3)$

eliminate $O(h^2)$ term

$\frac{4A_{21} - A_{11}}{3} = A_{22} = y + O(h^3)$

$\cdots$

$A_{j0} = y + O(h)$: computed using time step $h/2^j$

$A_{j1} = 2A_{j0} - A_{j-1,0} = y + O(h^2)$

$A_{j2} = \frac{4A_{j1} - A_{j-1,1}}{3} = y + O(h^3)$

$A_{j3} = \frac{8A_{j2} - A_{j-1,2}}{7} = y + O(h^4)$

example : $y' = y$, $y(0) = 1 \quad \Rightarrow \quad y(1) = e = 2.7182818$

j	h	A_{j0}	A_{j1}	A_{j2}	A_{j3}
0	0.1	2.5937425			
1	0.05	2.6532977	2.7128529		
2	0.025	2.6850638	2.7168299	2.7181556	
3	0.0125	2.7014849	2.7179060	2.7182647	2.7182803
		$O(h)$	$O(h^2)$	$O(h^3)$	$O(h^4)$

down a column : decreasing time step , fixed order of accuracy : h-refinement

across a row : fixed time step , increasing order of accuracy : p-refinement

Taylor series methods

$y' = f(y)$

$y_{n+1} = y_n + hy'_n + \dfrac{h^2}{2}\, y''_n + \cdots$

1st order Taylor series method

$y'_n = f(y_n)$

$u_{n+1} = u_n + hf(u_n)$: Euler's method

2nd order Taylor series method

$y''_n = f_y(y_n) \cdot y'_n = f_y(y_n) \cdot f(y_n)$

$u_{n+1} = u_n + hf(u_n) + \dfrac{h^2}{2} f_y(u_n) \cdot f(u_n)$

example

$y' = y\ ,\ y(0) = 1$

$u_{n+1} = u_n + hu_n + \dfrac{h^2}{2}\, u_n = \Big(1 + h + \dfrac{h^2}{2}\Big) u_n$

$t = 1\ ,\ h = 1/n$

h	u_n	$\lvert y_n - u_n \rvert$	$\lvert y_n - u_n \rvert / h^2$
0.1	2.7140808	0.0042010	0.4201
0.05	2.7171911	0.0010907	0.4363
0.025	2.7180039	0.0002779	0.4447
0.0125	2.7182117	0.0000701	0.4488

4
Tues
1/23

Runge-Kutta methods

$y' = f(y)$

$k_1 = f(u_n)$

$k_2 = f(u_n + ahk_1)$

$u_{n+1} = u_n + h(bk_1 + ck_2)$: 2-stage RK method

choose $a\,, b\,, c$ to minimize the local truncation error

$y_{n+1} = y_n + h(bf(y_n) + cf(y_n + ahf(y_n))) + \tau_n$

$f(y_n) = y_n'$

$f(y_n + ahf(y_n)) = f(y_n) + f_y(y_n) \cdot ahf(y_n) + O(h^2) = y_n' + ahy_n'' + O(h^2)$

(1) $y_{n+1} = y_n + h(by_n' + c(y_n' + ahy_n'' + O(h^2))) + \tau_n$

$\qquad = y_n + (b+c)hy_n' + ach^2y_n'' + O(h^3) + \tau_n$

(2) $y_{n+1} = y_n + hy_n' + \frac{1}{2}h^2y_n'' + O(h^3)$

equate powers of h : $b + c = 1\,,\ ac = \frac{1}{2}\,,\ \tau_n = O(h^3)$

$c = \frac{1}{2a}\,,\ b = 1 - \frac{1}{2a}$: 1-parameter family of 2nd order methods

midpoint method : $a = \frac{1}{2}\,,\ b = 0\,,\ c = 1$

$k_1 = f(u_n)$

$k_2 = f\Big(u_n + \frac{h}{2}k_1\Big)$

$u_{n+1} = u_n + hk_2 = u_n + hf\Big(u_n + \frac{h}{2}f(u_n)\Big)$

modified Euler method : $a = 1\,,\ b = \frac{1}{2}\,,\ c = \frac{1}{2}$

$k_1 = f(u_n)$

$k_2 = f(u_n + hk_1)$

$u_{n+1} = u_n + h\Big(\frac{1}{2}k_1 + \frac{1}{2}k_2\Big) = u_n + \frac{h}{2}\Big(f(u_n) + f(u_n + hf(u_n))\Big)$

4th order Runge-Kutta

$k_1 = f(u_n)$

$k_2 = f\Big(u_n + \dfrac{h}{2}k_1\Big)$

$k_3 = f\Big(u_n + \dfrac{h}{2}k_2\Big)$

$k_4 = f(u_n + hk_3)$

$u_{n+1} = u_n + \dfrac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$: 4 stages

example

$y' = y\ ,\ y(0) = 1$

$k_1 = u_n$

$k_2 = u_n + \dfrac{h}{2}k_1 = \Big(1 + \dfrac{h}{2}\Big)u_n$

$k_3 = u_n + \dfrac{h}{2}k_2 = \Big(1 + \dfrac{h}{2}\Big(1 + \dfrac{h}{2}\Big)\Big)u_n = \Big(1 + \dfrac{h}{2} + \dfrac{h^2}{4}\Big)u_n$

$k_4 = u_n + hk_3 = \Big(1 + h\Big(1 + \dfrac{h}{2} + \dfrac{h^2}{4}\Big)\Big)u_n = \Big(1 + h + \dfrac{h^2}{2} + \dfrac{h^3}{4}\Big)u_n$

$u_{n+1} = u_n + \dfrac{h}{6}\Big(1 + 2\Big(1 + \dfrac{h}{2}\Big) + 2\Big(1 + \dfrac{h}{2} + \dfrac{h^2}{4}\Big) + \Big(1 + h + \dfrac{h^2}{2} + \dfrac{h^3}{4}\Big)\Big)u_n$

$u_{n+1} = \Big(1 + h + \dfrac{h^2}{2} + \dfrac{h^3}{6} + \dfrac{h^4}{24}\Big)u_n$

$t = 1\ ,\ h = 1/n$

h	u_n	$\lvert y_n - u_n\rvert$	$\lvert y_n - u_n\rvert/h^4$
0.2	2.7182511	0.00003069185	0.01918
0.1	2.7182797	0.00000208432	0.02084
0.05	2.7182817	0.00000013580	0.02173

note : RK methods have the form $u_{n+1} = u_n + hF(u_n, h)$, where

(a) $F(u, 0) = f(u)$,

(b) $F(u, h)$ satisfies a Lipschitz condition in u, i.e. there exists a constant $\tilde{L}$ such that $|F(u, h) - F(v, h)| \leq \tilde{L}|u - v|$ for h sufficiently small.

example : modified Euler method

$$u_{n+1} = u_n + \frac{h}{2}\Big(f(u_n) + f(u_n + hf(u_n))\Big)$$

$$F(u, h) = \frac{1}{2}\Big(f(u) + f(u + hf(u))\Big)$$

(a) $F(u, 0) = f(u)$ ok

(b) $|F(u, h) - F(v, h)|$

$$= \frac{1}{2}|f(u) + f(u + hf(u)) - (f(v) + f(v + hf(v)))|$$

$$= \frac{1}{2}|f(u) - f(v) + f(u + hf(u)) - f(v + hf(v))|$$

$$\leq \frac{1}{2}|f(u) - f(v)| + \frac{1}{2}|f(u + hf(u)) - f(v + hf(v))|$$

$$\leq \frac{1}{2}L|u - v| + \frac{1}{2}L|u + hf(u) - (v + hf(v))|$$

$$= \frac{1}{2}L|u - v| + \frac{1}{2}L|u - v + h(f(u) - f(v))|$$

$$\leq \frac{1}{2}L|u - v| + \frac{1}{2}L|u - v| + \frac{1}{2}hL|f(u) - f(v)|$$

$$\leq L|u - v| + \frac{1}{2}hL \cdot L|u - v| = \Big(L + \frac{1}{2}hL^2\Big)|u - v|$$

$\Rightarrow |F(u, h) - F(v, h)| \leq \tilde{L}|u - v|$ for $0 \leq h \leq 1$, where $\tilde{L} = L + \frac{1}{2}L^2$ ok

hw : midpoint method

theorem

1. (a) $\Rightarrow \tau_n = O(h^2)$ (at least)
2. (b) $\Rightarrow$ the scheme is stable wrt initial data
3. (a) + (b) $\Rightarrow$ the scheme converges, i.e. $\lim_{h \to 0} u_n = y(t)$, $t = nh$

note

1. (a) $\Rightarrow$ the difference scheme approximates the correct differential equation

$u_{n+1} = u_n + hF(u_n, h)$

$y_{n+1} = y_n + hF(y_n, h) + \tau_n$

$\frac{y_{n+1} - y_n}{h} = F(y_n, h) + \frac{\tau_n}{h} \Rightarrow y_n' = f(y_n)$ as $h \to 0$

We say that the difference scheme is consistent with the differential equation.

2. The scheme is stable wrt initial data if there exists a constant C such that for all $n \geq 1$, $|u_n - v_n| \leq C|u_0 - v_0|$, where u_n, v_n are the numerical solutions starting from u_0, v_0, and $t = nh$ is fixed. The constant C may depend on t, but must be independent of n and h. In other words, a small change in the initial data of the difference scheme leads to a small change in the solution of the difference scheme.

3. The theorem says that consistency + stability $\Rightarrow$ convergence.

proof

5 Thurs 1/25

1. $u_{n+1} = u_n + hF(u_n, h)$

$y_{n+1} = y_n + hF(y_n, h) + \tau_n$

$\Rightarrow \cancel{y_n} + h\cancel{y_n'} + O(h^2) = \cancel{y_n} + h(F(\cancel{y_n}, 0) + O(h)) + \tau_n \Rightarrow \tau_n = O(h^2)$ ok

2. Let u_n, v_n be numerical solutions starting from u_0, v_0.

$$|u_{n+1} - v_{n+1}| = |u_n + hF(u_n, h) - (v_n + hF(v_n, h))|$$
$$\leq |u_n - v_n| + h|F(u_n, h) - F(v_n, h)|$$
$$\leq (1 + h\tilde{L})|u_n - v_n|$$

$\Rightarrow |u_n - v_n| \leq (1 + h\tilde{L})^n |u_0 - v_0| \leq e^{nh\tilde{L}}|u_0 - v_0| = e^{\tilde{L}t}|u_0 - v_0|$ ok

3. define $e_n = y_n - u_n$

$e_{n+1} = y_{n+1} - u_{n+1} = y_n + hF(y_n, h) + \tau_n - (u_n + hF(u_n, h)) = \cdots$

$|e_{n+1}| \leq (1 + h\tilde{L})|e_n| + |\tau_n| \Rightarrow |e_n| \leq e^{\tilde{L}t}|e_0| + \frac{e^{\tilde{L}t} - 1}{h\tilde{L}}\tau \ , \ \tau = \max|\tau_n|$ ok

---

1. A method of the form $u_{n+1} = u_n + hF(u_n, h)$ is an explicit 1-step method; this includes RK and Taylor series methods.

2. If $\tau = O(h^{p+1})$, then $|y_n - u_n| = O(h^p)$, i.e. the global error is one order lower than the local truncation error; we say that the method is p-th order accurate.

generalization

1. systems

$$\left.\begin{array}{l} y_1' = f_1(y_1, \ldots, y_N) \\ \quad \vdots \\ y_N' = f_N(y_1, \ldots, y_N) \end{array}\right\} \Rightarrow y' = f(y) \ , \ y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \ , \ f = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}$$

2. non-autonomous equations

$$y' = f(y,t) \ , \ \left.\begin{array}{l} y_1 = y \\ y_2 = t \end{array}\right\} \Rightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} f(y_1, y_2) \\ 1 \end{pmatrix}$$

3. higher order equations

$$y'' = f(y, y') \ , \ \left.\begin{array}{l} y_1 = y \\ y_2 = y' \end{array}\right\} \Rightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} y_2 \\ f(y_1, y_2) \end{pmatrix}$$

---

definition : A vector norm $||y||$ has the following properties.

1. $||y|| \geq 0 \ , \ ||y|| = 0 \Leftrightarrow y = 0$
2. $||\alpha y|| = |\alpha| \cdot ||y||$
3. $||y + u|| \leq ||y|| + ||u||$

example

$$||y||_1 = \sum_{i=1}^{N} |y_i| \quad , \quad ||y||_\infty = \max |y_i| \quad , \quad ||y||_2 = \left(\sum_{i=1}^{N} |y_i|^2 \right)^{1/2}$$

note

Given a vector norm $||y||$, the subordinate matrix norm is $||A|| = \max\limits_{y \neq 0} \dfrac{||Ay||}{||y||}$.

$||A||$ satisfies properties 1, 2, 3 above and 4, 5 below.

4. $||Ay|| \leq ||A|| \cdot ||y||$, 5. $||AB|| \leq ||A|| \cdot ||B||$

example

$||A||_1 = \max\limits_j \sum\limits_{i=1}^{N} |a_{ij}|$: max column sum

$||A||_\infty = \max\limits_i \sum\limits_{j=1}^{N} |a_{ij}|$: max row sum

$||A||_2 = \max\limits_i \sigma_i(A)$: max singular value , $\max\limits_i |\lambda_i(A)|$ if A is real symmetric

proof : Math 571

note : The previous results for a scalar equation $y' = f(y)$ also hold for a system of equations with $|y| \to ||y|| \, , \, L = \max |f_y| \to L = \max ||f_y|| \, , \, f_y = (\partial f_i / \partial y_j)$.

absolute stability

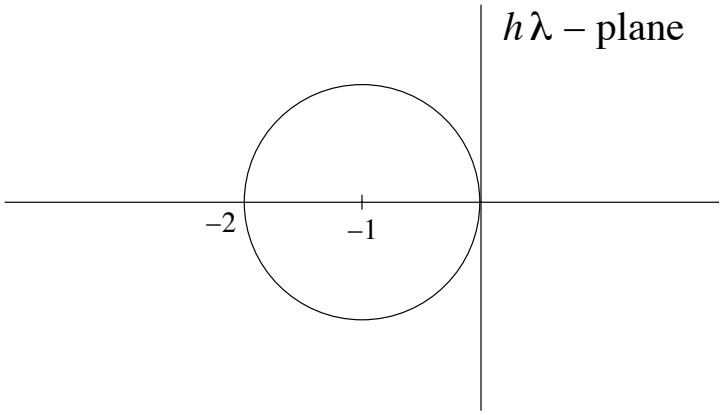
Consider $y' = \lambda y$, where y is a scalar and λ is a complex number with real$(\lambda) \le 0$ (test equation). The exact solution is $y(t) = y(0)e^{\lambda t}$, which is bounded as $t \to \infty$ for all initial data. A numerical method with time step h is absolutely stable if the numerical solution u_n is bounded as $n \to \infty$ for all initial data. This is different than stability wrt initial data which was concerned with $t_n = nh$ fixed and $h \to 0$. As we shall see, the region of absolute stability of a numerical method is a subset of the complex $h\lambda$-plane.

---

example : Euler's method

$y' = \lambda y \;\Rightarrow\; u_{n+1} = u_n + h\lambda u_n = (1 + h\lambda)u_n \;\Rightarrow\; u_n = (1 + h\lambda)^n u_0$

u_n is bounded as $n \to \infty \Leftrightarrow\; |1 + h\lambda| \le 1 \Leftrightarrow |h\lambda - (-1)| \le 1$



The region of absolute stability of Euler's method is the interior of a unit circle centered at -1 in the $h\lambda$-plane.

---

example : $y' = Ay$, where A is diagonalizable , $A = XDX^{-1}$

6
Tues
1/30

$D = \text{diag}(\lambda_1, \ldots, \lambda_N)$, $X = [x_1 \cdots x_N]$, $Ax_j = \lambda_j x_j$, $j = 1 : N$

$y(t) = \alpha_1 e^{\lambda_1 t} x_1 + \cdots + \alpha_N e^{\lambda_N t} x_N$

$u_{n+1} = u_n + hAu_n = (I + hA)u_n$: Euler's method

$u_n = (I + hA)^n u_0 = \alpha_1 (1 + h\lambda_1)^n x_1 + \cdots + \alpha_N (1 + h\lambda_N)^n x_N$

absolute stability $\Leftrightarrow |1 + h\lambda_j| \le 1$, $j = 1 : N$

Hence to ensure absolute stability of Euler's method applied to a diagonalizable linear system $y' = Ay$, it is necessary to ensure absolute stability for the scalar test equation $y' = \lambda y$ for every eigenvalue of A.

example

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} -11 & 9 \\ 9 & -11 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Rightarrow \lambda_1 = -20 \ , \ \lambda_2 = -2 \ : \ \underline{\text{stiff system}}$$

$y(t) = \alpha_1 e^{-20t} x_1 + \alpha_2 e^{-2t} x_2$: fast component + slow component

$u_n = \alpha_1 (1 - 20h)^n x_1 + \alpha_1 (1 - 2h)^n x_2$

absolute stability $\Leftrightarrow |1 - 20h| \leq 1 \ , \ |1 - 2h| \leq 1$

$\Leftrightarrow -1 \leq 1 - 20h \leq 1 \ , \ -1 \leq 1 - 2h \leq 1$

$\Leftrightarrow -1 \leq 1 - 20h \ , \ -1 \leq 1 - 2h \Leftrightarrow 20h \leq 2 \ , \ 2h \leq 2 \Leftrightarrow h \leq 0.1 \ , \ h \leq 1$

Hence we must choose $h \leq 0.1$ to ensure that Euler's method is absolutely stable. Note that the fast component of the exact solution decays rapidly and after some time essentially only the slow component is present, $y(t) \approx \alpha_2 e^{-2t} x_2$. By itself the slow component only requires $h \leq 1$ for absolute stability, but we must choose $h \leq 0.1$ to ensure absolute stability of Euler's method applied to the system. What happens if we don't?

---

backward Euler method

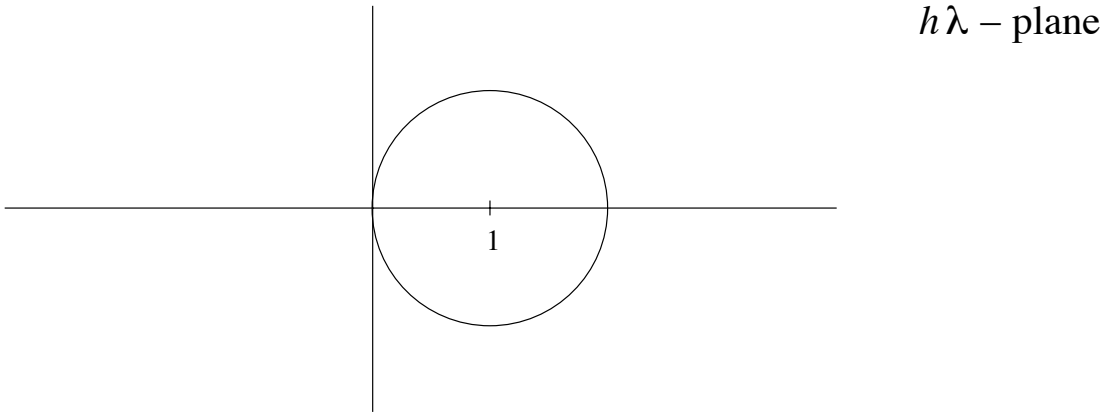
$$y' = f(y) \ , \ \frac{u_{n+1} - u_n}{h} = f(u_{n+1}) \Rightarrow u_{n+1} = u_n + h f(u_{n+1}) \ : \ \underline{\text{implicit}}$$

claim : $\tau_n = O(h^2)$, proof : hw

test equation : $y' = \lambda y$

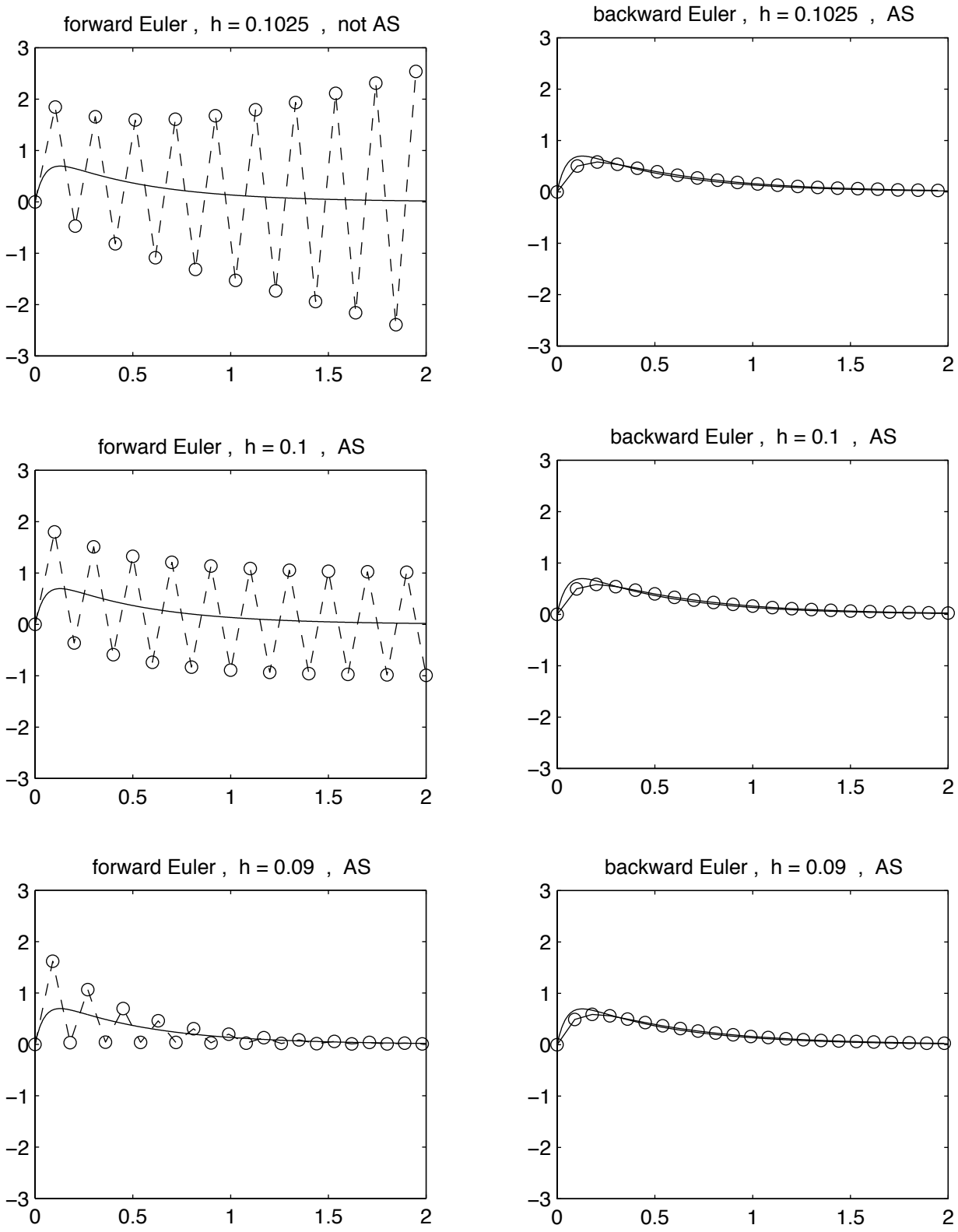
$$u_{n+1} = u_n + h\lambda u_{n+1} \Rightarrow (1 - h\lambda) u_{n+1} = u_n \Rightarrow u_{n+1} = \frac{1}{1 - h\lambda} \, u_n$$

absolute stability $\Leftrightarrow \left| \frac{1}{1 - h\lambda} \right| \leq 1 \Leftrightarrow |1 - h\lambda| \geq 1$



The region of absolute stability of backward Euler contains the left half of the $h\lambda$-plane. Hence for the previous example with $\lambda_1 = -20, \lambda_2 = -2$, backward Euler is absolutely stable for all h.

example : $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} -11 & 9 \\ 9 & -11 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}_0 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, solid line $= y_1(t)$



Even though $\lambda_1 = -20, \lambda_2 = -2$, the solution has an interval of transient growth; this is important in the context of hydrodynamic stability.

1. A scheme is A-stable if the region of absolute stability contains the left half of the $h\lambda$-plane. Backward Euler is A-stable, forward Euler is not A-stable. For an A-stable scheme, the time step h is not restricted by absolute stability and it can be chosen entirely on the basis of accuracy requirements.

2. $u_{n+1} = u_n + hf(u_{n+1})$ can be solved by iteration.

$u_{n+1}^{(0)} = u_n + hf(u_n)$, $u_{n+1}^{(m+1)} = u_n + hf(u_{n+1}^{(m)})$: predictor-corrector

Hence implicit schemes require more work per time step than explicit schemes.

3. $u_{n+1} = u_n + \frac{h}{2}\Big(f(u_n) + f(u_{n+1})\Big)$: trapezoid method , implicit

claim : $\tau_n = O(h^3)$, A-stable , proof : hw

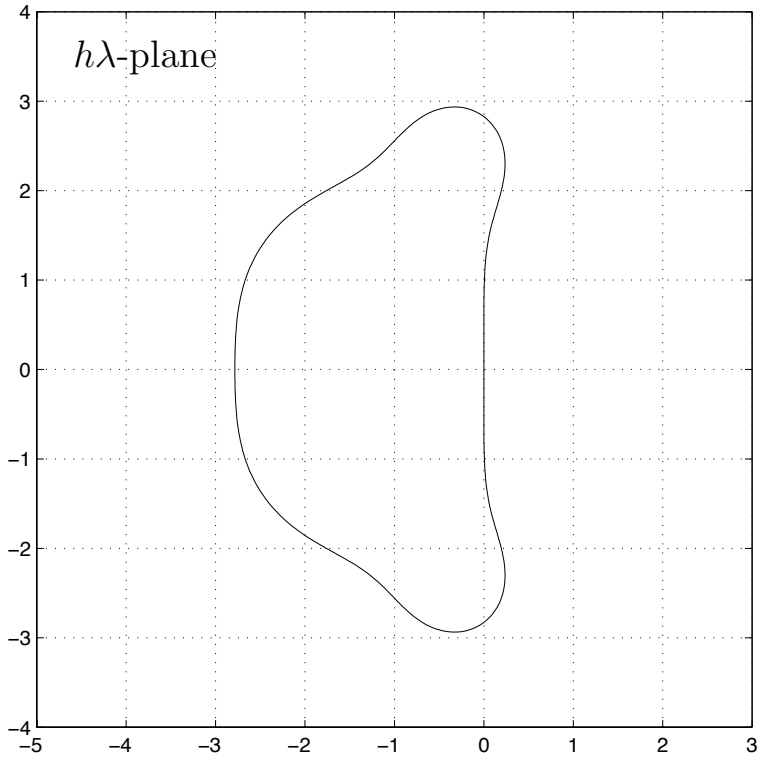
---

absolute stability of RK4

$k_1 = f(u_n)$, $k_2 = f\Big(u_n + \frac{h}{2}k_1\Big)$, $k_3 = f\Big(u_n + \frac{h}{2}k_2\Big)$, $k_4 = f(u_n + hk_3)$

$u_{n+1} = u_u + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$

$y' = \lambda y \Rightarrow u_{n+1} = \Big(1 + h\lambda + \frac{h^2\lambda^2}{2} + \frac{h^3\lambda^3}{6} + \frac{h^4\lambda^4}{24}\Big)u_n$



RK4 is not A-stable, but the region of absolute stability is larger than the region for forward Euler and it contains an interval on the imaginary axis.

multistep methods

Adams-Bashforth

$u_{n+1} = u_n + h(\beta_0 f(u_n) + \beta_1 f(u_{n-1}) + \cdots + \beta_k f(u_{n-k}))$

This is an explicit $(k+1)$-step method , $\beta_i = ?$

idea : $y' = f(y) \Rightarrow y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(y(t))\,dt$

approximate $f(y(t))$ by an interpolating polynomial $p(t)$

set $u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} p(t)\,dt$: need to find a formula for $p(t)$

$p(t_n) = f(u_n)$

$p(t_{n-1}) = f(u_{n-1})$

$\cdots$

$p(t_{n-k}) = f(u_{n-k})$

define $\nabla f_n = f_n - f_{n-1}$: backward difference

7
Thurs
2/1

$\nabla^2 f_n = \nabla(\nabla f_n) = \nabla(f_n - f_{n-1}) = \nabla f_n - \nabla f_{n-1} = f_n - f_{n-1} - (f_{n-1} - f_{n-2})$

$= f_n - 2f_{n-1} + f_{n-2}$

$\nabla^3 f_n = \cdots = f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}$

Newton form

$$p(t) = f_n + (t-t_n)\frac{\nabla f_n}{h} + (t-t_n)(t-t_{n-1})\frac{\nabla^2 f_n}{2h^2}$$

$$+ \cdots + \underbrace{(t-t_n)\cdots(t-t_{n-k+1})}_{k \text{ terms}}\frac{\nabla^k f_n}{k!\,h^k} \text{ : polynomial of degree } k$$

$$= f_n + \sum_{j=1}^{k} \underbrace{(t-t_n)\cdots(t-t_{n-j+1})}_{j \text{ terms}}\frac{\nabla^j f_n}{j!\,h^j} \ , \ k \geq 1$$

claim

1. $p(t_{n-j}) = f_{n-j}$, $j = 0 : k$

2. $f(y(t)) - p(t) = \dfrac{(t-t_n)\cdots(t-t_{n-k})}{(k+1)!}\dfrac{d^{k+1}}{dt^{k+1}} f(y(\alpha))$ for some $\alpha = \alpha(t)$

note : $p(t)$ interpolates $f(u_{n-j})$ in 1 and $f(y_{n-j})$ in 2

proof 1

$p(t_n) = f_n$

$p(t_{n-1}) = f_n + (t_{n-1} - t_n)\dfrac{\nabla f_n}{h}$

$\quad = f_n + (-h)\dfrac{\nabla f_n}{h} = f_n - \nabla f_n = f_n - (f_n - f_{n-1}) = f_{n-1}$

$p(t_{n-2}) = f_n + (t_{n-2} - t_n)\dfrac{\nabla f_n}{h} + (t_{n-2} - t_n)(t_{n-2} - t_{n-1})\dfrac{\nabla^2 f_n}{2h^2}$

$\quad = f_n + (-2h)\dfrac{\nabla f_n}{h} + (-2h)(-h)\dfrac{\nabla^2 f_n}{2h^2}$

$\quad = f_n - 2\nabla f_n + \nabla^2 f_n = f_n - 2(f_n - f_{n-1}) + (f_n - 2f_{n-1} + f_{n-2}) = f_{n-2}$

$\quad \cdots$

$$p(t_{n-k}) = f_n + \sum_{j=1}^{k} \underbrace{(t_{n-k} - t_n)(t_{n-k} - t_{n-1}) \cdots (t_{n-k} - t_{n-j+1})}_{j \text{ terms}} \frac{\nabla^j f_n}{j!\, h^j}$$

$$= f_n + \sum_{j=1}^{k} (-kh)(-(k-1)h) \cdots (-(k-j+1)h) \frac{\nabla^j f_n}{j!\, h^j}$$

$$= f_n + \sum_{j=1}^{k} (-1)^j \cdot k(k-1) \cdots (k-j+1) \cdot h^j \frac{\nabla^j f_n}{j!\, h^j} = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \nabla^j f_n$$

recall : $\dfrac{k!}{j!(k-j)!} = \binom{k}{j}$, $(a+b)^k = \displaystyle\sum_{j=0}^{k} \binom{k}{j} a^{k-j} b^j$: binomial expansion

$$p(t_{n-k}) = \sum_{j=0}^{k} \binom{k}{j} I^{k-j} (-\nabla)^j f_n = (I - \nabla)^k f_n = S_-^k f_n = f_{n-k} \quad \text{ok}$$

define $I f_n = f_n$: identity , $(I - \nabla) f_n = f_n - (f_n - f_{n-1}) = f_{n-1} = S_- f_n$: shift

---

recall : $u_{n+1} = u_n + \displaystyle\int_{t_n}^{t_{n+1}} p(t)\, dt$

$t - t_n = sh$

$t - t_{n-1} = t - t_n + t_n - t_{n-1} = sh + h = (s+1)h$

$\quad \cdots$

$t - t_{n-k+1} = \cdots = (s+k-1)h$

$$p(t) = f_n + s\nabla f_n + \frac{s(s+1)}{2}\nabla^2 f_n + \cdots + \frac{s(s+1)\cdots(s+k-1)}{k!}\nabla^k f_n$$

$$\int_{t_n}^{t_{n+1}} p(t)\, dt = \int_0^1 p(t(s))h\, ds = h\left(\gamma_0 f_n + \gamma_1 \nabla f_n + \gamma_2 \nabla^2 f_n + \cdots + \gamma_k \nabla^k f_n\right)$$

$\gamma_0 = \int_0^1 ds = 1$

$\gamma_1 = \int_0^1 s\, ds = \frac{1}{2}$

$\gamma_2 = \int_0^1 \frac{s(s+1)}{2}\, ds = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}$

$\cdots$

$\gamma_k = \int_0^1 \frac{s(s+1)\cdots(s+k-1)}{k!}\, ds \ , \ k \geq 1$

$\underline{k=0}$: 1-step AB

$u_{n+1} = u_n + h\gamma_0 f_n = u_n + hf_n$: Euler's method

$\underline{k=1}$: 2-step AB

$u_{n+1} = u_n + h(\gamma_0 f_n + \gamma_1 \nabla f_n) = u_n + h\Big(f_n + \frac{1}{2}(f_n - f_{n-1})\Big) = u_n + \frac{h}{2}(3f_n - f_{n-1})$

$\underline{k=2}$: 3-step AB

$u_{n+1} = u_n + h(\gamma_0 f_n + \gamma_1 \nabla f_n + \gamma_2 \nabla^2 f_n) = \cdots = u_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2})$

<u>note</u>

1. To get started we set $u_0 = y_0$, but we also need to compute $u_1, u_2, \ldots, u_k$.
2. To compute u_{n+1} from u_n, only 1 function evaluation $f_n = f(u_n)$ is needed.
3. Changing the step size h requires extra work.

<u>proof</u> 2

goal : $f(y(t)) - p(t) = \frac{(t-t_n)\cdots(t-t_{n-k})}{(k+1)!} \frac{d^{k+1}}{dt^{k+1}} f(y(\alpha))$

if $t = t_{n-j}, j = 0:k$, then <u>ok</u>, so assume $t \neq t_{n-j}$, $j = 0:k$

set $g(x) = f(y(x)) - p(x) + \frac{(x-t_n)\cdots(x-t_{n-k})}{(t-t_n)\cdots(t-t_{n-k})}(\, p(t) - f(y(t))\,)$

$g(t_n) = 0$, $g(t_{n-1}) = 0$, $\ldots$, $g(t_{n-k}) = 0$, $g(t) = 0$

$\Rightarrow$ $g(x)$ has $k+2$ distinct roots

$\Rightarrow$ $g'(x)$ has $k+1$ distinct roots , MVT : Math 451

$\cdots$

$\Rightarrow$ $g^{(k+1)}(x)$ has 1 root , say $g^{(k+1)}(\alpha) = 0$

then $0 = g^{(k+1)}(\alpha) = \frac{d^{k+1}}{dx^{k+1}} f(y(\alpha)) + \frac{(k+1)!(\, p(t) - f(y(t))\,)}{(t-t_n)\cdots(t-t_{n-k})}$ <u>ok</u>

8
Tues
2/6

local truncation error : $(k+1)$-step AB

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} p(t)\,dt + \tau_n$$

$$\tau_n = y_{n+1} - y_n - \int_{t_n}^{t_{n+1}} p(t)\,dt = \int_{t_n}^{t_{n+1}} (\, f(y(t) - p(t)\,)\,dt$$

$$= \int_{t_n}^{t_{n+1}} \frac{(t-t_n)\cdots(t-t_{n-k})}{(k+1)!}\frac{d^{k+1}}{dt^{k+1}} f(y(\alpha))\,dt \ , \ t - t_n = sh$$

$$= \int_0^1 \frac{s(s+1)\cdots(s+k)h^{k+1}}{(k+1)!} y^{(k+2)}(\alpha)\,h\,ds \ = \ \gamma_{k+1}\, y^{(k+2)}(\tilde{\alpha})\,h^{k+2}$$

Adams-Moulton

$$u_{n+1} = u_n + h(\beta_{-1}^* f(u_{n+1}) \ + \ \beta_0^* f(u_n) \ + \ \cdots \ + \ \beta_k^* f(u_{n-k}))$$

This is an implicit $(k+1)$-step method , $\beta_i^* = ?$

$p^*(t)$ interpolates f_{n+1} , f_n , $\ldots$, f_{n-k} : $k+2$ values , polynomial of degree $k+1$

$$p^*(t) \ = \ f_{n+1} \ + \ (t-t_{n+1})\frac{\nabla f_{n+1}}{h} \ + \ (t-t_{n+1})(t-t_n)\frac{\nabla^2 f_{n+1}}{2h^2}$$

$$+ \cdots + \underbrace{(t-t_{n+1})\cdots(t-t_{n-k+1})}_{k+1 \text{ terms}}\frac{\nabla^{k+1} f_{n+1}}{(k+1)!\,h^{k+1}}$$

$$t - t_n = sh \ \Rightarrow \ t - t_{n+1} \ = \ t - t_n + t_n - t_{n+1} = sh - h = (s-1)h$$

$$p^*(t) = f_{n+1} \ + \ (s-1)\nabla f_{n+1} \ + \ \frac{(s-1)s}{2}\nabla^2 f_{n+1}$$

$$+ \cdots + \ \frac{(s-1)\cdots(s+k-1)}{(k+1)!}\nabla^{k+1} f_{n+1}$$

$$u_{n+1} = u_n \ + \ \int_{t_n}^{t_{n+1}} p^*(t)\,dt = \int_0^1 p^*(t(s))h\,ds$$

$$u_{n+1} = u_n \ + \ h\left(\gamma_{-1}^* f_{n+1} \ + \ \gamma_0^*\,\nabla f_{n+1} \ + \ \cdots \ + \ \gamma_k^*\,\nabla^{k+1} f_{n+1}\right)$$

$$\gamma_k^* = \int_0^1 \frac{(s-1)s\cdots(s+k-1)}{(k+1)!}\,ds \ , \ k \ \geq \ 0$$

$$\gamma_{-1}^* = \int_0^1 ds = 1$$

$$\gamma_0^* = \int_0^1 (s-1)\,ds = -\frac{1}{2}$$

$$\gamma_1^* = \int_0^1 \frac{(s-1)s}{2}\,ds = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12}$$

$\underline{k=-1}$

$u_{n+1} = u_n + h\gamma^*_{-1} f_{n+1} = u_n + h f_{n+1}$: backward Euler

$\underline{k=0}$: 1-step AM

$$u_{n+1} = u_n + h\left(\gamma^*_{-1} f_{n+1} + \gamma^*_0 \nabla f_{n+1}\right) = u_n + h\Big(f_{n+1} - \frac{1}{2}(f_{n+1} - f_n)\Big)$$

$$u_{n+1} = u_n + \frac{h}{2}\Big(f_n + f_{n+1}\Big) \ : \text{ trapezoid method}$$

$\underline{k=1}$: 2-step AM

$$u_{n+1} = u_n + h(\gamma^*_{-1} f_{n+1} + \gamma^*_0 \nabla f_{n+1} + \gamma^*_1 \nabla f_{n+1})$$

$$= u_n + h\Big(f_{n+1} - \frac{1}{2}(f_{n+1} - f_n) - \frac{1}{12}(f_{n+1} - 2f_n + f_{n-1})\Big)$$

$$u_{n+1} = u_n + \frac{h}{12}\Big(5f_{n+1} + 8f_n - f_{n-1}\Big)$$

summary

AB

$u_{n+1} = u_n + h\left(\gamma_0 f_n + \gamma_1 \nabla f_n + \cdots + \gamma_k \nabla^k f_n\right)$: $(k+1)$-step

$$\gamma_k = \int_0^1 \frac{s(s+1)\cdots(s+k-1)}{k!}\, ds \ , \ k \geq 1 \ , \ \gamma_0 = 1$$

$$\tau_n = \gamma_{k+1}\, y^{(k+2)}(t)\, h^{k+2} + O(h^{k+3})$$

AM

$u_{n+1} = u_n + h\left(\gamma^*_{-1} f_{n+1} + \gamma^*_0 \nabla f_{n+1} + \cdots + \gamma^*_k \nabla^{k+1} f_{n+1}\right)$: $(k+1)$-step

$$\gamma^*_k = \int_0^1 \frac{(s-1)s\cdots(s+k-1)}{(k+1)!}\, ds \ , \ k \geq 0 \ , \ \gamma^*_{-1} = 1$$

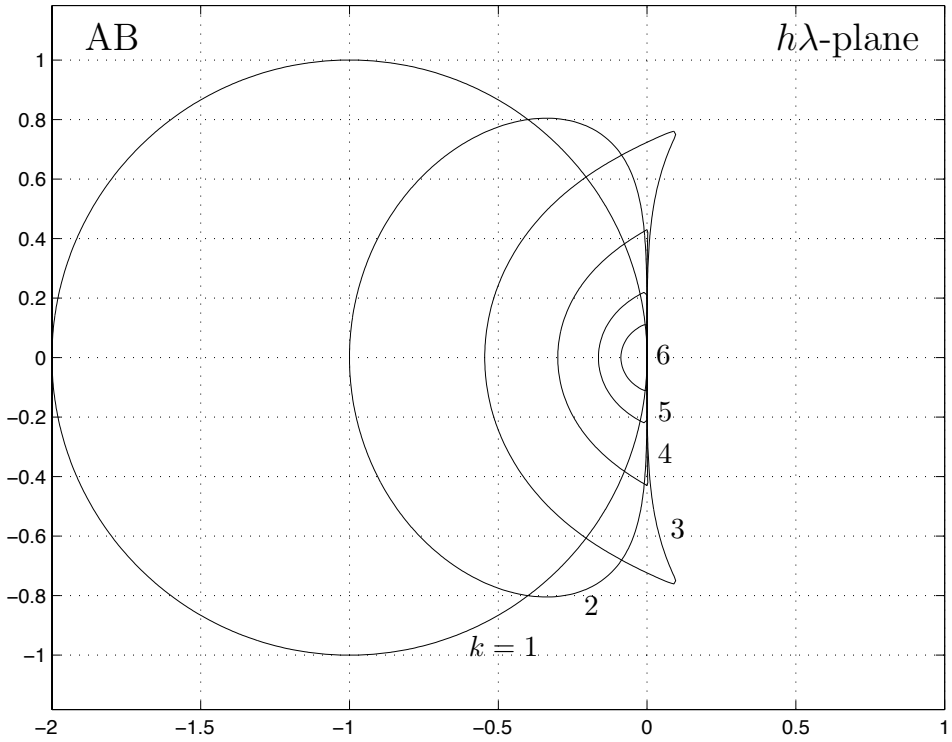
$$\tau_n = \gamma^*_{k+1}\, y^{(k+3)}(t)\, h^{k+3} + O(h^{k+4})$$

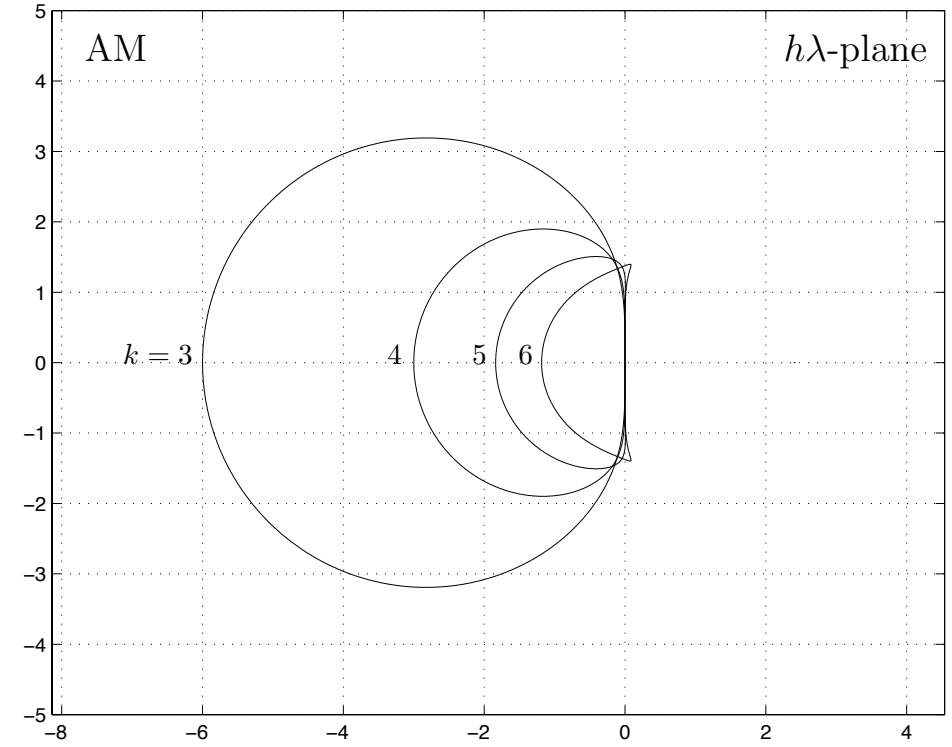
note

k-step AB has global order of accuracy k

k-step AM ..... " ..... $k+1$

region of absolute stability





The method of order k is absolutely stable inside the contour. Note the difference in scale; the AB regions are smaller than the AM regions. Higher order methods have smaller regions of absolute stability than lower order methods. The 1st order and 2nd order AM methods are not shown because they are A-stable.

general multistep methods

$$\alpha_0 u_n + \alpha_1 u_{n-1} + \cdots + \alpha_k u_{n-k} + h(\beta_0 f(u_n) + \beta_1 f(u_{n-1}) + \cdots + \beta_k f(u_{n-k})) = 0$$

: k-step method , assume $\alpha_0 \neq 0$, AB and AM are special cases

$$\sum_{i=0}^{k} (\alpha_i u_{n-i} + h\beta_i f(u_{n-i})) = 0$$

predictor : $\beta_0 = 0$

$$\alpha_0 u_n^{(0)} = -\sum_{i=1}^{k} (\alpha_i u_{n-i} + h\beta_i f(u_{n-i}))$$

corrector : $\beta_0 \neq 0$

$$\alpha_0 u_n^{(m+1)} = -h\beta_0 f(u_n^{(m)}) - \sum_{i=1}^{k} (\alpha_i u_{n-i} + h\beta_i f(u_{n-i}))$$

local truncation error

$$\sum_{i=0}^{k} (\alpha_i y_{n-i} + h\beta_i f(y_{n-i})) = \tau_n = O(h^{r+1})$$

$$y_{n-i} = y(t - ih) = \sum_{j=0}^{r+1} \frac{y^{(j)}(t)}{j!}(-ih)^j + O(h^{r+2})$$

$$\tau_n = \sum_{i=0}^{k} (\alpha_i y_{n-i} + h\beta_i y'_{n-i})$$

$$= \sum_{i=0}^{k} \left(\alpha_i \sum_{j=0}^{r+1} \frac{y^{(j)}(t)}{j!}(-ih)^j + h\beta_i \sum_{j=0}^{r} \frac{y^{(j+1)}(t)}{j!}(-ih)^j \right) + O(h^{r+2})$$

$$= \sum_{i=0}^{k} \left(\alpha_i \sum_{j=0}^{r+1} \frac{y^{(j)}(t)}{j!}(-ih)^j + h\beta_i \sum_{j=1}^{r+1} \frac{y^{(j)}(t)}{(j-1)!}(-ih)^{(j-1)} \right) + O(h^{r+2})$$

$$\tau_n = \sum_{j=0}^{r+1} C_j y^{(j)}(t) h^j + O(h^{r+2})$$

$$C_0 = \sum_{i=0}^{k} \alpha_i \ , \ C_j = \sum_{i=0}^{k} \left(\frac{(-i)^j \alpha_i}{j!} + \frac{(-i)^{j-1} \beta_i}{(j-1)!} \right) \ , \ j \geq 1$$

If $C_0 = \cdots = C_r = 0$ and $C_{r+1} \neq 0$, then $\tau_n = C_{r+1} y^{(r+1)}(t) h^{r+1} + O(h^{r+2})$ and we have a k-step method of order r.

9
Thurs
2/8

example

$y' = f(y)$

$\frac{u_{n+1} - u_{n-1}}{2h} = f(u_n)$: leap-frog method

$u_n - u_{n-2} - 2hf(u_{n-1}) = 0$: explicit , 2-step , needs u_0, u_1 to start

$\alpha_0 = 1$, $\alpha_1 = 0$, $\alpha_2 = -1$, $\beta_0 = 0$, $\beta_1 = -2$, $\beta_2 = 0$

$C_0 = \sum_{i=0}^{2} \alpha_i = \alpha_0 + \alpha_1 + \alpha_2 = 0$

$C_1 = \sum_{i=0}^{2} (-i\alpha_i + \beta_i) = -\cancel{\alpha_1} - 2\alpha_2 + \cancel{\beta_0} + \beta_1 + \cancel{\beta_2} = 0$

$C_2 = \sum_{i=0}^{2} \left(\frac{i^2 \alpha_i}{2} - i\beta_i \right) = \cancel{\frac{\alpha_1}{2}} + 2\alpha_2 - \beta_1 - \cancel{2\beta_2} = 0$

$C_3 = \sum_{i=0}^{2} \left(\frac{-i^3 \alpha_i}{6} + \frac{i^2 \beta_i}{2} \right) = \cancel{\frac{-\alpha_1}{6}} - \frac{4}{3}\alpha_2 + \frac{1}{2}\beta_1 + \cancel{2\beta_2} = \frac{4}{3} - 1 = \frac{1}{3}$

$\tau_n = \frac{1}{3} y^{(3)}(t) h^3 + O(h^4) \Rightarrow$ leap-frog is 2nd order accurate

---

question : What is the maximum order of a k-step multistep scheme?

$\alpha_0, \ldots, \alpha_k, \beta_0, \ldots, \beta_k$: $2k + 2$ coefficients

$\Rightarrow$ $2k + 1$ degrees of freedom

$\Rightarrow$ $C_0 = \cdots = C_{2k} = 0$, $C_{2k+1} \neq 0$

$\Rightarrow$ $\tau_n = O(h^{2k+1})$ is optimal

$\Rightarrow$ the maximum order of a k-step multistep scheme is $2k$

example

$k = 1$: trapezoid method has order 2

$k = 2$: 2-step AB has order 2 , 2-step AM has order 3

In fact there exists a 2-step scheme of order 4. (more later)

note : RK4 is a 1-step scheme of order 4, but there is no contradiction because it is not a multistep scheme of the form considered here.

characteristic polynomials of a multistep scheme

$\alpha_0 u_n + \alpha_1 u_{n-1} + \cdots + \alpha_k u_{n-k} + h(\beta_0 f(u_n) + \beta_1 f(u_{n-1}) + \cdots + \beta_k f(u_{n-k})) = 0$

definition

$\rho(\zeta) = \alpha_0 \zeta^k + \alpha_1 \zeta^{k-1} + \cdots + \alpha_k = \sum_{i=0}^{k} \alpha_i \zeta^{k-i}$

$\sigma(\zeta) = \beta_0 \zeta^k + \beta_1 \zeta^{k-1} + \cdots + \beta_k = \sum_{i=0}^{k} \beta_i \zeta^{k-i}$

note

$\rho(1) = \sum_{i=0}^{k} \alpha_i = C_0 \ , \ \sigma(1) = \sum_{i=0}^{k} \beta_i$

$\rho'(1) = \sum_{i=0}^{k} (k-i)\alpha_i = k \sum_{i=0}^{k} \alpha_i + \sum_{i=0}^{k} (-i\alpha_i + \beta_i) - \sum_{i=0}^{k} \beta_i = kC_0 + C_1 - \sigma(1)$

recall : a method is consistent $\Leftrightarrow \tau_n = O(h^{r+1})$ for some $r \geq 1$

$\Leftrightarrow C_0 = C_1 = 0 \Leftrightarrow \rho(1) = 0 \ , \ \rho'(1) + \sigma(1) = 0$

example : leap-frog

$u_n - u_{n-2} - 2hf(u_{n-1}) = 0 \ \Rightarrow \ \rho(\zeta) = \zeta^2 - 1 \ , \ \sigma(\zeta) = -2\zeta$

$\rho(1) = 0 \ , \ \rho'(1) + \sigma(1) = 0 \ \Rightarrow$ leap-frog is consistent

---

question : When is a multistep scheme stable? convergent? absolutely stable?

special case : $y' = 0$, test equation $y' = \lambda y$ with $\lambda = 0$

$\sum_{i=0}^{k} \alpha_i u_{n-i} = 0$: difference equation, linear, constant coefficient, homogeneous

$\alpha_0 u_n + \alpha_1 u_{n-1} + \cdots + \alpha_k u_{n-k} = 0 \ \ (*)$

given $u_0, u_1, \ldots, u_{k-1}$, use $(*)$ to solve for $u_k, u_{k+1}, \ldots$

note : If $u_n = \zeta^n$, then $\sum_{i=0}^{k} \alpha_i u_{n-i} = \sum_{i=0}^{k} \alpha_i \zeta^{n-i} = \zeta^{n-k} \sum_{i=0}^{k} \alpha_i \zeta^{k-i} = \zeta^{n-k} \rho(\zeta)$.

Hence if $\rho(\zeta) = 0$, then $u_n = \zeta^n$ is a solution of $(*)$.

theorem : If $\rho(\zeta)$ has j distinct roots $\zeta_1, \ldots, \zeta_j$ with multiplicities $m_1, \ldots, m_j$, then the solution of the difference equation $(*)$ has the form

$u_n = (a_{10} + a_{11} n + a_{12} n^2 + \cdots + a_{1,m_1-1} n^{m_1-1}) \zeta_1^n$

$+ \cdots + \ (a_{j0} + a_{j1} n + a_{j2} n^2 + \cdots + a_{j,m_j-1} n^{m_j-1}) \zeta_j^n,$

where $a_{10}, \ldots, a_{j,m_j-1}$ are determined by the initial data $u_0, \ldots, u_{k-1}$.

note : $\zeta_1, \ldots, \zeta_j$ are the characteristic roots of the difference equation $(*)$ and $\rho(\zeta) = \alpha_0 (\zeta - \zeta_1)^{m_1} \cdots (\zeta - \zeta_j)^{m_j}$

example

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

$\rho(\zeta) = \zeta^2 + 4\zeta - 5 = (\zeta - 1)(\zeta + 5) \Rightarrow \zeta_1 = 1\ ,\ \zeta_2 = -5\ ,\ m_1 = m_2 = 1$

$u_n = a_1\zeta_1^n + a_2\zeta_2^n = a_1 + a_2(-5)^n$, check ...

2. $u_n - 2u_{n-1} + u_{n-2} = 0$

$\rho(\zeta) = \zeta^2 - 2\zeta + 1 = (\zeta - 1)^2 \Rightarrow \zeta_1 = 1\ ,\ m_1 = 2$

$u_n = (a_1 + a_2 n)\zeta_1^n = a_1 + a_2 n$, check ...

There is an analogy between difference equations and differential equations.

$\alpha_0 y^{(k)} + \alpha_1 y^{(k-1)} + \cdots + \alpha_k y = 0\ ,\ y(t) = e^{\lambda t} \Rightarrow \rho(\lambda) = 0$

1. $y'' + 4y' - 5y = 0 \Rightarrow y(t) = a_1 e^t + a_2 e^{-5t}$

2. $y'' - 2y' + y = 0 \Rightarrow y(t) = (a_1 + a_2 t)e^t$

---

proof (sketch)

10
Tues
2/13

$(*) \Rightarrow u_n = -\frac{\alpha_1}{\alpha_0}u_{n-1} - \cdots - \frac{\alpha_k}{\alpha_0}u_{n-k} \Rightarrow u_k = -\frac{\alpha_1}{\alpha_0}u_{k-1} - \cdots - \frac{\alpha_k}{\alpha_0}u_0$

$$\begin{pmatrix} u_k \\ u_{k-1} \\ \vdots \\ u_1 \end{pmatrix} = \begin{pmatrix} -\frac{\alpha_1}{\alpha_0} & \cdots & \cdots & -\frac{\alpha_k}{\alpha_0} \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix} \begin{pmatrix} u_{k-1} \\ u_{k-2} \\ \vdots \\ u_0 \end{pmatrix}$$

$U_1 = AU_0 \Rightarrow U_n = AU_{n-1} = \cdots = A^n U_0$

$A = XJX^{-1}$: Jordan form $\Rightarrow A^n = XJ^nX^{-1}$

$\det(A - \lambda I) = \frac{(-1)^k}{\alpha_0}\rho(\lambda) = (-1)^k(\lambda - \zeta_1)^{m_1} \cdots (\lambda - \zeta_j)^{m_j}$

$$J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_j \end{pmatrix}\ ,\ J_i = \begin{pmatrix} \zeta_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \zeta_i \end{pmatrix}\ :\ m_i \times m_i\ \ldots \quad \underline{\text{ok}}$$

note

the difference scheme $(*)$ is absolutely stable

$\Leftrightarrow$ every solution u_n of $(*)$ is bounded as $n \to \infty$

$\Leftrightarrow$ $\rho(\zeta)$ satisfies the root condition : $\begin{cases} |\zeta_i| \le 1 \text{ for } i = 1 : j \\ \text{and} \\ \text{if } |\zeta_i| = 1, \text{ then } m_i = 1 \end{cases}$

proof : $\lim_{n\to\infty} n^p |\zeta|^n = \begin{cases} 0 & \text{if } |\zeta| < 1 \\ 1 & \text{if } |\zeta| = 1 \ , \ p = 0 \\ \infty & \text{if } |\zeta| > 1 \text{ or } |\zeta| = 1 \ , \ p \ge 1 \end{cases}$

example

$\rho(\zeta)$	root condition?	u_n
$\zeta^2 + 4\zeta - 5$	no	$a_1 + a_2(-5)^n$
$\zeta^2 - 2\zeta + 1$	no	$a_1 + a_2 n$
$\zeta^2 - 1$	yes	$a_1 + a_2(-1)^n$

---

test equation : $y' = \lambda y$

$$\sum_{i=0}^{k} (\alpha_i u_{n-i} + h\beta_i f(u_{n-i})) = 0 \Rightarrow \sum_{i=0}^{k} (\alpha_i + h\lambda\beta_i) u_{n-i} = 0$$

$$(\alpha_0 + h\lambda\beta_0)u_n + (\alpha_1 + h\lambda\beta_1)u_{n-1} + \cdots + (\alpha_k + h\lambda\beta_k)u_{n-k} = 0 \quad (**)$$

If $u_n = \zeta^n$, then $\sum_{i=0}^{k} (\alpha_i + h\lambda\beta_i) u_{n-i} = \cdots = \zeta^{n-k}(\rho(\zeta) + h\lambda\sigma(\zeta))$.

Hence if $\rho(\zeta) + h\lambda\sigma(\zeta) = 0$, then $u_n = \zeta^n$ is a solution of $(**)$.

theorem (existence of the principal root)

1. If $\rho(1) = 0$, $\rho'(1) \neq 0$, then $\rho(\zeta) + h\lambda\sigma(\zeta)$ has a root $\zeta_1(h)$ st $\lim_{h\to 0} \zeta_1(h) = 1$.

2. If in addition $\tau_n = O(h^{r+1})$, then $\zeta_1(h) = e^{h\lambda} + O(h^{r+1})$.

note

1. $\rho(1) = 0$, $\rho'(1) \neq 0 \Rightarrow$ $\rho(\zeta)$ has a simple root at $\zeta = 1$
This holds if the scheme is consistent and $\rho(\zeta)$ satisfies the root condition.

2. $\zeta_1(h)$ is called the principal root of the difference equation $(**)$; the other roots of $\rho(\zeta) + h\lambda\sigma(\zeta)$ are denoted $\zeta_2(h)$, ... and are called the extraneous roots.

3. Under the assumptions of the thm, the difference equation $(**)$ has a solution of the form $u_n = \zeta_1(h)^n = (e^{h\lambda} + O(h^{r+1}))^n = e^{\lambda t} + O(h^r)$, where $t = nh$; this justifies calling $\zeta_1(h)$ the principal root.

$(a+b)^n = a^n + na^{n-1}b + \frac{1}{2}n(n-1)a^{n-2}b^2 + \cdots$

$(e^{h\lambda} + O(h^{r+1}))^n = (e^{h\lambda})^n + n(e^{h\lambda})^{n-1}O(h^{r+1}) + \frac{1}{2}n(n-1)(e^{h\lambda})^{n-2}O(h^{r+1})^2 + \cdots$

example : $y' = \lambda y$, $y(0) = 1 \Rightarrow y(t) = e^{\lambda t}$

2-step AB : $u_{n+1} = u_n + \frac{1}{2}h(3f_n - f_{n-1})$, $u_0 = 1$, $u_1 = e^{h\lambda}$

$u_n - u_{n-1} - \frac{1}{2}h\lambda(3u_{n-1} - u_{n-2}) = 0$

$\rho(\zeta) = \zeta^2 - \zeta$, $\zeta = \{0, 1\}$, $\rho(1) = 0$, $\rho'(1) \neq 0$, $\tau_n = O(h^3)$, thm applies

$\rho(\zeta) + h\lambda\sigma(\zeta) = \zeta^2 - \zeta - \frac{1}{2}h\lambda(3\zeta - 1) = \zeta^2 - (1 + \frac{3}{2}h\lambda)\zeta + \frac{1}{2}h\lambda = 0$

$\zeta = \frac{1}{2}(1 + \frac{3}{2}h\lambda \pm (1 + \frac{9}{4}h^2\lambda^2 + 3h\lambda - 2h\lambda)^{1/2}) = \frac{1}{2} + \frac{3}{4}h\lambda \pm \frac{1}{2}(1 + h\lambda + \frac{9}{4}h^2\lambda^2)^{1/2}$

$(1 + x)^{1/2} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + O(x^3)$

$\zeta_1(h) = \cancel{\frac{1}{2}} + \frac{3}{4}\cancel{h\lambda} + \frac{1}{2}(\cancel{1} + \frac{1}{2}(\cancel{h\lambda} + \frac{9}{4}h^2\lambda^2) - \frac{1}{8}(h\lambda + \frac{9}{4}h^2\lambda^2)^2 + O(h^3))$

$= 1 + h\lambda + \frac{1}{2}h^2\lambda^2 + O(h^3) = e^{h\lambda} + O(h^3)$

$\zeta_2(h) = \frac{1}{2} + \frac{3}{4}h\lambda - \frac{1}{2}(1 + h\lambda + \frac{9}{4}h^2\lambda^2)^{1/2} = \frac{1}{2}h\lambda + O(h^2) = O(h)$

11 Thurs 2/15

$u_n = a_1\zeta_1(h)^n + a_2\zeta_2(h)^n$

$\zeta_1(h)^n = (e^{h\lambda} + O(h^3))^n = e^{\lambda t} + O(h^2)$, $\zeta_2(h)^n = O(h^n)$

$$\begin{pmatrix} 1 & 1 \\ \zeta_1 & \zeta_2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ e^{h\lambda} \end{pmatrix} \Rightarrow \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \frac{1}{\zeta_2 - \zeta_1}\begin{pmatrix} \zeta_2 & -1 \\ -\zeta_1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ e^{h\lambda} \end{pmatrix} = \frac{1}{\zeta_1 - \zeta_2}\begin{pmatrix} e^{h\lambda} - \zeta_2 \\ \zeta_1 - e^{h\lambda} \end{pmatrix}$$

$$a_1 = \frac{e^{h\lambda} - \zeta_2}{\zeta_1 - \zeta_2} = \frac{e^{h\lambda} - \zeta_1 + \zeta_1 - \zeta_2}{\zeta_1 - \zeta_2} = 1 + O(h^3) \ , \ a_2 = \frac{\zeta_1 - e^{h\lambda}}{\zeta_1 - \zeta_2} = O(h^3)$$

$u_n = (1 + O(h^3)) \cdot (e^{\lambda t} + O(h^2)) + O(h^3) \cdot O(h^n) \Rightarrow \lim_{h\to 0} u_n = e^{\lambda t}$

---

leap-frog : $u_{n+1} = u_{n-1} + 2hf_n$, $u_0 = 1$, $u_1 = e^{h\lambda}$

$u_n - u_{n-2} - 2h\lambda u_{n-1} = 0$

$\rho(\zeta) = \zeta^2 - 1$, $\zeta = \{\pm 1\}$, $\rho(1) = 0$, $\rho'(1) \neq 0$, $\tau_n = O(h^3)$, thm applies

$\rho(\zeta) + h\lambda\sigma(\zeta) = \zeta^2 - 1 - 2h\lambda\zeta = \zeta^2 - 2h\lambda\zeta - 1 = 0$

$\zeta = \frac{1}{2}(2h\lambda \pm (4h^2\lambda^2 + 4)^{1/2})$

$\zeta_1(h) = h\lambda + (1 + h^2\lambda^2)^{1/2} = h\lambda + 1 + \frac{1}{2}h^2\lambda^2 + O(h^4) = e^{h\lambda} + O(h^3)$

$\zeta_2(h) = h\lambda - (1 + h^2\lambda^2)^{1/2} = h\lambda - (1 + \frac{1}{2}h^2\lambda^2 + O(h^4)) = -e^{-h\lambda} + O(h^3)$

$u_n = a_1\zeta_1(h)^n + a_2\zeta_2(h)^n$

$\zeta_1(h)^n = (e^{h\lambda} + O(h^3))^n = e^{\lambda t} + O(h^2)$

$\zeta_2(h)^n = (-e^{-h\lambda} + O(h^3))^n = (-1)^n e^{-\lambda t} + O(h^2)$

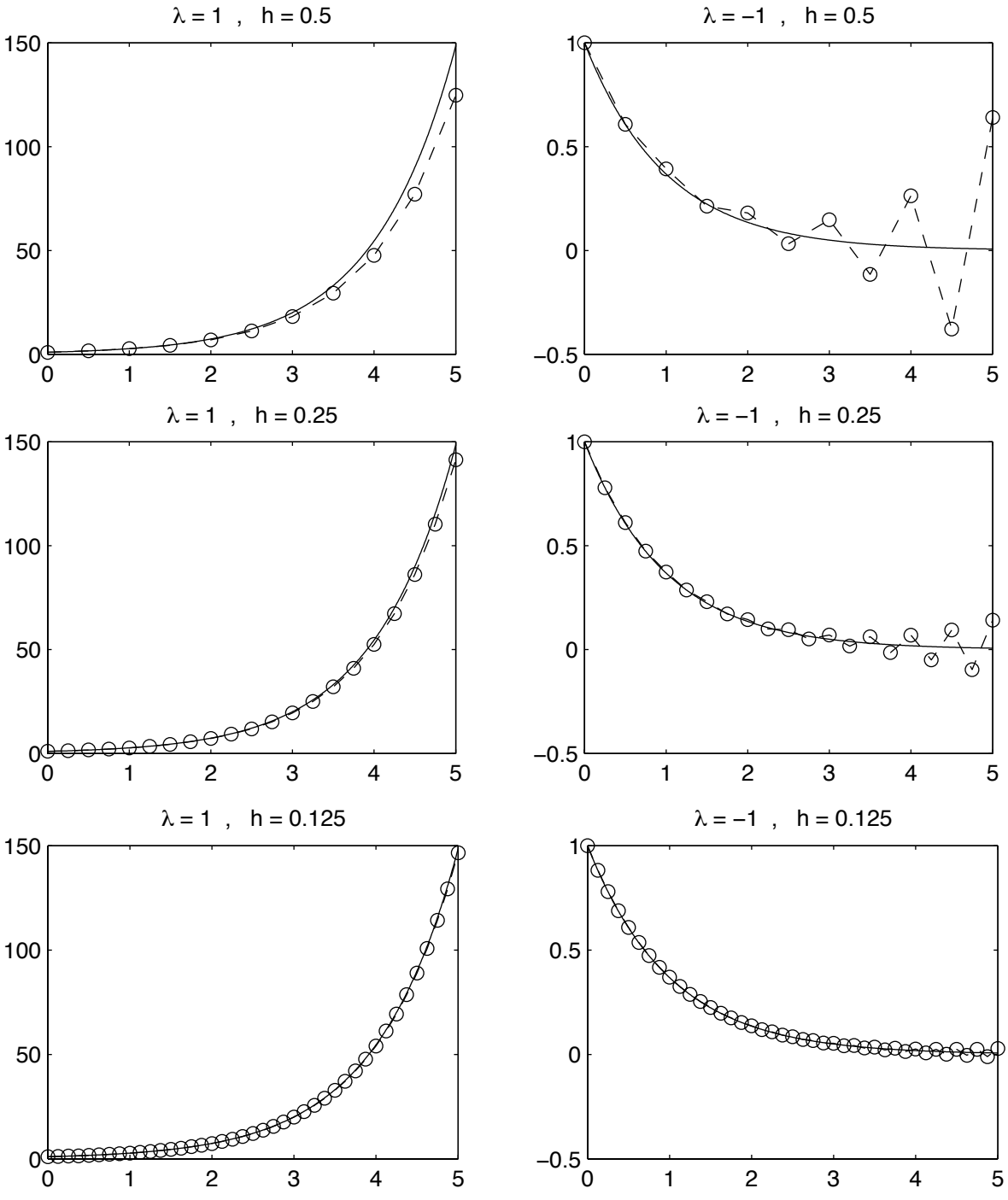
$a_1 = 1 + O(h^3)$, $a_2 = O(h^3)$

$u_n = (1 + O(h^3)) \cdot (e^{\lambda t} + O(h^2)) + O(h^3) \cdot ((-1)^n e^{-\lambda t} + O(h^2)) \Rightarrow \lim_{h\to 0} u_n = e^{\lambda t}$

note : if h is fixed and $n \to \infty$, then $O(h^3)(-1)^n e^{-\lambda t} \to \begin{cases} 0 & \text{if real}(\lambda) > 0 \\ \pm\infty & \text{if real}(\lambda) < 0 \end{cases}$

This is an example of weak instability which occurs when $\rho(\zeta)$ has two distinct roots on the unit circle; in this case the roots of $\rho(\zeta) + h\lambda\sigma(\zeta)$ may lie inside or outside the unit circle. For the leap-frog method with real$(\lambda) > 0$, the extraneous root $\zeta_2(h)$ lies inside the unit circle and is harmless, but with real$(\lambda) < 0$, $\zeta_2(h)$ lies outside the unit circle and $a_2\zeta_2(h)^n$ eventually dominates $a_1\zeta_1(h)^n$, even though $a_2 = O(h^3)$.

example : $y' = \lambda y$, $y(0) = 1$, $u_{n+1} = u_{n-1} + 2h\lambda u_n$, $u_0 = 1$, $u_1 = e^{h\lambda}$



absolute stability of leap-frog

$y' = \lambda y$, $u_{n+1} = u_{n-1} + 2h\lambda u_n$, what do we know already?

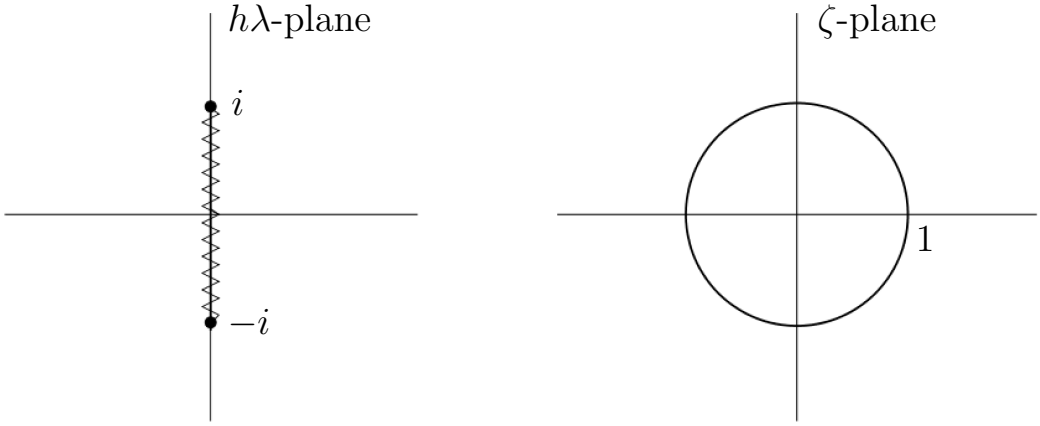
$\rho(\zeta) = \zeta^2 - 1 = 0 \Rightarrow \zeta = \pm 1$: root condition is satisfied

$\rho(\zeta) + h\lambda\sigma(\zeta) = \zeta^2 - 1 - 2h\lambda\zeta = \zeta^2 - 2h\lambda\zeta - 1 = (\zeta - \zeta_1)(\zeta - \zeta_2) = 0$

$\zeta = h\lambda \pm \left(1 + h^2\lambda^2\right)^{1/2} \Rightarrow \zeta_1 = e^{h\lambda} + O(h^3)$, $\zeta_2 = -e^{-h\lambda} + O(h^3)$

$h\lambda = \dfrac{\zeta^2 - 1}{2\zeta} = \dfrac{1}{2}\left(\zeta - \dfrac{1}{\zeta}\right)$: conformal mapping

12
Tues
2/20



1. There are branch points at $h\lambda = \pm i$ and a branch cut running between them.

2. The branch points $h\lambda = \pm i$ are fixed by the mapping and every other point in the $h\lambda$-plane corresponds to two points in the ζ-plane, e.g. $h\lambda = 0 \Rightarrow \zeta = \pm 1$.

3. $h\lambda = iy$, $|y| \le 1 \Rightarrow \zeta = iy \pm \left(1 - y^2\right)^{1/2}$: unit circle in ζ-plane

$\zeta = e^{i\theta}$, $|\theta| < \pi \Rightarrow h\lambda = \frac{1}{2}\left(e^{i\theta} - e^{-i\theta}\right) = i\sin\theta$: imaginary interval in $h\lambda$-plane

4. For any $h\lambda$ off the interval, the corresponding ζ_1, ζ_2 are off the circle; in fact since $\zeta_1\zeta_2 = -1$, one root lies inside the circle and the other lies outside. Hence the region of absolute stability of the leap-frog method is the interval in the $h\lambda$-plane between i and $-i$, excluding $\pm i$ (why?).

---

recall : theorem (existence of the principal root)

1. If $\rho(1) = 0$, $\rho'(1) \neq 0$, then $\rho(\zeta) + h\lambda\sigma(\zeta)$ has a root $\zeta_1(h)$ st $\lim_{h\to 0} \zeta_1(h) = 1$.

2. If in addition $\tau_n = O(h^{r+1})$, then $\zeta_1(h) = e^{h\lambda} + O(h^{r+1})$.

proof

1. set $F(\zeta, h) = \rho(\zeta) + h\lambda\sigma(\zeta) = 0$, $F(1,0) = \rho(1) = 0$, $F_\zeta(1,0) = \rho'(1) \neq 0$

implicit function thm $\Rightarrow$ there exists $\zeta_1(h)$ st $F(\zeta_1(h), h) = 0$, $\lim_{h\to 0} \zeta_1(h) = 1$

note : $F(\zeta, h) = F(1,0) + F_\zeta(1,0)(\zeta - 1) + F_h(1,0)h + \cdots = 0$

$\Rightarrow \zeta_1(h) = 1 - \dfrac{F_h(1,0)}{F_\zeta(1,0)}h + \cdots$

2. set $y(t) = e^{\lambda t} = e^{\lambda nh}$

$$\tau_n = \sum_{i=0}^{k} (\alpha_i + h\lambda\beta_i) y_{n-i} = \sum_{i=0}^{k} (\alpha_i + h\lambda\beta_i) e^{\lambda(n-i)h}$$

$$= e^{\lambda(n-k)h} \sum_{i=0}^{k} (\alpha_i + h\lambda\beta_i) e^{\lambda(k-i)h} = e^{\lambda(n-k)h} \left(\rho(e^{h\lambda}) + h\lambda\sigma(e^{h\lambda}) \right)$$

$$\Rightarrow \ \rho(e^{h\lambda}) + h\lambda\sigma(e^{h\lambda}) = O(h^{r+1})$$

set $\zeta_1(h) = e^{h\lambda} + \epsilon$, then $\lim_{h\to 0} \epsilon = \lim_{h\to 0} (\zeta_1(h) - e^{h\lambda}) = 0$

$$0 = \rho(\zeta_1(h)) + h\lambda\sigma(\zeta_1(h)) = \rho(e^{h\lambda} + \epsilon) + h\lambda\sigma(e^{h\lambda} + \epsilon)$$

$$= \rho(e^{h\lambda}) + \rho'(e^{h\lambda})\epsilon + O(\epsilon^2) + h\lambda\sigma(e^{h\lambda}) + O(h\epsilon)$$

$$= O(h^{r+1}) + O(\epsilon) + O(\epsilon^2) + O(h\epsilon) \ \Rightarrow \ \epsilon = O(h^{r+1}) \quad \underline{\text{ok}}$$

<u>convergence theory</u>

$y' = f(y)$, $y(0) = y_0$: well-posed IVP

$$\sum_{i=0}^{k} (\alpha_i u_{n-i} + h\beta_i f(u_{n-i})) = 0 \quad , \quad u_0, \ldots, u_{k-1} \ : \ \text{given}$$

<u>definition</u> : A multistep method is ...

... <u>convergent</u> if $u_n \to y(t)$ as $n \to \infty$ under the assumption that $t = nh$ is fixed and $u_0, \ldots, u_{k-1} \to y_0$.

... <u>stable</u> wrt initial data if there exists a constant C such that for all $n \geq k$, $|u_n - v_n| \leq C \max\{|u_0 - v_0|, \ldots, |u_{k-1} - v_{k-1}|\}$, where u_n, v_n are two solutions of the difference scheme, $t = nh$ is fixed, and the constant C may depend on t, but is independent of n and h.

<u>theorem</u>

1. stability $\Leftrightarrow$ root condition for $\rho(\zeta)$

2. if the scheme is consistent, then stability $\Leftrightarrow$ convergence

<u>recall</u> : A k-step scheme has order $\leq 2k$.

<u>theorem</u> (Dahlquist)

A stable k-step scheme has order $\begin{cases} \leq k+1 & \text{if } k \text{ is odd,} \\ \leq k+2 & \text{if } k \text{ is even.} \end{cases}$

If the scheme has order $k+2$, then the roots of $\rho(\zeta)$ all lie on the unit circle and so the method is weakly unstable.

<u>example</u>

$k = 1$: the trapezoid method is a stable 1-step scheme of order 2

$k = 2$: Milne's method is a stable 2-step scheme of order 4 , hw

13
Thurs
2/22

example

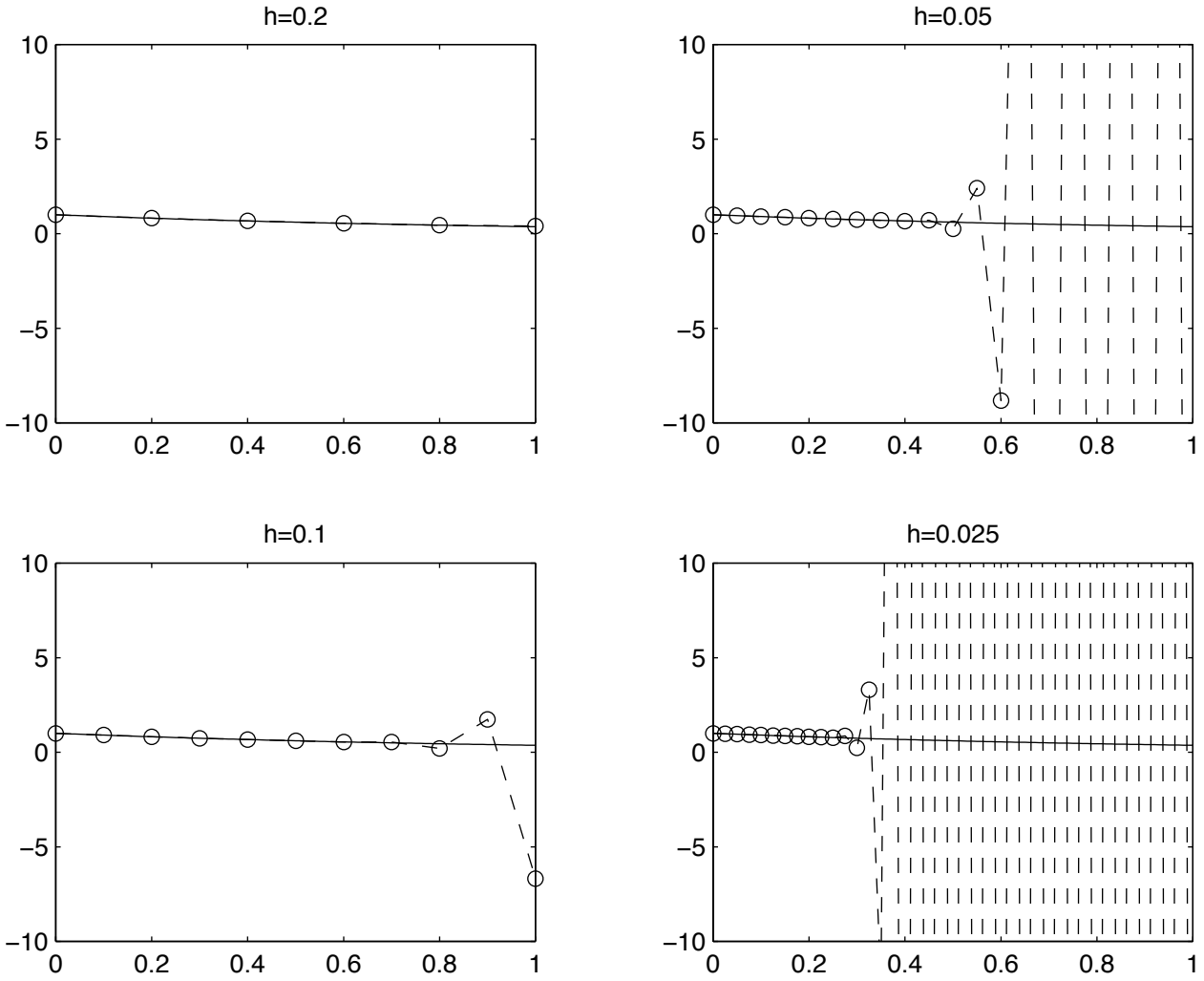
$u_n + 4u_{n-1} - 5u_{n-2} - h(4f(u_{n-1}) + 2f(u_{n-2})) = 0$: 2-step , explicit

hw : $\tau_n = O(h^4) \Rightarrow$ the scheme is consistent

$\rho(\zeta) = \zeta^2 + 4\zeta - 5 = (\zeta - 1)(\zeta + 5)$: root condition fails

$\Rightarrow$ the scheme is unstable $\Rightarrow$ the scheme is not convergent

consider $y' = -y$, $y(0) = 1 \Rightarrow y(t) = e^{-t}$, $u_0 = 1$, $u_1 = e^{-h\lambda}$



1. For any time step h, the numerical solution is accurate for some time, but the growing oscillations of the extraneous root eventually ruin the accuracy; this is seen in the expression,

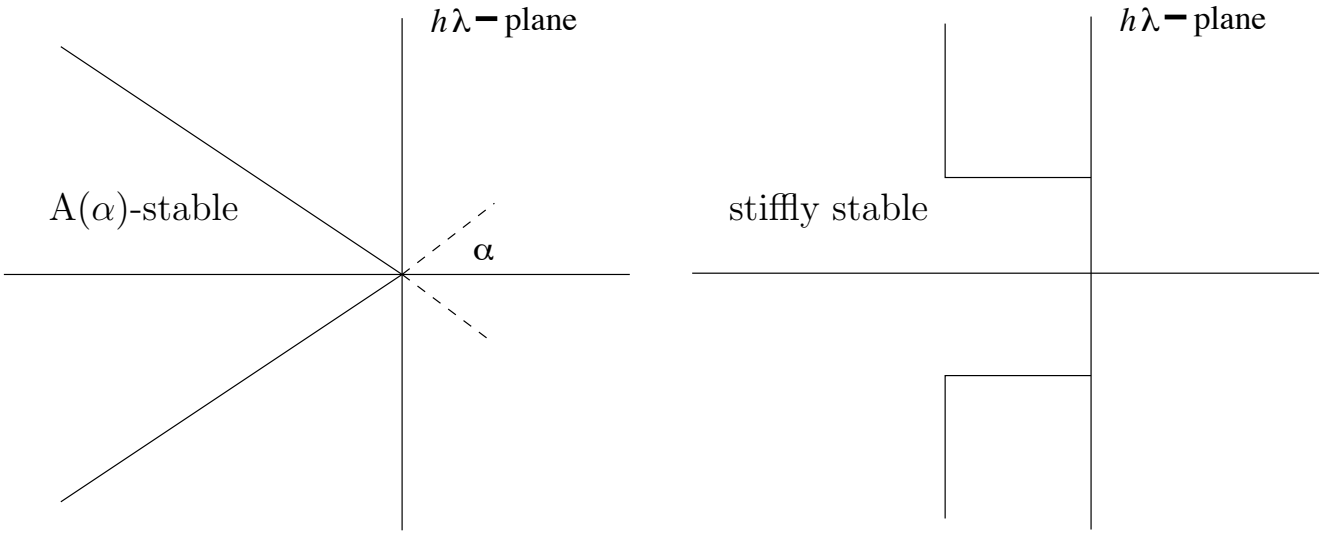
$u_n = (1 + O(h^4)) \cdot \left(e^{-t} + O(h^3)\right) + O(h^4) \cdot \left((-5)^n e^{3t/5} + O(h)\right)$ (hw).

2. The oscillations appear at an earlier time as h decreases; this strong instability is in contrast to weak instability (e.g. leap-frog), where the growing oscillations of the extraneous root appear at a later time as h decreases.

recall : A multistep scheme is A-stable if the region of absolute stability contains the left half-plane (e.g. backward Euler, trapezoid).

theorem (Dahlquist) An A-stable multistep scheme has order ≤ 2; among all A-stable schemes of order 2, the trapezoid method has the smallest $C_3 \, (= \frac{1}{12})$.

definition : A multistep method is A(α)-stable for $0 < \alpha < \pi/2$ if the region of absolute stability contains the wedge $|\arg(h\lambda) - \pi| \leq \alpha$.



theorem (Widlund) For any $k \leq 4$ and $0 < \alpha < \pi/2$, there exists a k-step method of order k which is A(α)-stable.

definition : A multistep method is stiffly stable if the region of absolute stability contains a domain of the type shown.

example : BDF methods (backward differentiation formula)

recall AB/AM : $y' = f(y) \Rightarrow y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(y(t))dt \ , \ p(t) \approx f(y(t)) \ , \ \ldots$

BDF : $p(t) = u_n + (t - t_n)\dfrac{\nabla u_n}{h} + (t - t_n)(t - t_{n-1})\dfrac{\nabla^2 u_n}{2h^2} \approx y(t)$

$$p'(t) = \frac{\nabla u_n}{h} + (2t - t_n - t_{n-1})\frac{\nabla^2 u_n}{2h^2} \Rightarrow p'(t_n) = \frac{\nabla u_n}{h} + \frac{\nabla^2 u_n}{2h} \approx y'(t_n)$$

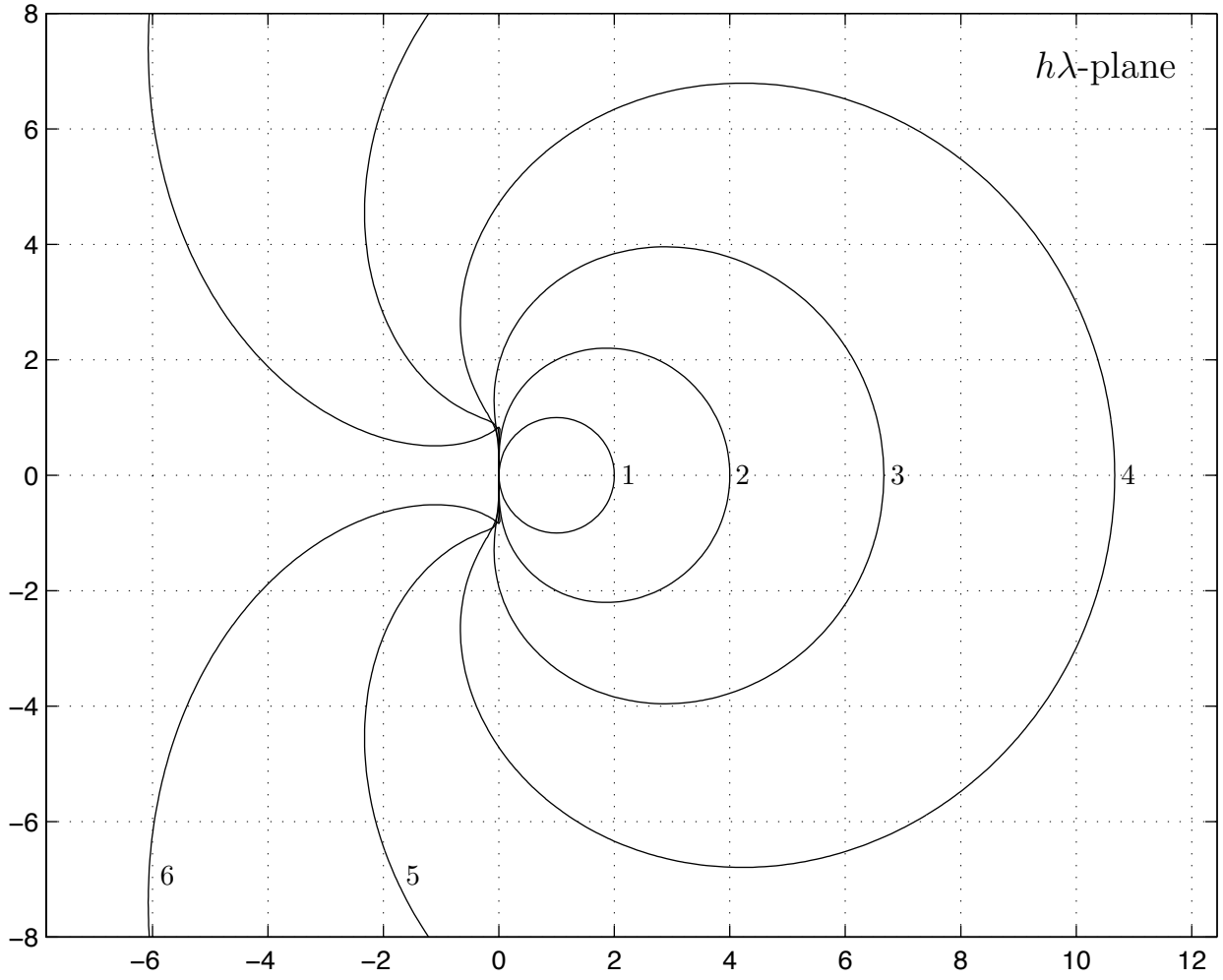
$$\nabla u_n + \frac{1}{2}\nabla^2 u_n = hf(u_n)$$

$$\frac{3}{2}u_n - 2u_{n-1} + \frac{1}{2}u_{n-2} - hf(u_n) = 0 \ : \ \text{2-step BDF, implicit}$$

$\tau_n = O(h^3)$: 2nd order accurate

$$\rho(\zeta) = \frac{3}{2}\zeta^2 - 2\zeta + \frac{1}{2} = \frac{1}{2}(\zeta - 1)(3\zeta - 1) \Rightarrow \zeta_1 = 1 \ , \ \zeta_2 = \frac{1}{3} \ : \ \text{stable}$$

<u>theorem</u> (Gear) The k-step BDF method is stiffly stable for $k \leq 6$.



For $k = 1:6$, the k-step BDF method is absolutely stable outside the contour.

<u>miscellaneous</u>

1. implicit schemes : $F(u_n) = 0$, fixed-point iteration , Newton's method
2. other methods for $y' = f(y)$: implicit RK , ...
3. software : www.netlib.org , Matlab : ode23 , ode45 , ...
5. adaptive error control : variable time step/order
6. mechanical systems : $My'' + Cy' + Ky = F$: Newmark , HHT , ...
7. Hamiltonian systems : $\begin{cases} p' = \quad H_q \\ q' = -H_p \end{cases}$, geometric/symplectic methods

2. IBVP for PDEs

heat equation

temperature : $v(x,t)$, heat flux : $\kappa\nabla v(x,t)$

conservation of energy

$$\frac{d}{dt}\int_D v\,dx = \int_{\partial D} \kappa\nabla v\cdot n\,dS \Rightarrow \int_D v_t\,dx = \int_D \nabla\cdot\kappa\nabla v\,dx \Rightarrow v_t = \nabla\cdot\kappa\nabla v$$

model problem

$v_t = v_{xx}$, $v(x,0) = f(x)$: find $v(x,t)$ for $t > 0$

1. $-\infty < x < \infty$: free-space BC

2. $0 \le x \le 1$, Dirichlet : $v(0,t) = v(1,t) = 0$

Neumann : $v_x(0,t) = v_x(1,t) = 0$

periodic : $v(0,t) = v(1,t)$, $v_x(0,t) = v_x(1,t)$

We assume the problem is well-posed. (Math 454, 556, 656)

finite-difference scheme

14 Tues 3/5

$h = \Delta x$, $x_j = jh$, $j = 0, \pm1, \pm2, \ldots$, $k = \Delta t$, $t_n = nk$, $n = 0, 1, 2, \ldots$

$u_j^n \approx v(x_j,t_n)$, $v_{xx}(x_j,t_n) \approx D_+D_-u_j^n$

$$D_+u_j^n = \frac{u_{j+1}^n - u_j^n}{h}\,,\ D_-u_j^n = \frac{u_j^n - u_{j-1}^n}{h}$$

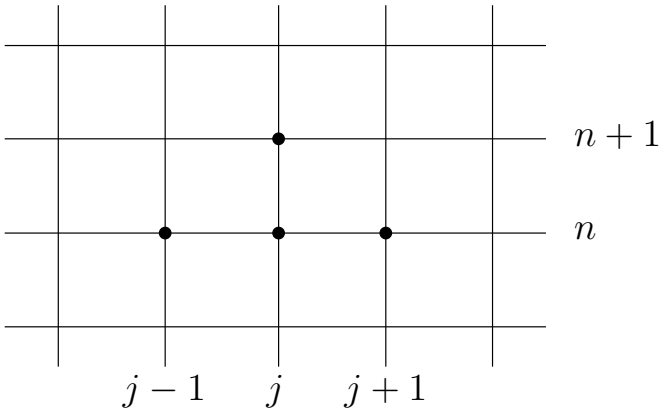
$$D_+D_-u_j^n = D_+\Big(\frac{u_j^n - u_{j-1}^n}{h}\Big) = \frac{1}{h}\Big(D_+u_j^n - D_+u_{j-1}^n\Big)$$

$$= \frac{1}{h}\Big(\frac{u_{j+1}^n - u_j^n}{h} - \frac{u_j^n - u_{j-1}^n}{h}\Big) = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}$$

$$v_t = v_{xx} \to \frac{u_j^{n+1} - u_j^n}{k} = D_+D_-u_j^n = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}$$

$$u_j^{n+1} = u_j^n + kD_+D_-u_j^n = u_j^n + \lambda(u_{j+1}^n - 2u_j^n + u_{j-1}^n)\,,\ \lambda = k/h^2$$

stencil



definition : Let $||u^n||_\infty = \max_j |u_j^n|$. The scheme is stable if $||u^n||_\infty \le ||u^0||_\infty$ for all $n \ge 1$ and all u^0; this property is also called the maximum principle.

theorem : the scheme is stable $\Leftrightarrow \lambda \le 1/2 \Leftrightarrow k \le h^2/2$

proof

$\Leftarrow$) $u_j^{n+1} = \lambda u_{j+1}^n + (1-2\lambda)u_j^n + \lambda u_{j-1}^n$

$|u_j^{n+1}| \le \lambda|u_{j+1}^n| + (1-2\lambda)|u_j^n| + \lambda|u_{j-1}^n|$

$\le \lambda||u^n||_\infty + (1-2\lambda)||u^n||_\infty + \lambda||u^n||_\infty = ||u^n||_\infty$

$\Rightarrow ||u^{n+1}||_\infty \le ||u^n||_\infty$ ok

$\Rightarrow$) assume $\lambda > 1/2$, we will show the scheme is unstable

consider $u_j^0 = (-1)^j$

$u_j^1 = \lambda u_{j+1}^0 + (1-2\lambda)u_j^0 + \lambda u_{j-1}^0 = \lambda(-1)^{j+1} + (1-2\lambda)(-1)^j + \lambda(-1)^{j-1}$

$= (-1)^j(-\lambda + (1-2\lambda) - \lambda) = (1-4\lambda)u_j^0$

$\lambda > 1/2 \Rightarrow 1-4\lambda < -1 \Rightarrow ||u^1||_\infty > ||u^0||_\infty$ ok

theorem (convergence)

$v_t = v_{xx}$, $0 \le x \le 1$, $v(x,0) = f(x)$, $v(0,t) = v(1,t) = 0$

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} \ , \ h = \frac{1}{N} \ , \ j = 1 : N-1$$

Let $x = x_j = jh, t = t_n = nk, \lambda = \dfrac{k}{h^2}$ be fixed, set $u_j^0 = f(x_j), u_0^n = u_N^n = 0$.

If $\lambda \le 1/2$, then $\lim\limits_{h\to 0} u_j^n = v(x,t)$.

proof 1. local truncation error

$$\frac{v_j^{n+1} - v_j^n}{k} = \frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{h^2} + \tau_j^n \ , \text{ contrast to ODEs } \begin{cases} y' = f(y) \\ u_{n+1} = u_n + hf(u_n) \\ y_{n+1} = y_n + hf(y_n) + \tau_n \end{cases}$$

let $v = v_j^n$, $v_t = (v_t)_j^n$, ...

$v_j^{n+1} = v + kv_t + \frac{1}{2}k^2 v_{tt} + O(k^3)$

$v_{j+1}^n = v + hv_x + \frac{1}{2}h^2 v_{xx} + \frac{1}{6}h^3 v_{xxx} + \frac{1}{24}h^4 v_{xxxx} + \frac{1}{120}h^5 v_{xxxxx} + O(h^6)$

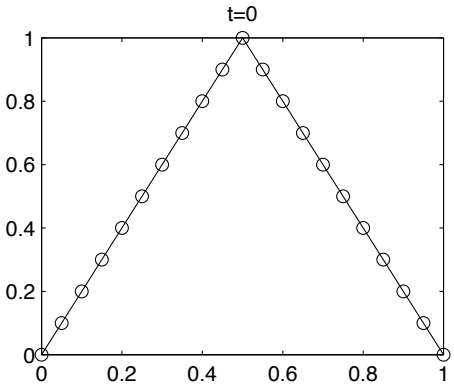
$v_{j-1}^n = v - hv_x + \frac{1}{2}h^2 v_{xx} - \frac{1}{6}h^3 v_{xxx} + \frac{1}{24}h^4 v_{xxxx} - \frac{1}{120}h^5 v_{xxxxx} + O(h^6)$

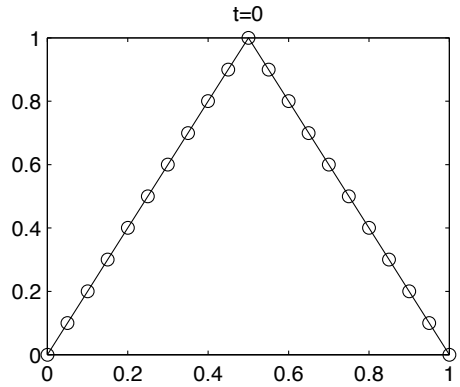
$\tau_j^n = \cancel{v_t} + \frac{1}{2}kv_{tt} + O(k^2) - (\cancel{v_{xx}} + \frac{1}{12}h^2 v_{xxxx} + O(h^4))$

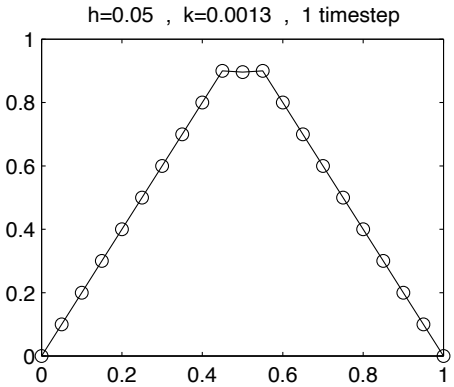
$v_t = v_{xx} \Rightarrow v_{tt} = v_{xxt} = v_{txx} = v_{xxxx} \Rightarrow \tau_j^n = (\frac{1}{2}k - \frac{1}{12}h^2)v_{xxxx} + O(k^2) + O(h^4)$

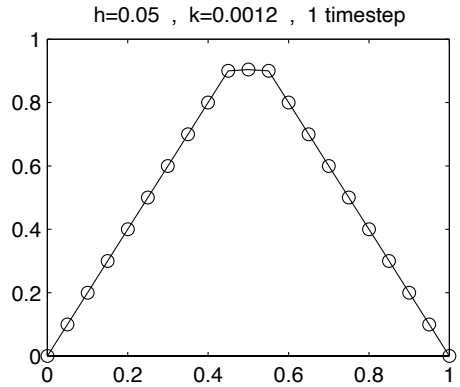
$\tau_j^n = O(k) + O(h^2)$: 1st order in time, 2nd order in space, scheme is consistent

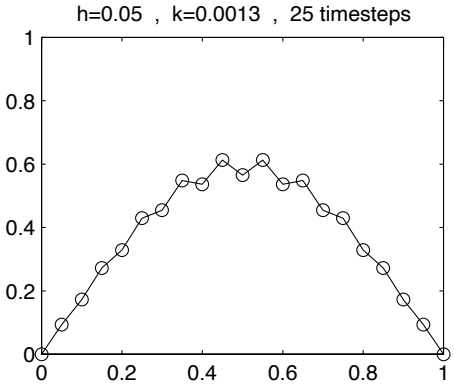
example : $u_j^{n+1} = u_j^n + kD_+D_-u_j^n$, $h = 0.05$, $\lambda = 1/2 \Rightarrow k_c = \lambda h^2 = 0.00125$

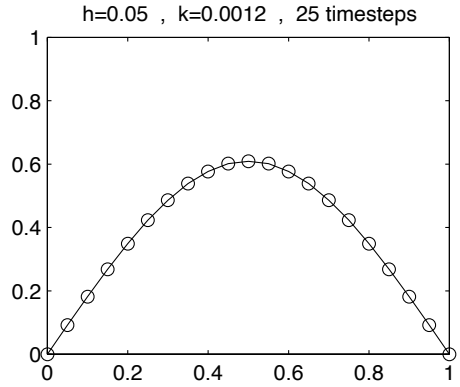


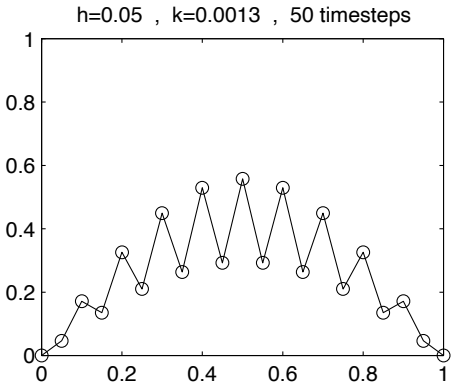


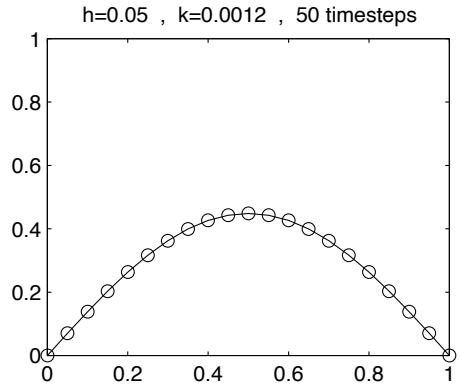












$k < k_c$: scheme is stable , $k > k_c$: scheme is unstable

2. error bound

$u_j^{n+1} = \lambda u_{j+1}^n + (1-2\lambda)u_j^n + \lambda u_{j-1}^n$

$v_j^{n+1} = \lambda v_{j+1}^n + (1-2\lambda)v_j^n + \lambda v_{j-1}^n + k\tau_j^n$

set $e_j^n = v_j^n - u_j^n$, $e_j^{n+1} = \lambda e_{j+1}^n + (1-2\lambda)e_j^n + \lambda e_{j-1}^n + k\tau_j^n$

$\lambda \le 1/2 \Rightarrow ||e^{n+1}||_\infty \le ||e^n||_\infty + k||\tau||_\infty \le ||e^{n-1}||_\infty + 2k||\tau||_\infty \le \cdots$

$||e^n||_\infty \le ||e^0||_\infty + nk||\tau||_\infty = t \cdot O(k)$ ok

note

1. We see (again) that for a consistent scheme, stability $\Rightarrow$ convergence.

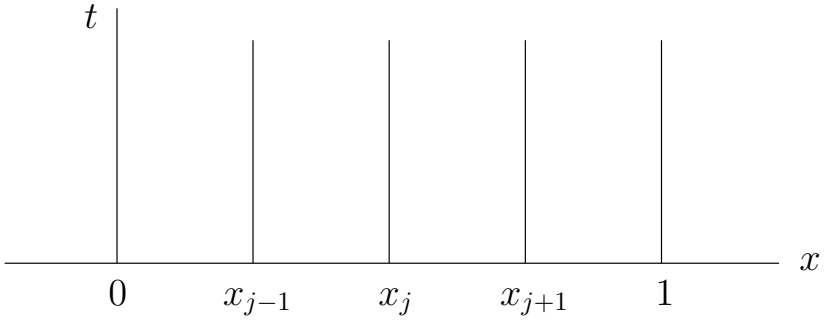
2. $u_j^n = v(x,t) + O(k) = v(x,t) + kE_j^n + O(k^2)$: hw

15
Tues
3/12

alternative approach : method of lines

$v_t = v_{xx}$, $0 \le x \le 1$, $v(x,0) = f(x)$, $v(0,t) = v(1,t) = 0$

define $u_j(t) \approx v(x_j,t)$, $x_j = jh$, $h = 1/N$, $j = 1 : N-1$



$u_j' = \dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2}$: system of ODEs , $u_0 = u_N = 0$

$$u' = Au \ , \ u = \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_{N-1} \end{pmatrix} \ , \ A = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{pmatrix}$$

A : tridiagonal , symmetric $\Rightarrow$ real e-values

Euler's method : $u^{n+1} = u^n + kAu^n$: finite-difference scheme

recall : absolute stability $\Leftrightarrow -2 \le k\mu \le 0$ for all e-values μ of A

theorem (Gershgorin)

If μ is an e-value of $A = (a_{ij})$, then there exists i such that $|\mu - a_{ii}| \le \sum_{j \ne i} |a_{ij}|$.

proof : Math 571 or ...

Gershgorin $\Rightarrow \left|\mu - \left(\frac{-2}{h^2}\right)\right| \le \frac{2}{h^2} \Rightarrow \frac{-4}{h^2} \le \mu \le 0$

hence, absolute stability holds if $-2 \le k \cdot \frac{-4}{h^2} \le 0 \Leftrightarrow \frac{k}{h^2} \le \frac{1}{2}$

---

e-values and e-vectors of A

$Au = \mu u$, $u = (u_1, \ldots, u_{N-1})^T$

$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = \mu u_j$, $u_0 = u_N = 0$

$u_{j+1} - (2 + \mu h^2)u_j + u_{j-1} = 0$: difference equation

$\rho(\zeta) = \zeta^2 - (2 + \mu h^2)\zeta + 1 = (\zeta - \zeta_1)(\zeta - \zeta_2) = 0$

$u_j = a_1\zeta_1^j + a_2\zeta_2^j$

$u_0 = 0 \Rightarrow a_1 + a_2 = 0 \Rightarrow a_2 = -a_1$

$u_N = 0 \Rightarrow a_1(\zeta_1^N - \zeta_2^N) = 0 \Rightarrow \zeta_1^N = \zeta_2^N \Rightarrow \left(\frac{\zeta_1}{\zeta_2}\right)^N = 1$

$\frac{\zeta_1}{\zeta_2} = e^{2\pi i m/N} = e^{2\pi i m h}$, $m = 1 : N-1$, $m = 0, N$ doesn't work ...

$\zeta_1\zeta_2 = 1 \Rightarrow \zeta_1^2 = e^{2\pi i m h} \Rightarrow \zeta_1 = e^{\pi i m h}$, $\zeta_2 = e^{-\pi i m h}$

$u_j = a_1\zeta_1^j + a_2\zeta_2^j = a_1\left(e^{\pi i j m h} - e^{-\pi i j m h}\right) = a_1 \cdot 2i \sin \pi j m h$

$\zeta_1 + \zeta_2 = 2 + \mu h^2 \Rightarrow \mu = \frac{\zeta_1 + \zeta_2 - 2}{h^2} = \frac{e^{\pi i m h} + e^{-\pi i m h} - 2}{h^2} = \frac{2(\cos \pi m h - 1)}{h^2}$

e-values of A : $\mu_m = \frac{2(\cos \pi m h - 1)}{h^2} = \frac{-4}{h^2}\sin^2 \frac{\pi m h}{2}$, $m = 1 : N-1$

e-vectors of A : $u_{m,j} = \sin \pi m x_j$, $j = 1 : N-1$: discrete Fourier modes

$\pi m = \xi$: wavenumber , wavelength $= \frac{2\pi}{\xi} = \frac{2}{m}$

---

matrix analysis of convergence

16
Thurs
3/14

$v_t = v_{xx}$

$u^{n+1} = u^n + kAu^n$: method of lines + Euler's method

$u^{n+1} = (I + kA)u^n$

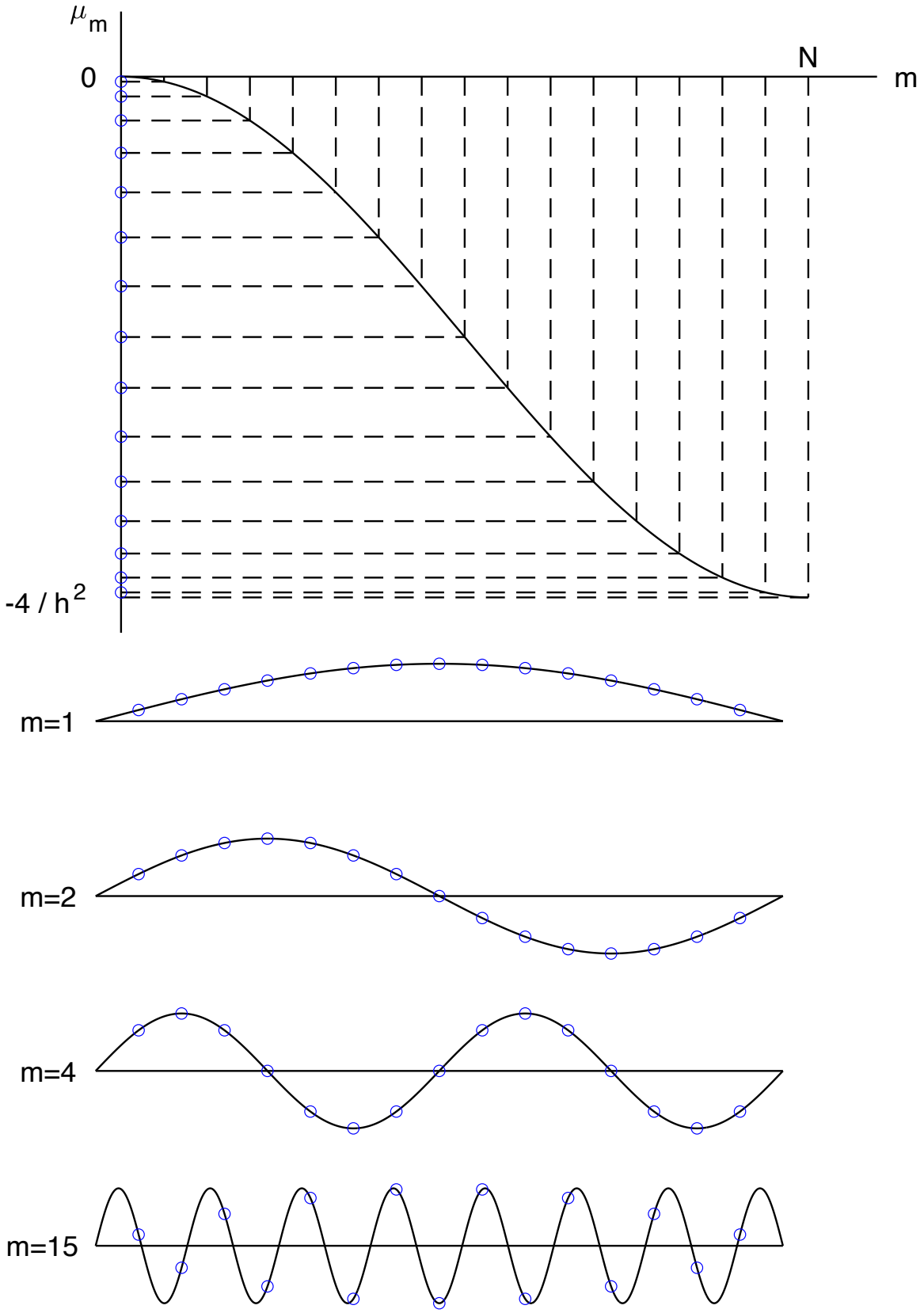
$v^{n+1} = (I + kA)v^n + k\tau^n$

$e^n = v^n - u^n \Rightarrow e^{n+1} = (I + kA)e^n + k\tau^n$

$||e^{n+1}|| \le ||I + kA|| \cdot ||e^n|| + k||\tau^n||$

example : $N = 16$, $A = D_+D_-$ + Dirichlet BC , $Au_m = \mu_m u_m$

$\mu_m = -(4/h^2)\sin^2(\pi m h/2)$, $u_{m,j} = \sin \pi m x_j$, $m = 1 : N-1$, $j = 1 : N-1$



$m \to 0$: long wavelength modes $\Rightarrow \mu \to 0$

$m \to N$: short ····· " ····· $\Rightarrow \mu \to -4/h^2$

$$I+kA = \begin{pmatrix} 1-2\lambda & \lambda & & & \\ \lambda & 1-2\lambda & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \lambda \\ & & & \lambda & 1-2\lambda \end{pmatrix}, \ \lambda = \frac{k}{h^2}$$

$$||I+kA||_\infty = ||I+kA||_1 = \begin{cases} 1 & \text{if } \lambda \le 1/2 \\ 4\lambda - 1 & \text{if } \lambda > 1/2 \end{cases}$$

$||I+kA||_2$ = max of absolute values of e-values of $I+kA$

recall : Gershgorin $\Rightarrow$ the e-values of A lie in the interval $[-4/h^2, 0]$

$$\Rightarrow \quad \cdots\cdots\cdots \ I+kA \ \cdots\cdots\cdots\cdots \ [1-4\lambda, 1]$$

hence $\lambda \le 1/2 \Rightarrow ||I+kA||_p \le 1$ for $p = 1, 2, \infty$

$||e^{n+1}|| \le ||e^n|| + k||\tau^n|| \Rightarrow ||e^n|| \le ||e^0|| + t||\tau||$, $||\tau|| = \max_n ||\tau^n||$ ok

---

Fourier analysis of PDE

$v_t = v_{xx}$, look for solutions of the form $v(x,t) = e^{\omega t + i\xi x}$: Fourier mode

ξ : wavenumber , ω : growth rate , $\omega = -\xi^2$: dispersion relation

$v(x,t) = e^{-\xi^2 t + i\xi x}$ $\begin{cases} \xi = 0 \ : \text{ constant mode} \\ \xi \neq 0 \ : \text{ oscillatory in space, decaying in time} \end{cases}$

free-space BC, $v(x,0) = f(x)$, $-\infty < x < \infty$

Fourier transform : $\widehat{f}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}dx$, $f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi)e^{i\xi x}d\xi$

Parseval relation : $\int_{-\infty}^{\infty} |f(x)|^2 dx = 2\pi \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2 d\xi$, $||f||_2^2 = 2\pi||\widehat{f}||_2^2$

theorem

1. $v(x,t) = \int_{-\infty}^{\infty} \widehat{f}(\xi)e^{-\xi^2 t + i\xi x}d\xi$: solution formula

2. $||v(\cdot,t)||_2 \le ||f||_2$ for all $t \ge 0$: L_2-stability

proof

1. $\widehat{v}(\xi,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} v(x,t)e^{-i\xi x}dx$

$\widehat{v}_t(\xi,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} v_t(x,t)e^{-i\xi x}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} v_{xx}(x,t)e^{-i\xi x}dx$

$= \frac{1}{2\pi} v_x(x,t)e^{-i\xi x}\Big|_{-\infty}^{\infty} - \frac{1}{2\pi}\int_{-\infty}^{\infty} v_x(x,t)(-i\xi)e^{-i\xi x}dx$

$= -\frac{1}{2\pi} v(x,t)(-i\xi)e^{-i\xi x}\Big|_{-\infty}^{\infty} + \frac{1}{2\pi}\int_{-\infty}^{\infty} v(x,t)(-i\xi)^2 e^{-i\xi x}dx = -\xi^2\widehat{v}(\xi,t)$

$\widehat{v}_t(\xi,t) = -\xi^2\widehat{v}(\xi,t)$, $\widehat{v}(\xi,0) = \widehat{f}(\xi) \Rightarrow \widehat{v}(\xi,t) = \widehat{f}(\xi)e^{-\xi^2 t}$

$\Rightarrow\ v(x,t) = \int_{-\infty}^{\infty} \hat{v}(\xi,t) e^{i\xi x} d\xi = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{-\xi^2 t} e^{i\xi x} d\xi$ ok

2. $||v(\cdot,t)||_2^2 = \int_{-\infty}^{\infty} |v(x,t)|^2 dx = 2\pi \int_{-\infty}^{\infty} |\hat{v}(\xi,t)|^2 d\xi$

$= 2\pi \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 e^{-2\xi^2 t} d\xi \ \le\ 2\pi \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 d\xi = \int_{-\infty}^{\infty} |f(x)|^2 dx \ =\ ||f||_2^2$ ok

Fourier analysis of difference scheme

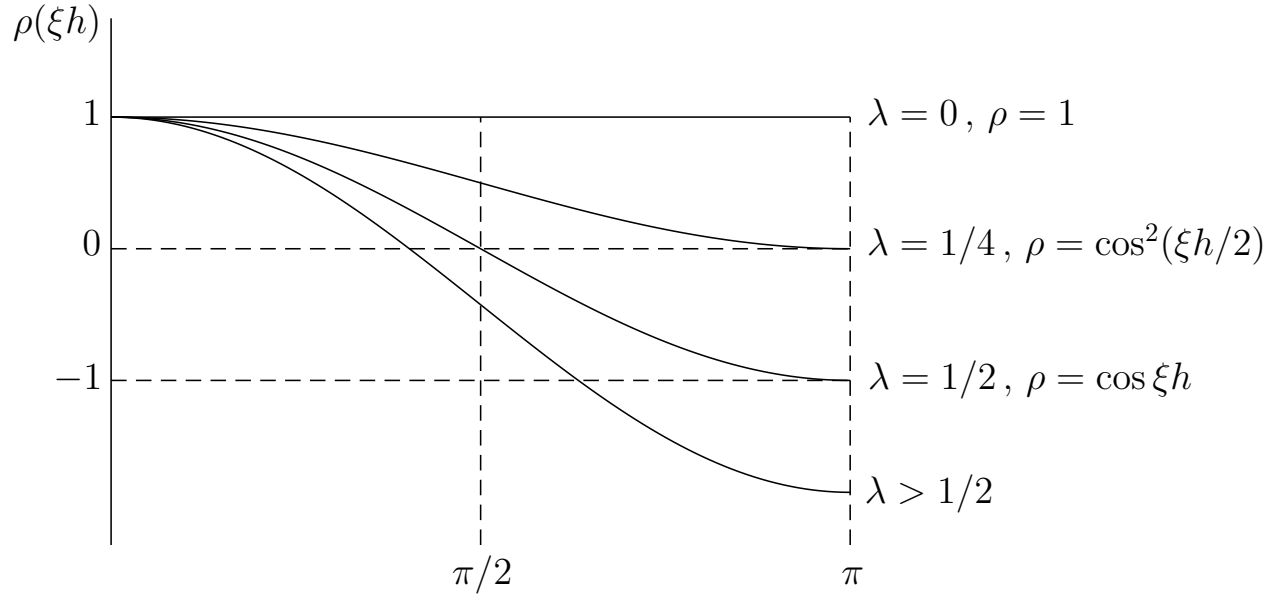
$u_j^{n+1} = u_j^n + \lambda \left(u_{j+1}^n - 2u_j^n + u_{j-1}^n \right)\ ,\ j = 0,\, \pm 1,\, \ldots$

look for solutions of the form $u_j^n = \zeta^n e^{i\xi jh}\ ,\ -\pi \le \xi h \le \pi$

$\zeta^{n+1} e^{i\xi jh} = \zeta^n e^{i\xi jh} + \lambda \left(\zeta^n e^{i\xi (j+1)h} - 2\zeta^n e^{i\xi jh} + \zeta^n e^{i\xi (j-1)h} \right)$

$\zeta = 1 + \lambda(e^{i\xi h} - 2 + e^{-i\xi h}) = 1 + 2\lambda(\cos \xi h - 1) = 1 - 4\lambda \sin^2(\xi h/2) = \rho(\xi h)$

$u_j^{n+1} = \rho(\xi h) u_j^n\ ,\ \rho(\xi h)\ :$ amplification factor



1. $|\rho(\xi h)| \le 1$ for all $\xi h \ \Leftrightarrow\ \lambda \le 1/2$

17
Tues
3/19

2. $u_j^n = \rho(\xi h)^n u_j^0$

$0 \le \lambda \le 1/4$: all modes decay monotonically in time

$1/4 \le \lambda \le 1/2$: { long waves decay monotonically in time
short waves oscillate in sign, amplitude decays

$\lambda > 1/2$: { long waves decay monotonically in time
intermediate waves oscillate in sign, amplitude decays
short waves oscillate in sign, amplitude grows

note : $\xi h = \pi\ \Rightarrow\ u_j^0 = e^{i\pi j} = (-1)^j$, $\rho(\pi) = 1 - 4\lambda$, recall ...

3. amplification factor of PDE $= e^{-\xi^2 k} = e^{-\lambda(\xi h)^2} = \rho(\xi h) + O((\xi h)^4)$; for the PDE all modes decay monotonically in time, unlike for the difference scheme

Consider the difference scheme with free-space BC.

$u_j^{n+1} = u_j^n + kD_+D_-u_j^n \ , \ u_j^0 = f(x_j) \ , \ x_j = jh \ , \ j = 0, \pm 1, \ldots$

We will derive results for u_j^n analogous to the results for $v(x,t)$.

Fourier coefficients : $\widehat{f}_m = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}dx \ , \ f(x) = \sum_{m=-\infty}^{\infty} \widehat{f}_m \, e^{imx}$

Parseval relation : $\int_{-\pi}^{\pi} |f(x)|^2 dx = 2\pi \sum_{m=-\infty}^{\infty} |\widehat{f}_m|^2$

change notation : $m \to j \ , \ \widehat{f}_m \to u_j^n \ , \ x \to \xi h \ , \ f(x) \to \widehat{u}^n(\xi h)$

$u_j^n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{u}^n(\xi h)e^{-ij\xi h}d(\xi h) \ , \ \widehat{u}^n(\xi h) = \sum_{j=-\infty}^{\infty} u_j^n \, e^{ij\xi h} \ , \ \text{hw5}$

$\int_{-\pi}^{\pi} |\widehat{u}^n(\xi h)|^2 d(\xi h) = 2\pi \sum_{j=-\infty}^{\infty} |u_j^n|^2 \ , \ \text{hw5}$

<u>theorem</u>

1. $u_j^n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{u}^0(\xi h)\rho(\xi h)^n e^{-ij\xi h}d(\xi h)$: solution formula

2. $\lambda \le 1/2 \Rightarrow ||u^n||_2 \le ||u^0||_2$ for all $n \ge 0$: ℓ_2-stability

<u>proof</u>

1. $\widehat{u}^{n+1}(\xi h) = \sum_{j=-\infty}^{\infty} u_j^{n+1} e^{ij\xi h} = \sum_{j=-\infty}^{\infty} (u_j^n + \lambda(u_{j+1}^n - 2u_j^n + u_{j-1}^n))e^{ij\xi h}$

$$= \sum_{j=-\infty}^{\infty} \left(u_j^n e^{ij\xi h} + \lambda \left(\underbrace{u_{j+1}^n e^{ij\xi h}}_{u_j^n e^{i(j-1)\xi h}} - 2u_j^n e^{ij\xi h} + \underbrace{u_{j-1}^n e^{ij\xi h}}_{u_j^n e^{i(j+1)\xi h}} \right) \right)$$

$= \sum_{j=-\infty}^{\infty} u_j^n e^{ij\xi h}(1 + \lambda(e^{-i\xi h} - 2 + e^{i\xi h})) = \rho(\xi h)\widehat{u}^n(\xi h) \, , \, \rho(\xi h) = 1 - 4\lambda \sin^2(\xi h/2)$

$\widehat{u}^{n+1}(\xi h) = \rho(\xi h)\widehat{u}^n(\xi h) \Rightarrow \widehat{u}^n(\xi h) = \rho(\xi h)^n \widehat{u}^0(\xi h)$ <u>ok</u>

2. $||u^n||_2^2 = \sum_{j=-\infty}^{\infty} |u_j^n|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\widehat{u}^n(\xi h)|^2 d(\xi h) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\rho(\xi h)^n \widehat{u}^0(\xi h)|^2 d(\xi h)$

$\le \frac{1}{2\pi}\int_{-\pi}^{\pi} |\widehat{u}^0(\xi h)|^2 d(\xi h) = \sum_{j=-\infty}^{\infty} |u_j^0|^2 = ||u^0||_2^2$ <u>ok</u>

implicit schemes

$v_t = v_{xx}$, $0 \le x \le 1$, $v(0,t) = v(1,t) = 0$: Dirichlet BC

method of lines : $u' = Au$, $A = (1/h^2)\text{trid}(1,-2,1)$

the e-values of A lie in the interval $(-4/h^2, 0)$: stiff system

forward Euler : conditionally stable , i.e. stable $\Leftrightarrow \lambda \le 1/2$

backward Euler : unconditionally stable , i.e. stable for all $\lambda > 0$, hw5

Crank-Nicolson : central differencing in space + trapezoid in time

$u^{n+1} = u^n + \frac{1}{2}k(Au^n + Au^{n+1})$: error is $O(h^2) + O(k^2)$, 2nd order accurate

$(I - \frac{1}{2}kA)u^{n+1} = (I + \frac{1}{2}kA)u^n$: linear system (more later)

theorem : CN is unconditionally stable and convergent in ℓ_2-norm

proof

1. $u^{n+1} = (I - \frac{1}{2}kA)^{-1}(I + \frac{1}{2}kA)u^n$

$\Rightarrow ||u^{n+1}||_2 \le ||(I - \frac{1}{2}kA)^{-1}||_2 \cdot ||I + \frac{1}{2}kA||_2 \cdot ||u^n||_2$

the e-values of $I - \frac{1}{2}kA$ lie in the interval $(1, 1 + 2\lambda)$

..... " ..... $I + \frac{1}{2}kA$ ..... " ..... $(1 - 2\lambda, 1)$

$||(I - \frac{1}{2}kA)^{-1}||_2 \le 1$, $||I + \frac{1}{2}kA||_2 \le \max\{1, |1 - 2\lambda|\} \le 1 \Leftrightarrow \lambda \le 1$

then $\lambda \le 1 \Rightarrow ||u^{n+1}||_2 \le ||u^n||_2$: we can do better

if μ is an e-value of A , then $\dfrac{1 + \frac{1}{2}k\mu}{1 - \frac{1}{2}k\mu}$ is an e-value of $(I - \frac{1}{2}kA)^{-1}(I + \frac{1}{2}kA)$

$\mu < 0 \Rightarrow ||(I - \frac{1}{2}kA)^{-1}(I + \frac{1}{2}kA)||_2 \le 1 \Rightarrow ||u^{n+1}||_2 \le ||u^n||_2$ for all $\lambda > 0$ ok

2. $u^{n+1} = u^n + \frac{1}{2}k(Au^n + Au^{n+1})$

$v^{n+1} = v^n + \frac{1}{2}k(Av^n + Av^{n+1}) + k\tau^n$, $\tau^n = O(k^2)$

18
Thurs
3/21

$e^n = v^n - u^n \;\Rightarrow\; e^{n+1} = e^n + \frac{1}{2}k(Ae^n + Ae^{n+1}) + k\tau^n$

$e^{n+1} = (I - \frac{1}{2}kA)^{-1}(I + \frac{1}{2}kA)e^n + (I - \frac{1}{2}kA)^{-1}k\tau^n$

$||e^{n+1}||_2 \le ||e^n||_2 + k||\tau^n||_2 \;\Rightarrow\; ||e^n||_2 \le ||e^0||_2 + t \cdot O(k^2)$ ok

note : proving stability of CN in other norms requires different analysis

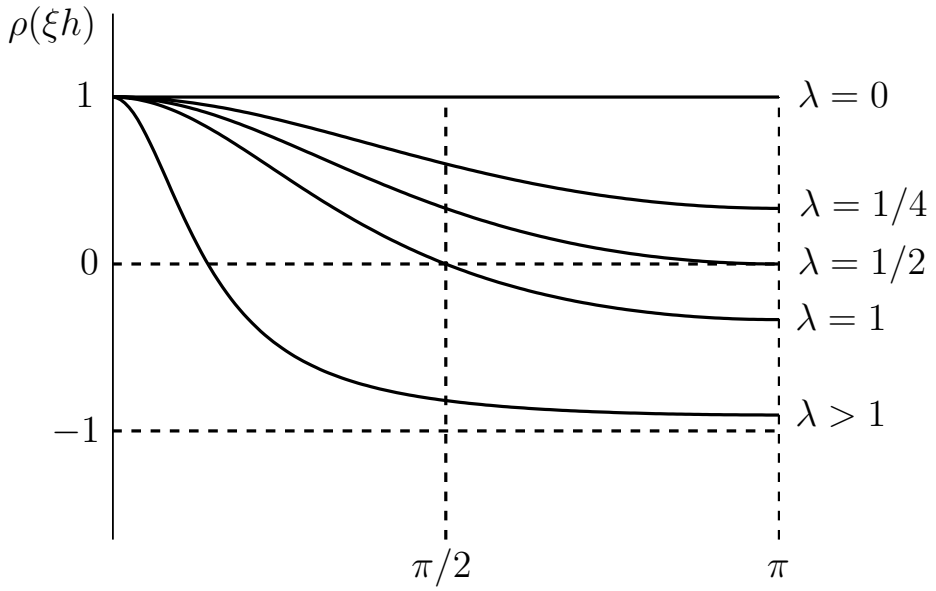
Fourier analysis (Crank-Nicolson)

$u_j^{n+1} = u_j^n + \frac{1}{2}k(D_+D_-u_j^n + D_+D_-u_j^{n+1})$

$u_j^{n+1} - \frac{1}{2}\lambda(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) = u_j^n + \frac{1}{2}\lambda(u_{j+1}^n - 2u_j^n + u_{j-1}^n)$

look for $u_j^n = \zeta^n e^{i\xi jh}$, $j = 0, \pm 1, \ldots$

$$\zeta = \frac{1 + \frac{1}{2}\lambda(e^{i\xi h} - 2 + e^{-i\xi h})}{1 - \frac{1}{2}\lambda(e^{i\xi h} - 2 + e^{-i\xi h})} = \frac{1 - 2\lambda\sin^2(\xi h/2)}{1 + 2\lambda\sin^2(\xi h/2)} = \rho(\xi h) \ , \ u_j^n = \rho(\xi h)^n u_j^0$$



1. $0 \le \lambda \le 1/2$: all modes decay monotonically in time

$\lambda > 1/2$: { long waves decay monotonically in time; short waves oscillate in sign, amplitude decays

2. $|\rho(\xi h)| \le 1$ for all ξh and all $\lambda > 0$

$\Rightarrow$ CN is unconditionally ℓ_2-stable (with free-space BC)

The proof follows as for conditional ℓ_2-stability of the forward Euler/central difference scheme on page 45.

recall CN : $(I - \frac{1}{2}kA)u^{n+1} = (I + \frac{1}{2}kA)u^n$

write this as $Ax = f$, not the same A , assume $N \times N$

<u>tridiagonal Gaussian elimination</u> : special case

$$A = \begin{pmatrix} a & b & & & \\ b & a & b & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & b \\ & & & b & a \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ \beta_2 & 1 & & & \\ & \beta_3 & 1 & & \\ & & \ddots & \ddots & \\ & & & \beta_N & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 & b & & & \\ & \alpha_2 & b & & \\ & & \ddots & \ddots & \\ & & & \ddots & b \\ & & & & \alpha_N \end{pmatrix}$$

$A = LU$, L : unit lower triangular , U : upper triangular

<u>steps</u>

1. find L, U
2. solve $Ly = f$
3. solve $Ux = y$

check $Ax = LUx = Ly = f$ <u>ok</u>

<u>find L, U</u>

$a = \alpha_1$, $b = \beta_k \alpha_{k-1} \Rightarrow \beta_k = \dfrac{b}{\alpha_{k-1}}$, $a = \beta_k b + \alpha_k \Rightarrow \alpha_k = a - \beta_k b$, $k = 2 : N$

<u>solve $Ly = f$</u> : forward elimination

$y_1 = f_1$, $f_k = \beta_k y_{k-1} + y_k \Rightarrow y_k = f_k - \beta_k y_{k-1}$, $k = 2 : N$

<u>solve $Ux = y$</u> : back substitution

$y_N = \alpha_N x_N \Rightarrow x_N = \dfrac{y_N}{\alpha_n}$, $y_k = \alpha_k x_k + b x_{k+1} \Rightarrow x_k = \dfrac{y_k - b x_{k+1}}{\alpha_k}$, $k = N - 1 : 1$

note : memory and operation count are $O(N)$

summary on Fourier analysis : we are using 4 types of transforms

1. $f(x)/\widehat{f}(\xi)$, $-\infty < x < \infty \,/\, -\infty < \xi < \infty$

2. $f(x)/\{\widehat{f}_m\}$, $0 \le x \le 1 \,/\, m = 0, \pm 1, \ldots$

3. $\{u_j\}/\widehat{u}(\xi h)$, $j = 0, \pm 1, \ldots / -\pi \le \xi h \le \pi$

4. $\{u_j\}/\{\widehat{u}_m\}$, $j = 1 : N-1 \,/\, m = 1 : N-1$

The specific form of the transform pair depends on the domain and BC, and in each case there is a Parseval relation.

example : stability analysis of the forward Euler/central difference scheme

$v_t = v_{xx}$, $0 \le x \le 1$, $v(0,t) = v(1,t) = 0$: Dirichlet BC

$u^{n+1} = (I + kA)u^n$, $u^n = \{u_j^n\}$, $j = 1 : N-1$, $u_0^n = u_N^n = 0$, $h = 1/N$

$A = (1/h^2)\text{trid}(1, -2, 1)$

previously we used transform pair #3 , now use #4

e-values of A : $\mu_m = (-4/h^2)\sin^2(\pi m h/2)$, $m = 1 : N-1$

e-vectors of A : $q_m = \{\sin \pi m j h\}$, $j = 1 : N-1$, orthonormal basis , check ...

any vector $u = \{u_j\}$ can be expanded in this basis

$$u = \sum_{m=1}^{N-1} \widehat{u}_m q_m \Rightarrow u_j = \sum_{m=1}^{N-1} \widehat{u}_m \sin \pi m j h \ : \text{ inversion formula}$$

$$\widehat{u}_m = u \cdot q_m = \sum_{j=1}^{N-1} u_j \sin \pi m j h \ : \text{ discrete Sine transform}$$

$$||u||_2^2 = \sum_{j=1}^{N-1} u_j^2 = \sum_{m=1}^{N-1} \widehat{u}_m^2 = ||\widehat{u}||_2^2 \ : \text{ Parseval relation}$$

now consider the difference scheme

19
Tues
3/26

$$u^0 = \sum_{m=1}^{N-1} \widehat{u}_m^0 q_m$$

$$u^n = (I + kA)^n u^0 = (I + kA)^n \sum_{m=1}^{N-1} \widehat{u}_m^0 q_m = \sum_{m=1}^{N-1} \widehat{u}_m^0 (1 + k\mu_m)^n q_m = \sum_{m=1}^{N-1} \widehat{u}_m^n q_m$$

$|1 + k\mu_m| = |1 - 4\lambda \sin^2(\pi m h/2)| \le 1$ for $\lambda \le 1/2$

then $||u^n||_2 = ||\widehat{u}^n||_2 \le ||\widehat{u}^0||_2 = ||u^0||_2$: ℓ_2-stability

hw : Neumann BC , $v_x(0,t) = v_x(1,t) = 0$

energy method : alternative method for proving 2-norm stability

$v_t = v_{xx}$, $-\infty < x < \infty$, ok too for $0 \le x \le 1$ + Dirichlet BC

define $||v(\cdot,t)||_2 = \Big(\int_{-\infty}^{\infty} v(x,t)^2 \, dx \Big)^{1/2}$: energy

theorem : $||v(\cdot,t)||_2 \le ||v(\cdot,0)||_2$ for all $t \ge 0$: L_2-stability

proof

$$\frac{d}{dt}||v(\cdot,t)||_2^2 = \frac{d}{dt}\int_{-\infty}^{\infty} v^2 \, dx = \int_{-\infty}^{\infty} 2vv_t \, dx = 2\int_{-\infty}^{\infty} vv_{xx} \, dx$$

$$= \cancel{2vv_x}\Big|_{-\infty}^{\infty} - 2\int_{-\infty}^{\infty} (v_x)^2 \, dx \le 0 \quad \text{ok}$$

integration by parts

define $(f,g) = \int_{-\infty}^{\infty} f(x)g(x) \, dx$: inner product , $(f,f) = ||f||_2^2$

theorem : $(f,g') = -(f',g)$

proof

$(fg)' = fg' + f'g$

$$\int_{-\infty}^{\infty} (fg)' \, dx = \int_{-\infty}^{\infty} (fg' + f'g) \, dx = (f,g') + (f',g) = fg\Big|_{-\infty}^{\infty} = 0 \quad \text{ok}$$

summation by parts

define $(f,g)_h = \sum_{j=-\infty}^{\infty} f_j g_j$, $(f,f)_h = ||f||_{2,h}^2$

theorem : $(f, D_- g) = -(D_+ f, g)$

proof

$$D_+(fg)_j = \frac{f_{j+1}g_{j+1} - f_j g_j}{h} = \frac{f_{j+1}g_{j+1} - f_{j+1}g_j + f_{j+1}g_j - f_j g_j}{h}$$

$$= f_{j+1} D_+ g_j + (D_+ f_j) g_j$$

$$\sum_{j=-\infty}^{\infty} D_+(fg)_j = \sum_{j=-\infty}^{\infty} (f_{j+1} D_+ g_j + (D_+ f_j) g_j) = \sum_{j=-\infty}^{\infty} (f_j D_+ g_{j-1} + (D_+ f_j) g_j)$$

$$D_+ g_{j-1} = \frac{g_j - g_{j-1}}{h} = D_- g_j$$

$$\sum_{j=-\infty}^{\infty} D_+(fg)_j = (f, D_- g) + (D_+ f, g) = \sum_{j=-\infty}^{\infty} \frac{f_{j+1}g_{j+1} - f_j g_j}{h} = 0 \quad \text{ok}$$

$\uparrow$

telescoping sum

application to difference schemes

$||u^n||_2 = (u^n, u^n)^{1/2} = \left(\sum_{j=-\infty}^{\infty} (u_j^n)^2 \right)^{1/2}$: discrete energy

1. forward Euler/central difference

$u^{n+1} = u^n + kD_+D_-u^n$

theorem : $\lambda \le 1/2 \Rightarrow ||u^n||_2 \le ||u^0||_2$: conditionally stable

proof

$||u^{n+1}||^2 - ||u^n||^2 = (u^{n+1}, u^{n+1}) - (u^n, u^n) = (u^{n+1} + u^n, u^{n+1} - u^n)$

$= (2u^n + kD_+D_-u^n, kD_+D_-u^n) = 2k(u^n, D_+D_-u^n) + k^2||D_+D_-u^n||^2 = a + b$

$a = -2k(D_-u^n, D_-u^n) = -2k||D_-u^n||^2$

$b = k^2 \left|\left| \dfrac{S_+D_-u^n - D_-u^n}{h} \right|\right|^2 \le \dfrac{k^2}{h^2} \left(||S_+D_-u^n|| + ||D_-u^n|| \right)^2 = k\lambda \cdot 4||D_-u^n||^2$

$||u^{n+1}||^2 - ||u^n||^2 \le (-2k + 4k\lambda)||D_-u^n||^2 = 2k(2\lambda - 1)||D_-u^n||^2 \le 0$ ok

2. backward Euler/central difference

theorem : $u^{n+1} = u^n + kD_+D_-u^{n+1}$: unconditionally stable , pf : hw

3. Crank-Nicolson

theorem : $u^{n+1} = u^n + \frac{1}{2}kD_+D_- \left(u^n + u^{n+1} \right)$: unconditionally stable

proof

$||u^{n+1}||^2 - ||u^n||^2 = (u^{n+1} + u^n, u^{n+1} - u^n) = (u^{n+1} + u^n, \frac{1}{2}kD_+D_-(u^n + u^{n+1}))$

$= -\frac{1}{2}k(D_-(u^{n+1} + u^n), D_-(u^n + u^{n+1})) = -\frac{1}{2}k||D_-(u^n + u^{n+1})||^2 \le 0$ ok

---

2D heat equation

$v_t = v_{xx} + v_{yy}$, $v(x, y, 0) = f(x, y)$, $-\infty < x, y < \infty$

$u_{j,l}^n \approx v(jh, \ell h, nk)$, $j, \ell = 0, \pm 1, \ldots$, $n = 0, 1, \ldots$

$D_+^x u_{j,\ell}^n = \dfrac{u_{j+1,\ell}^n - u_{j,\ell}^n}{h}$, $D_-^x u_{j,\ell}^n = \dfrac{u_{j,\ell}^n - u_{j-1,\ell}^n}{h}$

$D_+^y u_{j,\ell}^n = \dfrac{u_{j,\ell+1}^n - u_{j,\ell}^n}{h}$, $D_-^y u_{j,\ell}^n = \dfrac{u_{j,\ell}^n - u_{j,\ell-1}^n}{h}$

forward Euler/central difference

$u_{j,\ell}^{n+1} = u_{j,\ell}^n + k(D_+^x D_-^x + D_+^y D_-^y)u_{j,\ell}^n$

$u_{j,\ell}^{n+1} = u_{j,\ell}^n + \lambda(u_{j+1,\ell}^n - 2u_{j,\ell}^n + u_{j-1,\ell}^n + u_{j,\ell+1}^n - 2u_{j,\ell}^n + u_{j,\ell-1}^n)$

1. maximum principle

$u_{j,\ell}^{n+1} = (1-4\lambda)u_{j,\ell}^n + \lambda(u_{j+1,\ell}^n + u_{j-1,\ell}^n + u_{j,\ell+1}^n + u_{j,\ell-1}^n)$

theorem : $||u^n||_\infty \leq ||u^0||_\infty \Leftrightarrow \lambda \leq 1/4$

proof : as before

2. Fourier analysis , look for $u_{j,l}^n = \zeta^n e^{i(\xi j + \eta l)h}$

20
Thurs
3/28

$\rho(\xi h, \eta h) = 1 + 2\lambda(\cos \xi h + \cos \eta h - 2) = 1 - 4\lambda(\sin^2(\xi h/2) + \sin^2(\eta h/2))$

$|\rho(\xi h, \eta h)| \leq 1 \Leftrightarrow \lambda \leq 1/4$

$u_{j,\ell}^n = \dfrac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \widehat{u}^n(\xi h, \eta h) e^{i(\xi j + \eta \ell)h} d(\xi h) d(\eta h)$

$\widehat{u}^n(\xi h, \eta h) = \sum_{j,\ell=-\infty}^{\infty} u_{j,\ell}^n e^{-i(\xi j + \eta \ell)h}$

$\sum_{j,\ell=-\infty}^{\infty} |u_{j,\ell}^n|^2 = \dfrac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} |\widehat{u}^n(\xi h, \eta h)|^2 d(\xi h) d(\eta h)$

theorem : $\lambda \leq 1/4 \Rightarrow ||u^n||_2 \leq ||u^0||_2$

proof : $\widehat{u}^{n+1}(\xi h, \eta h) = \rho(\xi h, \eta h)\widehat{u}^n(\xi h, \eta h)$ ... as before

---

backward Euler/central difference

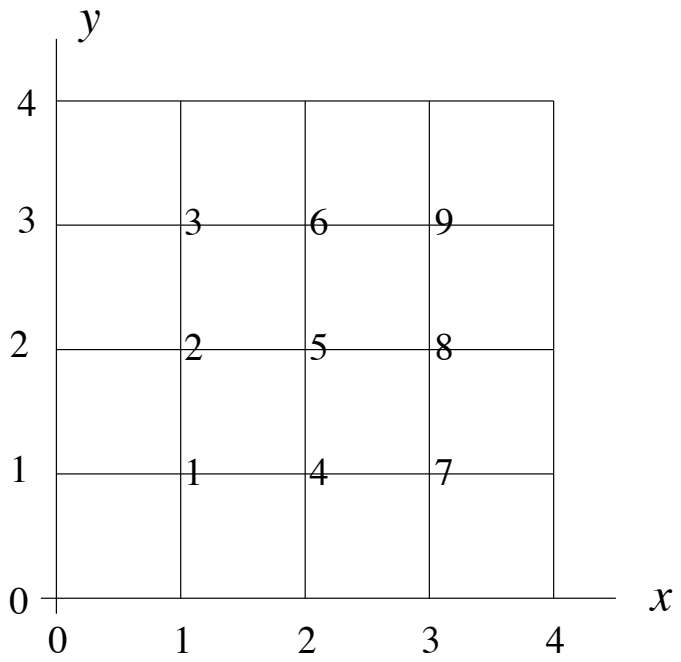
$u^{n+1} = u^n + k(D_+^x D_-^x + D_+^y D_-^y)u^{n+1}$

$(I - k(D_+^x D_-^x + D_+^y D_-^y))u^{n+1} = u^n$: linear system

$(1+4\lambda)u_{j,\ell}^{n+1} - \lambda(u_{j+1,\ell}^{n+1} + u_{j-1,\ell}^{n+1} + u_{j,\ell+1}^{n+1} + u_{j,\ell-1}^{n+1}) = u_{j,\ell}^n$

consider the problem on $D = [0,1]^2$

$v(x,y,0) = f(x,y)$ on D , $v(x,y,t) = 0$ on ∂D



1	2	3	4	5	6	7	8	9
u_{11}	u_{12}	u_{13}	u_{21}	u_{22}	u_{23}	u_{31}	u_{32}	u_{33}
$1+4\lambda$	$-\lambda$		$-\lambda$					
$-\lambda$	$1+4\lambda$	$-\lambda$		$-\lambda$				
	$-\lambda$	$1+4\lambda$			$-\lambda$			
$-\lambda$			$1+4\lambda$	$-\lambda$		$-\lambda$		
	$-\lambda$		$-\lambda$	$1+4\lambda$	$-\lambda$		$-\lambda$	
		$-\lambda$		$-\lambda$	$1+4\lambda$			$-\lambda$
			$-\lambda$			$1+4\lambda$	$-\lambda$	
				$-\lambda$		$-\lambda$	$1+4\lambda$	$-\lambda$
					$-\lambda$		$-\lambda$	$1+4\lambda$

1. symmetric, block tridiagonal, positive definite

methods : Cholesky, block LU, onjugate gradient, FFT, SOR, multigrid, ...

2. $\rho(\xi h, \eta h) = \ldots$, 1D case is on hw

$|\rho(\xi h, \eta h)| \leq 1$ for all λ : unconditionally stable in 2-norm

(can also show by energy method)

3. global error is $O(k) + O(h^2)$

goal : find a method that requires only 1D tridiagonal solves, is unconditionally stable, and has global error $O(k^2) + O(h^2)$

operator splitting

$y' = Ay \Rightarrow y(t) = e^{At}y_0$

$e^{At} = I + At + \frac{1}{2}A^2t^2 + \cdots$

define $S_A(t) = e^{At}$: solution operator

$y(t) = S_A(t)y_0$

$S_A(t) = e^{At} = e^{Ank} = (e^{Ak})^n = (S_A(k))^n$

consider $y' = (A + B)y$

theorem

$S_{A+B}(k) \neq S_A(k)S_B(k)$, unless $AB = BA$

proof

$S_{A+B}(k) = e^{(A+B)k} = I + (A + B)k + \frac{1}{2}(A + B)^2k^2 + \cdots$

$= I + (A + B)k + \frac{1}{2}(A^2 + AB + BA + B^2)k^2 + \cdots$

$S_A(k)S_B(k) = e^{Ak}e^{Bk} = (I + Ak + \frac{1}{2}(Ak)^2 + \cdots)(I + Bk + \frac{1}{2}(Bk)^2 + \cdots)$

$= I + (A + B)k + (\frac{1}{2}A^2 + AB + \frac{1}{2}B^2)k^2 + \cdots$

$S_{A+B}(k) - S_A(k)S_B(k) = \frac{1}{2}(BA - AB)k^2 + \cdots$ ok

$S_{A+B}(k) = S_A(k)S_B(k) + O(k^2)$

$S_{A+B}(nk) = (S_{A+B}(k))^n = (S_A(k)S_B(k) + O(k^2))^n = (S_A(k)S_B(k))^n + O(k)$

$\Rightarrow$ global error is $O(k)$, this holds more generally

application : backward Euler/central difference scheme for 2D heat equation

$u^{n+1} = u^n + k(D_+^xD_-^x + D_+^yD_-^y)u^{n+1}$

$(I - k(D_+^xD_-^x + D_+^yD_-^y))u^{n+1} = u^n$

$u^{n+1} = S_{A+B}(k)u^n$, $S_{A+B}(k) = (I - k(D_+^xD_-^x + D_+^yD_-^y))^{-1}$

set $S_A(k) = (I - kD_+^xD_-^x)^{-1}$, $S_B(k) = (I - kD_+^yD_-^y)^{-1}$

theorem : $S_{A+B}(k) = S_A(k)S_B(k) + O(k^2)$

proof

$u^n = (I - k(D_+^x D_-^x + D_+^y D_-^y))u^{n+1}$

$\quad = (I - kD_+^y D_-^y)(I - kD_+^x D_-^x)u^{n+1} - k^2 D_+^y D_-^y D_+^x D_-^x u^{n+1}$

$u^{n+1} = (I - kD_+^x D_-^x)^{-1}(I - kD_+^y D_-^y)^{-1}u^n + O(k^2)$ ok

The backward Euler/central difference scheme can be replaced by the following.

fractional-step method

$(I - kD_+^y D_-^y)u^{n+\frac{1}{2}} = u^n$

$(I - kD_+^x D_-^x)u^{n+1} = u^{n+\frac{1}{2}}$

21
Tues
4/2

note : If the vector components are correctly ordered, then only 1D tridiagonal systems need to be solved.

1st half-step : u_1 , u_2 , u_3 , u_4 , u_5 , u_6 , u_7 , u_8 , u_9

2nd half-step : u_1 , u_4 , u_7 , u_2 , u_5 , u_8 , u_3 , u_6 , u_9

unconditionally stable in 2-norm , global error is $O(k) + O(h^2)$

Crank-Nicolson

$u^{n+1} = u^n + \frac{1}{2}k(D_+^x D_-^x + D_+^y D_-^y)(u^n + u^{n+1})$

$(I - \frac{1}{2}k(D_+^x D_-^x + D_+^y D_-^y))u^{n+1} = (I + \frac{1}{2}k(D_+^x D_-^x + D_+^y D_-^y))u^n$

matrix is symmetric, positive definite, block-tridiagonal

$$\rho(\xi h, \eta h) = \frac{1 - 2\lambda(\sin^2(\xi h/2) + \sin^2(\eta h/2))}{1 + 2\lambda(\sin^2(\xi h/2) + \sin^2(\eta h/2))} \Rightarrow |\rho(\xi h, \eta h)| \leq 1 \text{ for all } \lambda > 0$$

unconditionally stable in 2-norm , global error is $O(k^2) + O(h^2)$

fractional-step/Crank-Nicolson

$(I - \frac{1}{2}kD_+^y D_-^y)u^{n+1/2} = (I + \frac{1}{2}k(D_+^x D_-^x + D_+^y D_-^y))u^n$

$(I - \frac{1}{2}kD_+^x D_-^x)u^{n+1} = u^{n+1/2}$

1D tridiagonal systems, unconditionally stable, $O(k) + O(h^2)$

alternating direction implicit

$(I - \frac{1}{2}kD_+^yD_-^y)u^{n+1/2} = (I + \frac{1}{2}kD_+^xD_-^x)u^n$

$(I - \frac{1}{2}kD_+^xD_-^x)u^{n+1} \quad = (I + \frac{1}{2}kD_+^yD_-^y)u^{n+1/2}$

1D tridiagonal systems

stability of ADI

$$\rho_1(\xi h, \eta h) = \frac{1 - 2\lambda \sin^2(\xi h/2)}{1 + 2\lambda \sin^2(\eta h/2)} \ , \ \ \rho_1(\pi, 0) = 1 - 2\lambda$$

$$\rho_2(\xi h, \eta h) = \frac{1 - 2\lambda \sin^2(\eta h/2)}{1 + 2\lambda \sin^2(\xi h/2)} \ , \ \ \rho_2(0, \pi) = 1 - 2\lambda$$

$\Rightarrow$ each fractional step is conditionally stable : $\lambda \leq 1$

$$\rho(\xi h, \eta h) = \rho_1(\xi h, \eta h) \cdot \rho_2(\xi h, \eta h) = \frac{1 - 2\lambda \sin^2(\xi h/2)}{1 + 2\lambda \sin^2(\eta h/2)} \cdot \frac{1 - 2\lambda \sin^2(\eta h/2)}{1 + 2\lambda \sin^2(\xi h/2)}$$

$\Rightarrow |\rho(\xi h, \eta h)| \leq 1$ for all ξ, η, λ : each full step is unconditionally stable

accuracy of ADI

$u^{n+1/2} - u^n = \frac{1}{2}k(D_+^yD_-^yu^{n+1/2} + D_+^xD_-^xu^n)$

$u^{n+1} - u^{n+1/2} = \frac{1}{2}k(D_+^xD_-^xu^{n+1} + D_+^yD_-^yu^{n+1/2})$

- -

$u^{n+1} - u^n = \frac{1}{2}kD_+^xD_-^x(u^{n+1} + u^n) + kD_+^yD_-^yu^{n+1/2}$

$2u^{n+1/2} - (u^{n+1} + u^n) = \frac{1}{2}kD_+^xD_-^x(u^n - u^{n+1})$

- -

$u^{n+1/2} = \frac{1}{2}(u^{n+1} + u^n) - \frac{1}{4}kD_+^xD_-^x(u^{n+1} - u^n)$

$$\underbrace{u^{n+1} - u^n = \frac{1}{2}k(D_+^xD_-^x + D_+^yD_-^y)(u^{n+1} + u^n)}_{\text{Crank-Nicolson}} \ - \ \underbrace{\frac{1}{4}k^2D_+^yD_-^yD_+^xD_-^x(u^{n+1} - u^n)}_{= \frac{k^3}{4}v_{xxyyt} + \cdots}$$

$\Rightarrow$ the global error of ADI is $O(k^2) + O(h^2)$

hyperbolic equations

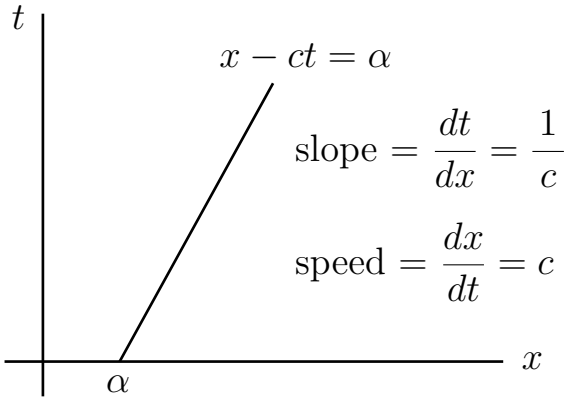
wave propagation : sound , light , water , gravitational , ...

scalar wave equation

$v_t + cv_x = 0$, $-\infty < x < \infty$, $v(x,0) = f(x)$, assume $c > 0$

definition

The line $x - ct = \alpha$ in the xt-plane is called a characteristic.



theorem

1. The solution $v(x,t)$ is constant on characteristics.

2. The solution is $v(x,t) = f(x - ct)$.

proof

1. $v(x,t) = v(\alpha + ct, t)$

$\frac{d}{dt} v(\alpha + ct, t) = v_x(\alpha + ct, t) \cdot c + v_t(\alpha + ct, t) = 0$ ok

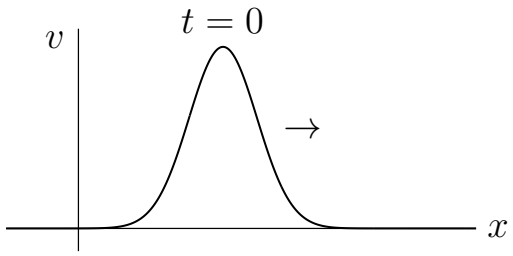
2. $v(x,t) = v(\alpha, 0) = f(\alpha) = f(x - ct)$ ok

check : if $v(x,t) = f(x - ct)$, then $v_t + cv_x = f' \cdot -c + c \cdot f' = 0$ ok

note

1. The solution is a traveling wave with wave speed c.

2. The domain of dependence of the PDE is $\{\alpha\}$.



22
Thurs
4/4

linear system of 1st order PDEs

$$v_t + Av_x = 0 \ , \ v(x,0) = f(x) \ , \ v = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} \ , \ f = \begin{pmatrix} f_1 \\ \vdots \\ f_p \end{pmatrix}$$

The system is called hyperbolic if A is diagonalizable and has real e-values, $A = TDT^{-1}$, $D = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_i \in \mathbb{R}$.

$v_t + Av_x = v_t + TDT^{-1}v_x = 0 \Rightarrow T^{-1}v_t + DT^{-1}v_x = 0$

$w = T^{-1}v \Rightarrow w_t + Dw_x = 0 \Rightarrow (w_i)_t + \lambda_i (w_i)_x = 0 \ , \ i = 1:p$

Hence the e-values of A are the wave speeds of the system.

example

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}_t + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}_x = 0 \ , \ \begin{matrix} v_1(x,0) = f_1(x) \\ v_2(x,0) = f_2(x) \end{matrix}$$

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = TDT^{-1} \ , \ D = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \ , \ T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \ , \ T = T^{-1}$$

$$w_t + Dw_x = 0 \Rightarrow \begin{matrix} (w_1)_t + (w_1)_x = 0 \\ (w_2)_t - (w_2)_x = 0 \end{matrix} \Rightarrow \begin{matrix} w_1(x,t) = w_1(x-t,0) \\ w_2(x,t) = w_2(x+t,0) \end{matrix}$$

$$w = T^{-1}v \Rightarrow \begin{pmatrix} w_1(x,0) \\ w_2(x,0) \end{pmatrix} = T^{-1} \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} f_1(x) + f_2(x) \\ f_1(x) - f_2(x) \end{pmatrix}$$

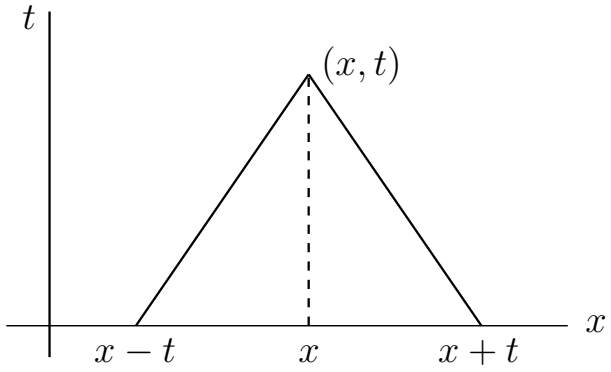
$$v = Tw \Rightarrow \begin{pmatrix} v_1(x,t) \\ v_2(x,t) \end{pmatrix} = T \begin{pmatrix} w_1(x,t) \\ w_2(x,t) \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} w_1(x,t) + w_2(x,t) \\ w_1(x,t) - w_2(x,t) \end{pmatrix}$$

$$= \frac{1}{\sqrt{2}} \begin{pmatrix} w_1(x-t,0) + w_2(x+t,0) \\ w_1(x-t,0) - w_2(x+t,0) \end{pmatrix}$$

$$= \frac{1}{2} \begin{pmatrix} f_1(x-t) + f_2(x-t) + f_1(x+t) - f_2(x+t) \\ f_1(x-t) + f_2(x-t) - f_1(x+t) + f_2(x+t) \end{pmatrix}$$

The solution is a superposition of traveling waves with wave speeds $c = \pm 1$.

The domain of dependence of the PDE is $\{x-t, x+t\}$.



example : 2nd order scalar wave equation

$v_{tt} = c^2 v_{xx}$, $v(x,0) = f(x)$, $v_t(x,0) = g(x)$, assume $c > 0$

$v_1 = v_x \;\Rightarrow\; (v_1)_t = v_{xt} = v_{tx} = (v_2)_x$

$v_2 = v_t \qquad (v_2)_t = v_{tt} = c^2 v_{xx} = c^2 (v_1)_x$

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}_t + \begin{pmatrix} 0 & -1 \\ -c^2 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}_x = 0$$

$$A = \begin{pmatrix} 0 & -1 \\ -c^2 & 0 \end{pmatrix} = TDT^{-1}$$

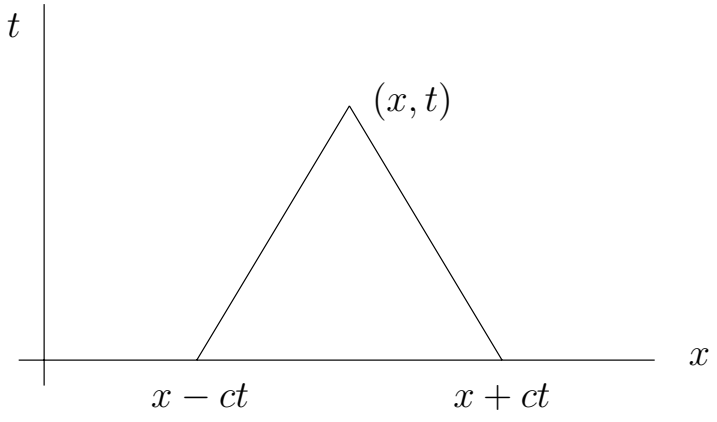
$$D = \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix} \;,\; T = \begin{pmatrix} 1 & 1 \\ -c & c \end{pmatrix} \;,\; T^{-1} = \frac{1}{2c} \begin{pmatrix} c & -1 \\ c & 1 \end{pmatrix}$$

$\dots$

$$v_1(x,t) = \frac{1}{2}\left(f_1(x+ct) + f_1(x-ct)\right) + \frac{1}{2c}\left(f_2(x+ct) - f_2(x-ct)\right)$$

$$v_x(x,t) = \frac{1}{2}\left(f'(x+ct) + f'(x-ct)\right) + \frac{1}{2c}\left(g(x+ct) - g(x-ct)\right)$$

$$v(x,t) \;=\; \frac{1}{2}\left(f(x+ct) + f(x-ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\,ds \;:\; \underline{\text{d'Alembert's formula}}$$

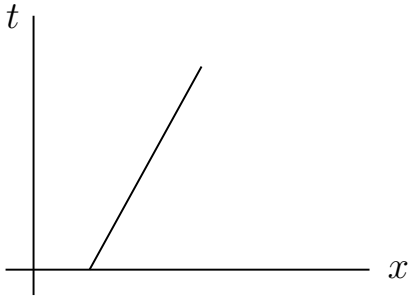


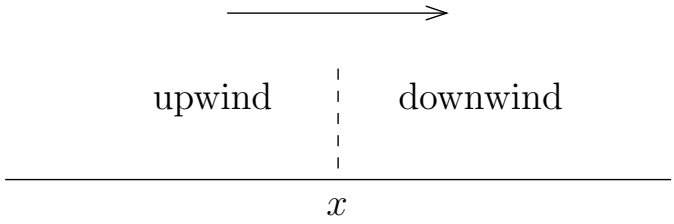
The domain of dependence of the PDE is $[x-ct, x+ct]$.

difference schemes

$v_t + cv_x = 0\ ,\ v(x,0) = f(x)$

$v(x,t) = f(x-ct)\ ,\ c > 0\ :$ right-moving wave





downwind scheme

$$\frac{u_j^{n+1} - u_j^n}{k} + cD_+u_j^n = 0$$

$$\frac{u_j^{n+1} - u_j^n}{k} + c\,\frac{u_{j+1}^n - u_j^n}{h} = 0 \Rightarrow u_j^{n+1} = (1+c\lambda)u_j^n - c\lambda\, u_{j+1}^n\ ,\ \lambda = \frac{k}{h}$$

theorem

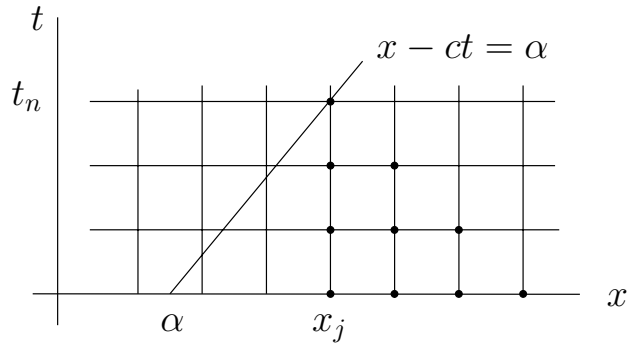
The downwind scheme is unconditionally unstable in the ∞-norm.

proof

set $u_j^0 = (-1)^j$

then $u_j^1 = (1+c\lambda)u_j^0 - c\lambda\, u_{j+1}^0 = (1+c\lambda)(-1)^j - c\lambda(-1)^{j+1} = (-1)^j(1+2c\lambda)$

$||u^1||_\infty = (1+2c\lambda)||u^0||_\infty$ ok



The domain of dependence of the PDE is $\{\alpha\}$.

The ····· " ····· downwind scheme is $\{x_j, x_{j+1}, \ldots, x_{j+n}\}$.

23
Tues
4/9

upwind scheme

$$\frac{u_j^{n+1}-u_j^n}{k}+cD_-u_j^n=0$$

$$\frac{u_j^{n+1}-u_j^n}{k}+c\,\frac{u_j^n-u_{j-1}^n}{h}=0 \;\Rightarrow\; u_j^{n+1}=(1-c\lambda)u_j^n+c\lambda\,u_{j-1}^n$$

theorem

The upwind scheme is stable in the ∞-norm $\Leftrightarrow c\lambda\le 1$.

proof

$\Leftarrow$) $u_j^{n+1}=(1-c\lambda)u_j^n+c\lambda\,u_{j-1}^n$

$|u_j^{n+1}|\le(1-c\lambda)|u_j^n|+c\lambda\,|u_{j-1}^n|\le(1-c\lambda)||u^n||_\infty+c\lambda||u^n||_\infty=||u^n||_\infty$

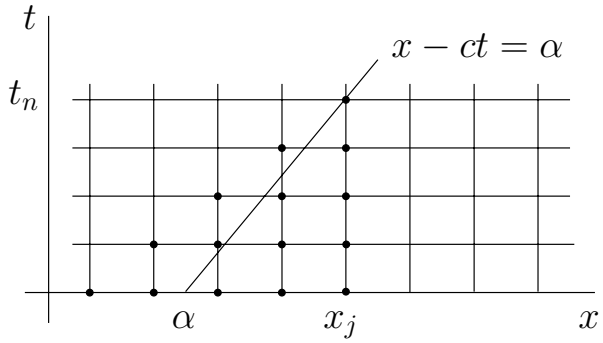
then $||u^{n+1}||_\infty\le||u^n||_\infty$ ok

$\Rightarrow$) assume $c\lambda>1$, set $u_j^0=(-1)^j$

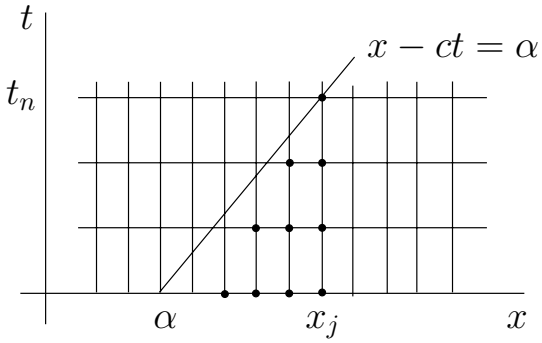
then $u_j^1=(1-c\lambda)u_j^0+c\lambda\,u_{j-1}^0=(1-c\lambda)(-1)^j+c\lambda(-1)^{j-1}=(-1)^j(1-2c\lambda)$

hence if $c\lambda>1$, then $||u^1||_\infty=(2c\lambda-1)||u^0||_\infty$ and the scheme is unstable ok

$c\lambda<1$: small k/large h



$c\lambda>1$: large k/small h



The domain of dependence of the upwind scheme is $\{x_{j-n},\dots,x_j\}$.

$x_{j-n}\le\alpha\Leftrightarrow \cancel{x_j}-x_n\le\cancel{x_j}-ct_n\Leftrightarrow \cancel{nh}\ge cn\cancel{k}\Leftrightarrow c\lambda\le 1$

CFL condition : Courant-Friedrichs-Levy (1928)

The domain of dependence of the PDE is contained in the domain of dependence of the difference scheme.

1. downwind scheme violates CFL for all λ
2. upwind scheme satisfies CFL $\Leftrightarrow c\lambda\le 1$

local truncation error : upwind scheme

$v_t + cv_x = 0$

$$\frac{v_j^{n+1} - v_j^n}{k} + c\,\frac{v_j^n - v_{j-1}^n}{h} = \tau_j^n$$

$$\tau = \frac{\cancel{v} + kv_t + \frac{1}{2}k^2 v_{tt} + O(k^3) - \cancel{v}}{k} + c\frac{\cancel{v} - (\cancel{v} - hv_x + \frac{1}{2}h^2 v_{xx} + O(h^3))}{h}$$

$$= v_t + \tfrac{1}{2}kv_{tt} + O(k^2) + c(v_x - \tfrac{1}{2}hv_{xx} + O(h^2))$$

$$= v_t + cv_x + \tfrac{1}{2}h(\lambda v_{tt} - cv_{xx}) + O(h^2) = \tfrac{1}{2}h(\lambda c^2 v_{xx} - cv_{xx}) + O(h^2)$$

$v_t = -cv_x \ , \ v_{tt} = -cv_{xt} = -cv_{tx} = -c(-cv_x)_x = c^2 v_{xx}$

$\tau = \frac{1}{2}hc(c\lambda - 1)v_{xx} + O(h^2) = O(k)$ for fixed $\lambda = k/h$

convergence : upwind scheme , $c\lambda \le 1$

$u_j^{n+1} = (1 - c\lambda)u_j^n + c\lambda u_{j-1}^n$

$v_j^{n+1} = (1 - c\lambda)v_j^n + c\lambda v_{j-1}^n + k\tau_j^n$

$e_j^n = v_j^n - u_j^n$

$e_j^{n+1} = (1 - c\lambda)e_j^n + c\lambda e_{j-1}^n + k\tau_j^n$

$\cdots$

$||e^{n+1}||_\infty \le ||e^n||_\infty + k||\tau^n||_\infty$

$\cdots$

$||e^n||_\infty \le ||e^0||_\infty + nk||\tau||_\infty = t \cdot O(k)$ if $e^0 = 0$

$$\Rightarrow\ e^n \to 0 \ \text{ if } \begin{cases} u_j^0 = f(x_j) \\ h, k \to 0 \ \text{ with } \lambda = k/h \ \text{ fixed such that } \ c\lambda \le 1 \end{cases}$$

note : The upwind scheme is exact for $c\lambda = 1$.

proof

$u_j^{n+1} = (1 - c\lambda)u_j^n \ + \ c\lambda u_{j-1}^n = u_{j-1}^n$

$u_j^n = u_{j-1}^{n-1} = \cdots = u_{j-n}^0 = f(x_{j-n})$

$$x_{j-n} = x_j - nh = x_j - \frac{t_n}{k}\,h = x_j - \frac{t_n}{\lambda} = x_j - ct_n$$

$u_j^n = f(x_{j-n}) = f(x_j - ct_n) = v(x_j, t_n)$ ok

notation

$L_h u_j^n = \dfrac{u_j^{n+1} - u_j^n}{k} + cD_- u_j^n = 0$, L_h : difference operator for upwind scheme

we know that $v_t + cv_x = 0 \Rightarrow \begin{cases} 1.\ L_h v_j^n = \tau_j^n = \frac{1}{2}hc(c\lambda - 1)v_{xx} + O(h^2) = O(h) \\ 2.\ u_j^n = v_j^n + O(h) \text{ for } c\lambda \leq 1 \end{cases}$

theorem

24 Thurs 4/11

Consider $\phi_t + c\phi_x = \frac{1}{2}hc(1 - c\lambda)\phi_{xx}$: model equation for upwind scheme.

1. $L_h \phi_j^n = O(h^2)$
2. $u_j^n = \phi_j^n + O(h^2)$ for $c\lambda \leq 1$

proof : as on previous page , future hw

1. The model equation gives insight into the behavior of the numerical solution.
2. $\frac{1}{2}hc(1 - c\lambda)\phi_{xx}$: artificial viscosity
3. Stability of the difference scheme is equivalent to requiring $||L_h^{-1}|| \leq 1$ for all h sufficiently small. (more later)

Fourier analysis of PDE

$v_t + cv_x = 0$, look for solutions of the form $v(x,t) = e^{\omega t + i\xi x}$: Fourier mode

$\omega + ic\xi = 0 \Rightarrow \omega = -ic\xi$: dispersion relation $\Rightarrow v(x,t) = e^{-ic\xi t + i\xi x} = e^{i\xi(x-ct)}$

All modes travel with phase speed c and constant amplitude independent of wavenumber ξ.

IVP with free-space BC

$v_t + cv_x = 0$, $v(x,0) = f(x)$, $-\infty < x < \infty$, goal : solution formula, stability

$$\hat{v}(\xi,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} v(x,t)e^{-i\xi x}dx$$

$$\hat{v}_t(\xi,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} v_t(x,t)e^{-i\xi x}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} -cv_x(x,t)e^{-i\xi x}dx$$

$$= -c \cdot \frac{1}{2\pi} \cancel{v(x,t)e^{-i\xi x}}\Big|_{-\infty}^{\infty} + \ c \cdot \frac{1}{2\pi}\int_{-\infty}^{\infty} v(x,t)(-i\xi)e^{-i\xi x}dx$$

$\hat{v}_t(\xi,t) = -ic\xi\hat{v}(\xi,t)$, $\hat{v}(\xi,0) = \hat{f}(\xi) \Rightarrow \hat{v}(\xi,t) = \hat{f}(\xi)e^{-ic\xi t}$

$$v(x,t) = \int_{-\infty}^{\infty} \hat{v}(\xi,t)e^{i\xi x}d\xi = \int_{-\infty}^{\infty} \hat{f}(\xi)e^{-ic\xi t}e^{i\xi x}d\xi$$

$$= \int_{-\infty}^{\infty} \hat{f}(\xi)e^{i\xi(x-ct)}d\xi = f(x-ct) \ , \ ||v(\cdot,t)||_2 = ||f||_2 \quad \text{ok}$$

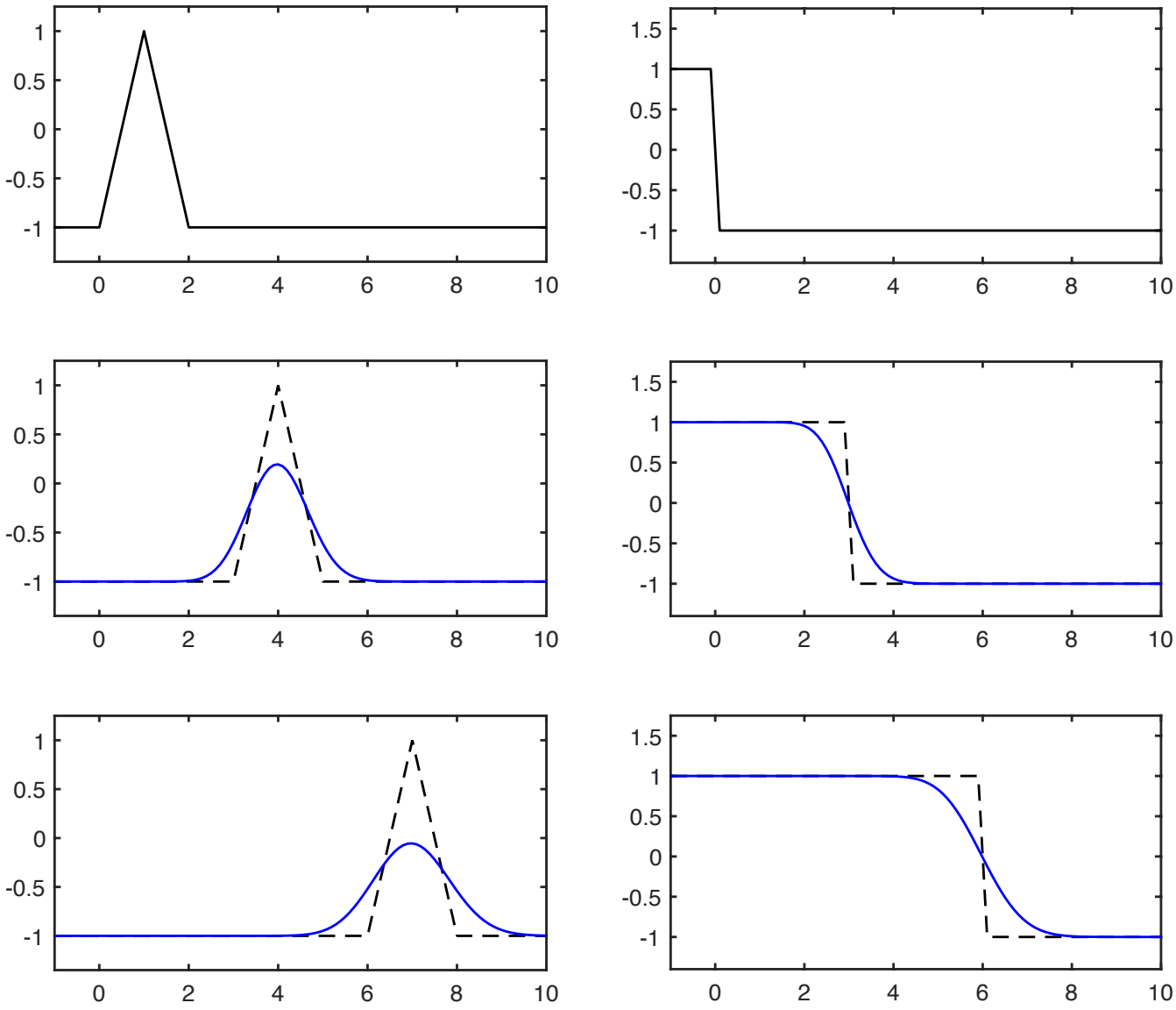
example

$v_t + cv_x = 0$: scalar wave equation , assume $c > 0$

$u_j^{n+1} = (1 - c\lambda)u_j^n + c\lambda u_{j-1}^n$, $\lambda = k/h$: upwind scheme

$\phi_t + c\phi_x = \frac{1}{2}hc(1 - c\lambda)\phi_{xx}$: model equation

$c = 1$, $k = 0.01$, $h = 0.1$, $c\lambda = 0.1$, $\frac{1}{2}hc(1 - c\lambda) = 0.045$



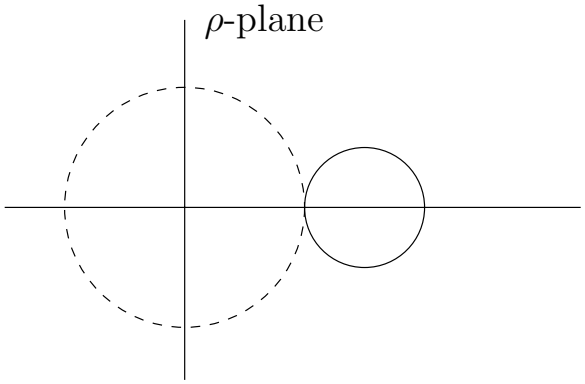
Fourier analysis of difference schemes

look for solutions of the form $u_j^n = \rho(\xi h)^n e^{ij\xi h}$, $-\pi \le \xi h \le \pi$

downwind scheme

$u_j^{n+1} = (1 + c\lambda)u_j^n - c\lambda u_{j+1}^n \Rightarrow \rho(\xi h) = 1 + c\lambda - c\lambda e^{i\xi h}$

$\rho(\xi h)$ maps $[-\pi, \pi]$ into the circle centered at $1 + c\lambda$ with radius $c\lambda$



$\max|\rho(\xi h)| = 1 + 2c\lambda \Rightarrow \begin{cases} \text{unconditionally unstable in 2-norm} \\ \text{short waves } (\xi h = \pm\pi) \text{ are amplified the most} \end{cases}$

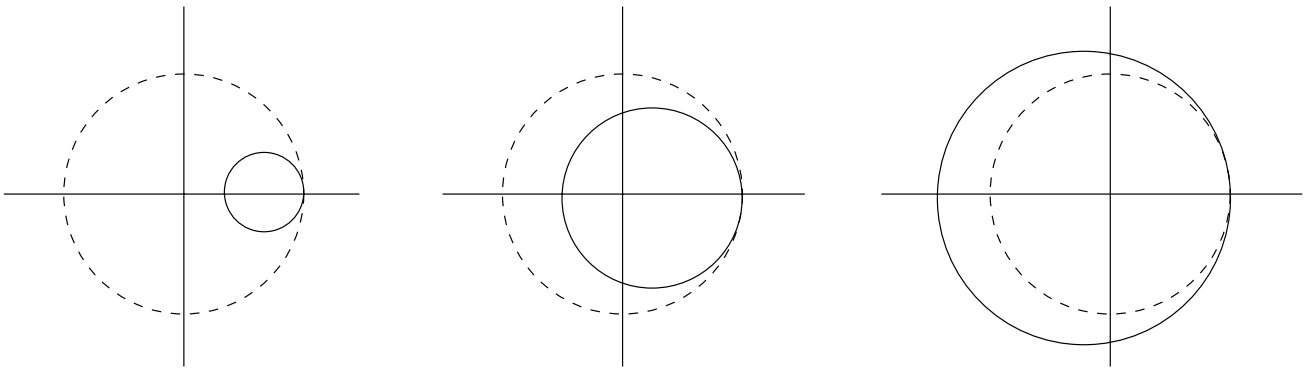
---

upwind scheme

$u_j^{n+1} = (1 - c\lambda)u_j^n + c\lambda u_{j-1}^n \Rightarrow \rho(\xi h) = 1 - c\lambda + c\lambda e^{-i\xi h}$

$\rho(\xi h)$ maps $[-\pi, \pi]$ into the circle centered at $1 - c\lambda$ with radius $c\lambda$

$0 < c\lambda < 1/2$ $\qquad$ $1/2 < c\lambda < 1$ $\qquad$ $c\lambda > 1$



$0 < c\lambda < 1 \Rightarrow |\rho(\xi h)| \le 1 \Rightarrow \begin{cases} \text{scheme is stable in 2-norm} \\ \text{short waves are damped the most} \end{cases}$

$c\lambda > 1 \Rightarrow \max|\rho(\xi h)| = 2c\lambda - 1 \Rightarrow \begin{cases} \text{scheme is unstable in 2-norm} \\ \text{short waves are amplified the most} \end{cases}$

question : In a system with positive and negative wave speeds, e.g. $v_{tt} = c^2 v_{xx}$, there is no unique upwind direction; is there is a stable scheme in this case?

central difference scheme

$v_t + cv_x = 0$, where c can be positive or negative

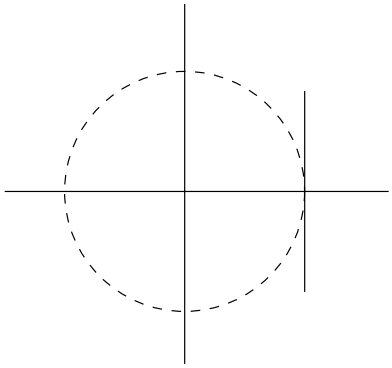
$$\frac{u_j^{n+1} - u_j^n}{k} + cD_0 u_j^n = 0 \ , \ D_0 u_j^n = \frac{u_{j+1}^n - u_{j-1}^n}{2h}$$

$$u_j^{n+1} = u_j^n - \tfrac{1}{2} c\lambda (u_{j+1}^n - u_{j-1}^n)$$

theorem

1. The CFL condition is satisfied if $|c|\lambda \leq 1$.
2. The scheme is unconditionally unstable in the 2-norm.

proof 1. ... ok , 2. $\rho(\xi h) = 1 - \frac{1}{2} c\lambda (e^{i\xi h} - e^{-i\xi h}) = 1 - ic\lambda \sin \xi h$ ok



Lax-Friedrichs

25
Tues
4/16

$v_t + cv_x = 0$

$$\frac{u_j^{n+1} - \frac{1}{2}(u_{j+1}^n + u_{j-1}^n)}{k} + cD_0 u_j^n = 0$$

$$\frac{u_j^{n+1} - u_j^n}{k} + cD_0 u_j^n = -\frac{1}{k} u_n^n + \frac{1}{2k}(u_{j+1}^n + u_{j-1}^n) = \frac{h^2}{2k} D_+ D_- u_j^n$$

$$\frac{h^2}{2k} = \frac{h}{2\lambda} = \frac{k}{2\lambda^2}$$

theorem

1. The CFL condition is satisfied if $|c|\lambda \leq 1$.
2. LF is stable in the 2-norm if $|c|\lambda \leq 1$.
3. LF is 1st order accurate, i.e. error $= O(k)$.
4. The model equation for LF is $\phi_t + c\phi_x = \dfrac{h}{2\lambda}(1 - c^2\lambda^2)\phi_{xx}$. , proof : hw

Hence the central difference scheme is stabilized by adding artificial viscosity.

question : can we get 2nd order accuracy and still keep the scheme explicit?

Lax-Wendroff

$v_t + cv_x = 0 \Rightarrow v_t = -cv_x \ , \ v_{tt} = c^2 v_{xx}$

$$\frac{v_j^{n+1} - v_j^n}{k} = \frac{v + kv_t + \frac{1}{2}k^2 v_{tt} + O(k^3) - v}{k} = v_t + \tfrac{1}{2}kv_{tt} + O(k^2)$$

$$= -cv_x + \tfrac{1}{2}kc^2 v_{xx} + O(k^2)$$

$$\frac{u_j^{n+1} - u_j^n}{k} + cD_0 u_j^n = \tfrac{1}{2}kc^2 D_+ D_- u_j^n$$

Hence the central difference scheme is stabilized by adding artificial viscosity and the viscosity coefficient is chosen to obtain 2nd order accuracy.

theorem

1. The CFL condition is satisfied if $|c|\lambda \leq 1$.
2. LW is stable in the 2-norm if $|c|\lambda \leq 1$.
3. LW is 2nd order accurate, i.e. error $= O(k^2)$.
4. The model equation for LW is $\phi_t + c\phi_x = \underbrace{\frac{1}{6}ch^2(1 - c^2\lambda^2)\phi_{xxx}}$.

proof (partial) artificial dispersion (more later)

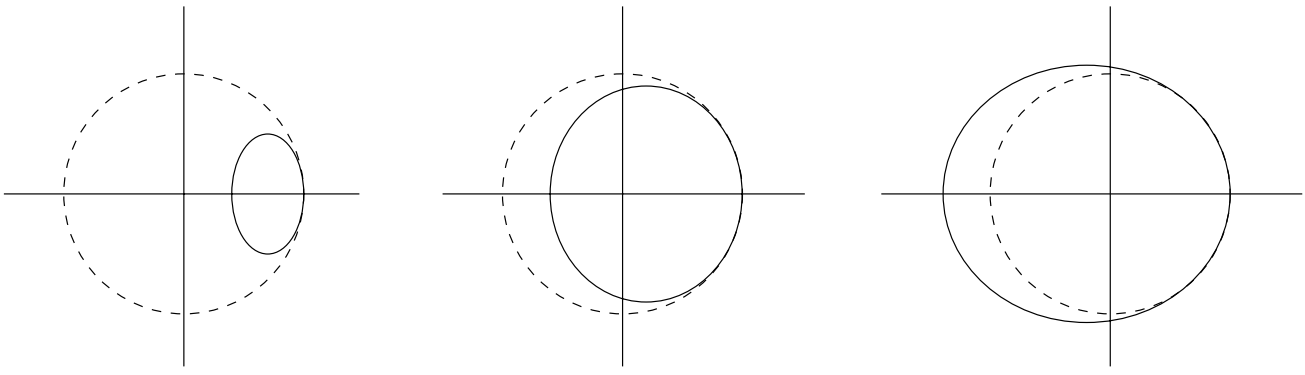
$u_j^{n+1} = (1 - c^2\lambda^2)u_j^n - \frac{1}{2}c\lambda(1 - c\lambda)u_{j+1}^n + \frac{1}{2}c\lambda(1 + c\lambda)u_{j-1}^n$

$\rho(\xi h) = 1 - c^2\lambda^2 - \frac{1}{2}c\lambda(1 - c\lambda)e^{i\xi h} + \frac{1}{2}c\lambda(1 + c\lambda)e^{-i\xi h}$

$= 1 - ic\lambda \sin \xi h - c^2\lambda^2(1 - \cos \xi h)$: compare to central difference scheme

$= 1 - c^2\lambda^2 + c^2\lambda^2 \cos \xi h - ic\lambda \sin \xi h$: ellipse

$0 < |c|\lambda < 1/2$ $\quad$ $1/2 < |c|\lambda < 1$ $\quad$ $|c|\lambda > 1$

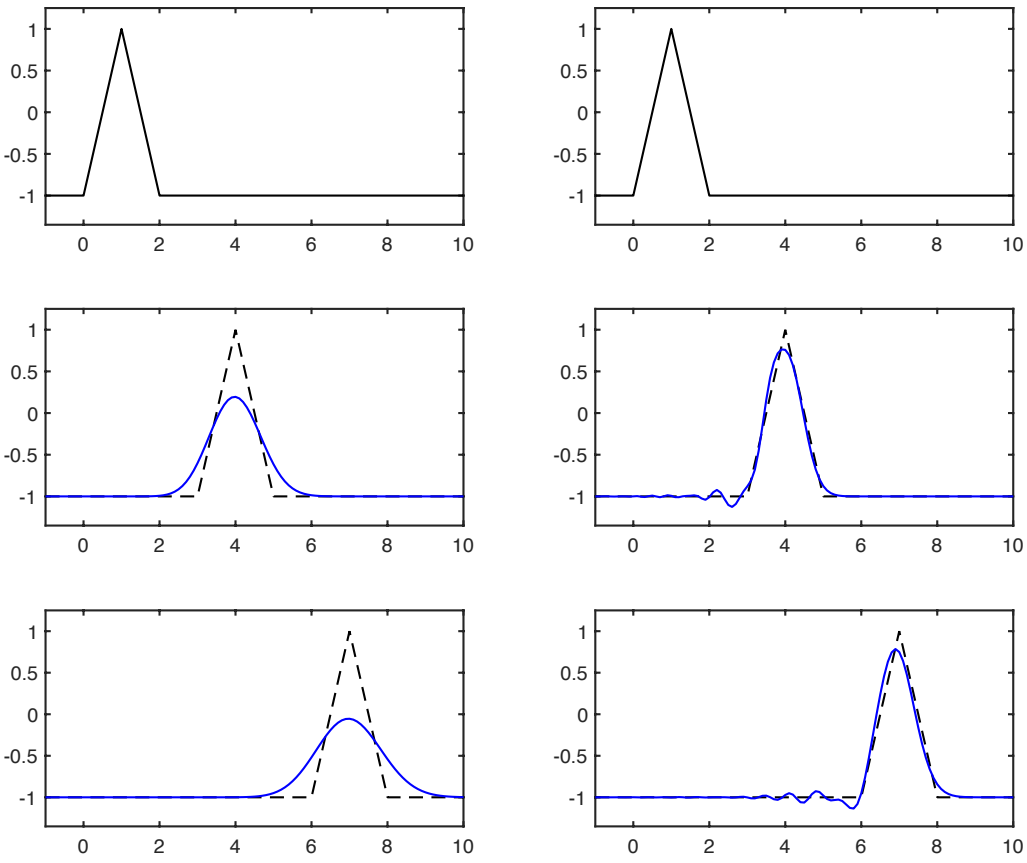


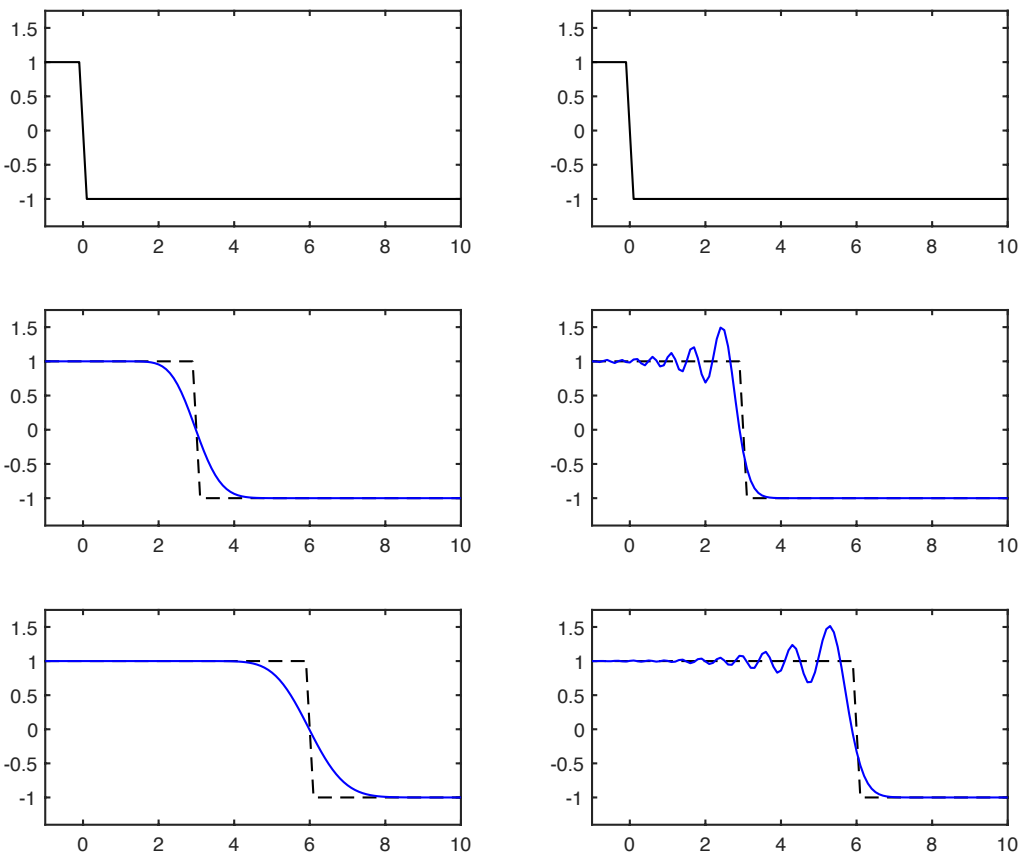
1. $\rho(\xi h) = 1 - ic\lambda \sin \xi h - 2c^2\lambda^2 \sin^2(\xi h/2)$
2. If $|c|\lambda \leq 1$, upwind and LF are stable in the ∞-norm, but LW is not.

$c = 1$
$k = 0.01$
$h = 0.1$
$c\lambda = 0.1$

upwind scheme

Lax-Wendroff





phase error

compare : $v(x,t) = e^{i\xi(x-ct)}$, $u_j^n = \rho(\xi h)^n e^{ij\xi h}$

$\rho(\xi h) = |\rho(\xi h)| e^{i\theta(\xi h)}$, $\theta(\xi h) = \arg \rho(\xi h)$, $-\pi < \theta(\xi h) < \pi$

$u_j^n = |\rho(\xi h)|^n e^{in\theta(\xi h)} e^{ij\xi h} = |\rho(\xi h)|^n e^{i(j\xi h + n\theta(\xi h))}$

$$i(j\xi h + n\theta(\xi h)) = i\xi\left(x_j + \frac{t_n\,\theta(\xi h)}{k\xi}\right) = i\xi(x_j - \tilde{c}\,t_n)$$

$$\tilde{c} = \tilde{c}(\xi h) = -\frac{\theta(\xi h)}{k\xi} = -\frac{\theta(\xi h)}{\lambda\,\xi h}$$: numerical wave speed

$u_j^n = |\rho(\xi h)|^n e^{i\xi(x_j - \tilde{c}\,t_n)}$: discrete traveling wave (amplitude can grow/decay)

definition : A PDE or numerical method is called dispersive if it has traveling wave solutions for which the wave speed depends on the wavelength.

1. $v_t + cv_x = 0$: non-dispersive

$v(x,t) = e^{i\xi(x-ct)}$: all waves travel at the same speed

2. $\phi_t = \phi_{xxx}$: dispersive

$\phi(x,t) = e^{\omega t + i\xi x} \Rightarrow \omega = (i\xi)^3 = -i\xi^3 \Rightarrow \phi(x,t) = e^{i\xi(x-\xi^2 t)}$

$\Rightarrow c = c(\xi) = \xi^2$: short waves ($\xi \to \pm\infty$) travel faster than long waves ($\xi \to 0$)

3. Lax-Wendroff

26 Thurs 4/18

The dispersive character of LW is already seen in the model equation and numerical results; here it is analyzed through the numerical wave speed.

$$\rho(\xi h) = 1 - ic\lambda \sin \xi h - 2c^2\lambda^2 \sin^2(\xi h/2)\ ,\ \tilde{c}(\xi h) = -\frac{\theta(\xi h}{\lambda \xi h} = ?$$

consider long waves , $\xi h \to 0$

$$\rho(\xi h) = 1 - ic\lambda(\xi h - \tfrac{1}{6}(\xi h)^3 + \cdots) - 2c^2\lambda^2 \cdot (\xi h/2)^2 + \cdots$$

$$\tan\theta = \frac{-c\lambda\xi h + \frac{1}{6}c\lambda(\xi h)^3 + \cdots}{1 - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots}$$

$$= (-c\lambda\xi h + \tfrac{1}{6}c\lambda(\xi h)^3 + \cdots)(1 + \tfrac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots)$$

$$= -c\lambda\xi h + (\tfrac{1}{6}c\lambda - \tfrac{1}{2}c^3\lambda^3)(\xi h)^3 + \cdots$$

$$= \alpha_1\epsilon + \alpha_2\epsilon^2 + \alpha_3\epsilon^3 + \cdots\ ,\ \epsilon = \xi h$$

$$\theta = \beta_1\epsilon + \beta_2\epsilon^2 + \beta_3\epsilon^3 + \cdots$$

$$\tan\theta = \theta + \tfrac{1}{3}\theta^3 + \cdots$$

$$= (\beta_1\epsilon + \beta_2\epsilon^2 + \beta_3\epsilon^3 + \cdots) + \tfrac{1}{3}(\beta_1\epsilon + \cdots)^3 + \cdots$$

$$= \beta_1\epsilon + \beta_2\epsilon^2 + (\beta_3 + \tfrac{1}{3}\beta_1^3)\epsilon^3 + \cdots$$

$\beta_1 = \alpha_1 = -c\lambda \ , \ \beta_2 = \alpha_2 = 0$

$\beta_3 + \frac{1}{3}\beta_1^3 = \alpha_3 \Rightarrow \beta_3 = (\frac{1}{6}c\lambda - \frac{1}{2}c^3\lambda^3) - \frac{1}{3}(-c\lambda)^3 = \frac{1}{6}c\lambda(1 - c^2\lambda^2)$

$\theta = -c\lambda\xi h + \frac{1}{6}c\lambda(1 - c^2\lambda^2)(\xi h)^3 + \cdots$

$\tilde{c}(\xi h) = \dfrac{\theta(\xi h)}{-\lambda\xi h} \ = \ c(1 - \frac{1}{6}(1 - c^2\lambda^2)(\xi h)^2 \ + \ \cdots)$

Hence discrete traveling waves have the correct speed in the long wave limit ($\xi h \to 0$), but for $\xi h \neq 0$ and $|c|\lambda < 1$, their speed is less than the exact speed; this explains the oscillations in the numerical solution.

<u>convection-diffusion equation</u>

$v_t + cv_x = \epsilon v_{xx}$, assume $c > 0$

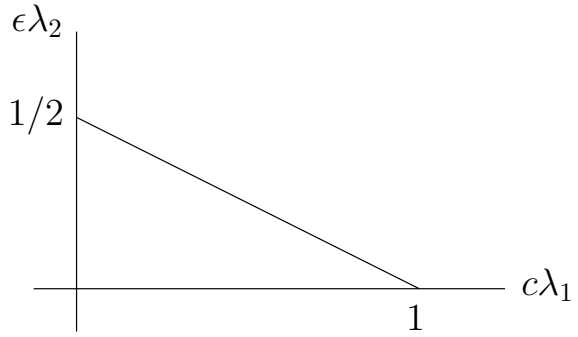
$\dfrac{u_j^{n+1} - u_j^n}{k} + cD_-u_j^n = \epsilon D_+D_-u_j^n$: upwind/central difference

$u_j^{n+1} = u_j^n - c\lambda_1(u_j^n - u_{j-1}^n) + \epsilon\lambda_2(u_{j+1}^n - 2u_j^n + u_{j-1}^n)$

$\lambda_1 = k/h \ , \ \lambda_2 = k/h^2$

$u_j^{n+1} = (1 - c\lambda_1 - 2\epsilon\lambda_2)u_j^n + (c\lambda_1 + \epsilon\lambda_2)u_{j-1}^n + \epsilon\lambda_2 u_{j+1}^n$

then $||u^{n+1}||_\infty \leq ||u^n||_\infty \Leftrightarrow 1 - c\lambda_1 - 2\epsilon\lambda_2 \geq 0$, proof ...



$$\Rightarrow \begin{cases} c\lambda_1 \leq 1 - 2\epsilon\lambda_2 \ : \text{ more restrictive than for } v_t + cv_x = 0\,,\ c\lambda_1 \leq 1 \\ \epsilon\lambda_2 \leq \frac{1}{2} - \frac{1}{2}c\lambda_1 \ : \ \cdots\cdots\cdots\text{''}\cdots\cdots\cdots \ v_t = \epsilon v_{xx}\,,\ \epsilon\lambda_2 \leq 1/2 \end{cases}$$

note : if $\epsilon << 1$, then artificial viscosity may overwhelm physical viscosity

systems

$v_t + Av_x = 0\ ,\ v = v(x,t) = \begin{pmatrix} v_1(x,t) \\ \vdots \\ v_p(x,t) \end{pmatrix}$

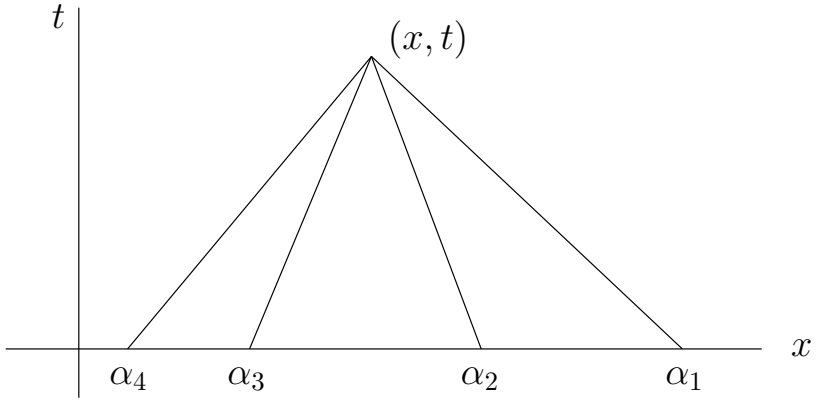
definition : the system is called hyperbolic if A is diagonalizable, $A = TDT^{-1}$, where $D = \text{diag}(c_1, \ldots, c_p)$ and the c_i are real

then $v_t + TDT^{-1}v_x = 0 \Rightarrow T^{-1}v_t + DT^{-1}v_x = 0$

set $w = T^{-1}v$, then $w_t + Dw_x = 0 \Rightarrow (w_i)_t + c_i(w_i)_x = 0$ $\begin{cases} \text{decoupled scalar} \\ \text{wave equations} \\ \text{with wave speeds } c_i \end{cases}$

characteristics : $x - c_i t = \alpha_i\ ,\ w_i(x,t) = w_i(x - c_i t, 0)\ ,\ v_i(x,t) = (Tw)_i(x,t)$

example : $p = 4\ ,\ c_1 < c_2 < 0 < c_3 < c_4 \Rightarrow c_2^{-1} < c_1^{-1} < 0 < c_4^{-1} < c_3^{-1}$: slopes



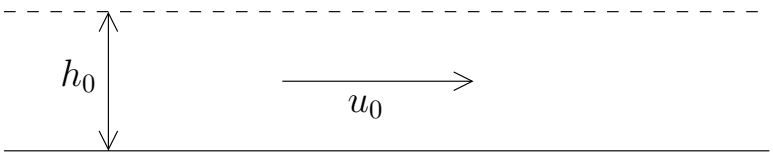
domain of dependence of $v(x,t)$: $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$

27
Tues
4/23

consider Lax-Wendroff : $u_j^{n+1} = u_j^n - kAD_0u_j^n + \frac{1}{2}k^2A^2D_+D_-u_j^n$

CFL condition is satisfied if $\max_i |c_i|\lambda \leq 1$

example : linearized shallow water equations



$h_0\ ,\ u_0$: equilibrium free surface height, horizontal velocity

$h(x,t)\ ,\ u(x,t)$: perturbations

conservation of mass, momentum

$\begin{matrix} h_t + u_0h_x + h_0u_x = 0 \\ u_t + u_0u_x + gh_x \ \ = 0 \end{matrix} \Rightarrow \begin{pmatrix} h \\ u \end{pmatrix}_t + \begin{pmatrix} u_0 & h_0 \\ g & u_0 \end{pmatrix} \begin{pmatrix} h \\ u \end{pmatrix}_x = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

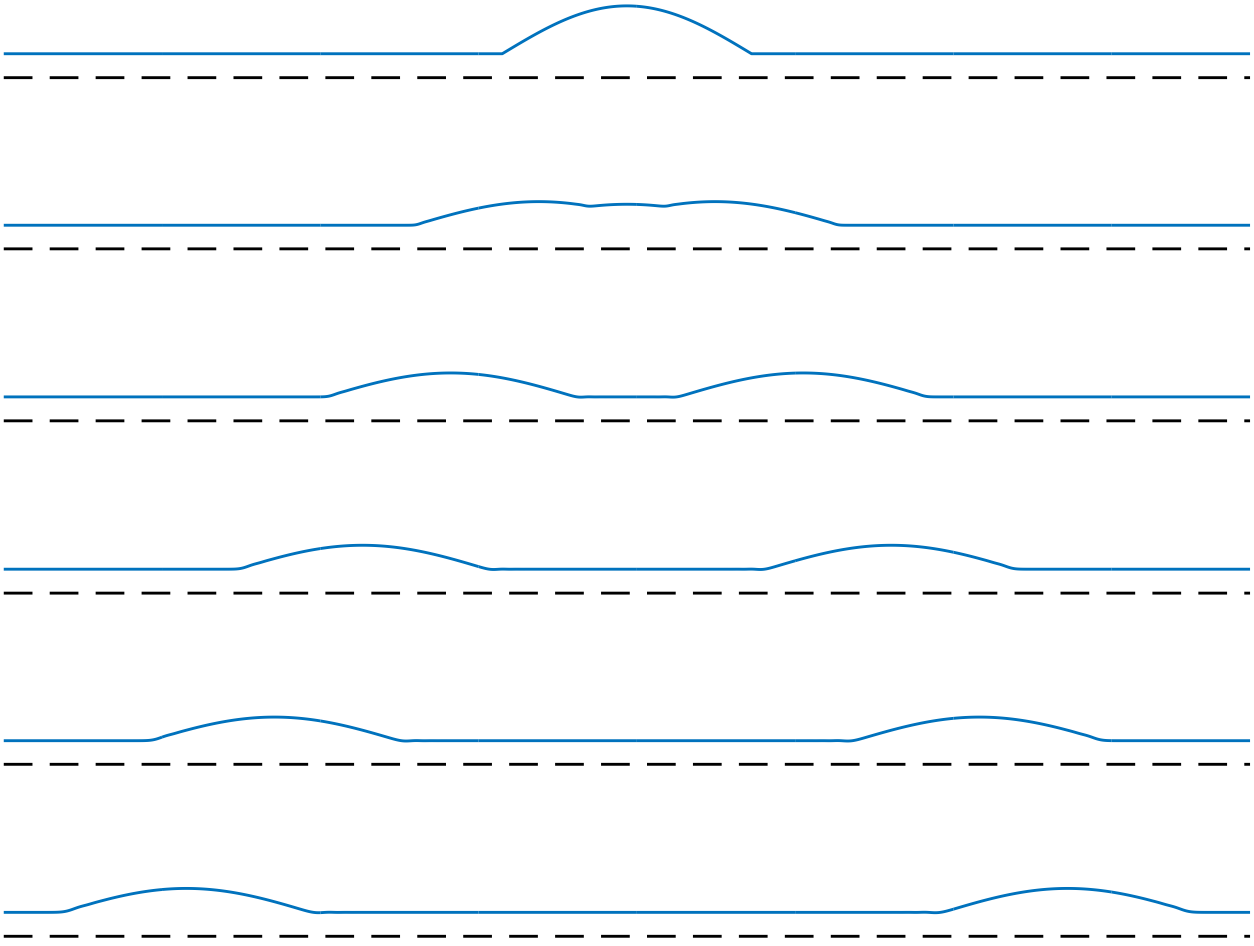
$A = \begin{pmatrix} u_0 & h_0 \\ g & u_0 \end{pmatrix} \Rightarrow \det(A - cI) = (u_0 - c)^2 - gh_0 \Rightarrow c_{1,2} = u_0 \pm \sqrt{gh_0}$

linearized shallow water equations

numerical solution by Lax-Wendroff scheme

dashed black line : ocean bottom

solid blue line : free surface height



cutoff for final exam

2-norm stability for hyperbolic systems

$v_t + Av_x = 0\,,\ v = v(x,t) = \begin{pmatrix} v_1(x,t) \\ \vdots \\ v_p(x,t) \end{pmatrix}$

consider $u_j^{n+1} = Qu_j^n\ :\ Q = Q_1 S_+ + Q_0 I + Q_{-1} S_-\ :$ 3-point difference operator

example

$Q_{LW} = -\frac{1}{2}\lambda A\,(I - \lambda A)\,S_+ + \left(I - \lambda^2 A^2\right) I + \frac{1}{2}\lambda A\,(I + \lambda A)\,S_-$

Fourier analysis

$u_j^n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{u}^n(\xi h) e^{-ij\xi h} d(\xi h)\ ,\ \widehat{u}^n(mh) = \sum_{j=-\infty}^{\infty} u_j^n\, e^{ijmh}$

$\widehat{u}^{n+1}(\xi h) = G(\xi h)\widehat{u}^n(\xi h)\,,\ G(\xi h) = Q_1 e^{i\xi h} + Q_0 + Q_{-1} e^{-i\xi h}\ :$ amplification matrix

$\widehat{u}^n(\xi h) = G^n(\xi h)\widehat{u}^0(\xi h)$

$||u^n||_2^2 = \sum_{j=-\infty}^{\infty} ||u_j^n||_2^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} ||\widehat{u}^n(\xi h)||_2^2\, d(\xi h) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ||G^n(\xi h)\widehat{u}^0(\xi h)||_2^2\, d(\xi h)$

$\leq \max_{\xi h} ||G^n(\xi h)||_2^2 \cdot \frac{1}{2\pi}\int_{-\pi}^{\pi} ||\widehat{u}^0(\xi h)||_2^2\, d(\xi h)$

$||u^n||_2 \leq \max_{\xi h} ||G^n(\xi h)||_2 \cdot ||u^0||_2$

definition

$G(\xi h)$ is uniformly power bounded if $\max_{\xi h} ||G^n(\xi h)||_2 \leq K$ for all $n \geq 0$, where K is independent of h, k, n (but may depend on $\lambda = k/h$ and $t = nk$).

note : The scheme is stable in the 2-norm $\Leftrightarrow$ $G(\xi h)$ is unif power bdd.

example: Lax-Wendroff for hyperbolic systems

$G(\xi h) = -\frac{1}{2}\lambda A\,(I - \lambda A)\,e^{i\xi h} + \left(I - \lambda^2 A^2\right) + \frac{1}{2}\lambda A\,(I + \lambda A)\,e^{-i\xi h}$

$= I - i\lambda A \sin \xi h - 2\lambda^2 A^2 \sin^2(\xi h/2)\,,\ A = TDT^{-1}$

$= T(I - i\lambda D \sin \xi h - 2\lambda^2 D^2 \sin^2(\xi h/2))T^{-1}$

$G^n(\xi h) = T(I - i\lambda D \sin \xi h - 2\lambda^2 D^2 \sin^2(\xi h/2))^n T^{-1}$

$||G^n(\xi h)||_2 \leq ||T||_2 \cdot ||T^{-1}||_2$ if $\max_i |c_i|\lambda \leq 1$

$\Rightarrow G(\xi h)$ is unif power bdd and hence LW for a hyperbolic system is stable in the 2-norm if the CFL condition is satisfied.

example : leap-frog/central difference scheme for scalar wave equation

$v_t + cv_x = 0$

$\dfrac{u_j^{n+1} - u_j^{n-1}}{2k} + cD_0 u_j^n = 0$: 2-step method

CFL is satisfied if $|c|\lambda \leq 1$; does this ensure stability in the 2-norm?

$\begin{pmatrix} u_j^{n+1} \\ u_j^n \end{pmatrix} = \begin{pmatrix} -2ckD_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u_j^n \\ u_j^{n-1} \end{pmatrix}$

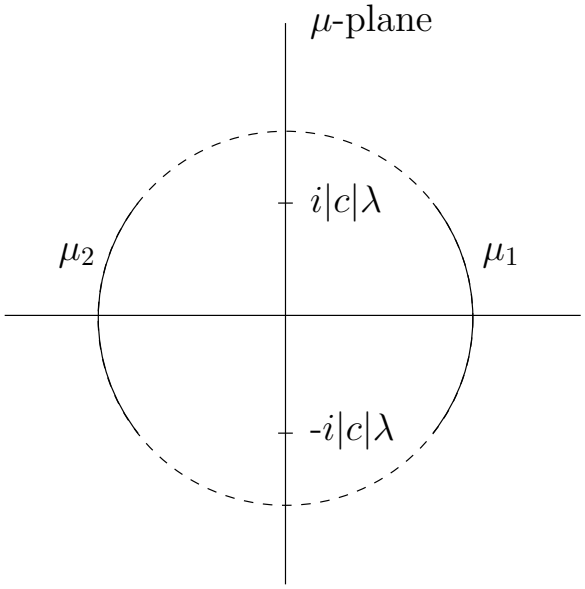
$u_j^{n+1} = Qu_j^n \ , \ Q = Q_1 S_+ + Q_0 I + Q_{-1} S_-$

$\widehat{u}^{n+1} = G(\xi h)\widehat{u}^n \ , \ G(\xi h) = Q_1 e^{i\xi h} + Q_0 + Q_{-1} e^{-i\xi h} = \begin{pmatrix} -2ic\lambda \sin \xi h & 1 \\ 1 & 0 \end{pmatrix}$

question : is $G(\xi h)$ unif power bdd?

$\det\left(G(\xi h) - \mu I\right) = \mu^2 + 2ic\lambda \sin \xi h \cdot \mu - 1 = (\mu - \mu_1)(\mu - \mu_2) = 0$

$\mu_{1,2} = -ic\lambda \sin \xi h \pm \sqrt{1 - c^2\lambda^2 \sin^2 \xi h} \Rightarrow |\mu| = 1$ if $|c|\lambda \leq 1$



case 1 : $|c|\lambda < 1$

$\mu_1 \neq \mu_2 \ , \ |\mu_1| = |\mu_2| = 1 \ , \ \mu_1 + \mu_2 = -2ic\lambda \sin \xi h \ , \ \mu_1 \, \mu_2 = -1$

$G(\xi h) = \begin{pmatrix} \mu_1 + \mu_2 & 1 \\ 1 & 0 \end{pmatrix} = TDT^{-1}$

$T = \begin{pmatrix} 1 & 1 \\ -\mu_2 & -\mu_1 \end{pmatrix} \ , \ D = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} \ , \ T^{-1} = \dfrac{1}{\mu_2 - \mu_1} \begin{pmatrix} -\mu_1 & -1 \\ \mu_2 & 1 \end{pmatrix}$

$G^n(\xi h) = TD^n T^{-1} \Rightarrow ||G^n(\xi h)||_2 \leq ||T||_2 \cdot ||T^{-1}||_2$

theorem : If A is a $p \times p$ matrix, then $||A||_2 \leq \sqrt{p}\,||A||_\infty$, $||A||_\infty \leq \sqrt{p}\,||A||_2$.

proof

1. $||Ax||_2^2 = \sum_{i=1}^{p} |(Ax)_i|^2 \leq p\,||Ax||_\infty^2 \leq p\,(\,||A||_\infty \cdot ||x||_\infty)^2 \leq p\,||A||_\infty^2 \cdot ||x||_2^2$

$\Rightarrow \dfrac{||Ax||_2}{||x||_2} \leq \sqrt{p}\;||A||_\infty$ ok

2. ...

$||T||_\infty \leq 2$, $||T||_2 \leq \sqrt{2} \cdot 2$

$||T^{-1}||_\infty \leq \dfrac{2}{|\mu_2 - \mu_1|} = \dfrac{1}{\sqrt{1 - c^2\lambda^2 \sin^2 \xi h}} \leq \dfrac{1}{\sqrt{1 - c^2\lambda^2}}$

$||T^{-1}||_2 \leq \sqrt{2} \cdot \dfrac{1}{\sqrt{1 - c^2\lambda^2}}$

$\Rightarrow \max\limits_{\xi h} ||G^n(\xi h)||_2 \leq \dfrac{4}{\sqrt{1 - c^2\lambda^2}}$: unif power bdd if $|c|\lambda < 1$

case 2 : $|c|\lambda = 1$

previous bound fails , for example consider $c\lambda = -1$, $\xi h = \pi/2$

$$G(\pi/2) = \begin{pmatrix} 2i & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -i \end{pmatrix}$$

$$\begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix} = i \begin{pmatrix} 1 & -i \\ 0 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix}^n = i^n \begin{pmatrix} 1 & -ni \\ 0 & 1 \end{pmatrix}$$

$$G^n(\pi/2) = \begin{pmatrix} i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix}^n \begin{pmatrix} 0 & 1 \\ 1 & -i \end{pmatrix} = \ldots = i^n \begin{pmatrix} n+1 & -ni \\ -ni & 1-n \end{pmatrix}$$

$\max\limits_{\xi h} ||G^n(\xi h)||_2 \geq ||G^n(\pi/2)||_2 \geq \frac{1}{\sqrt{2}}\,||G^n(\pi/2)||_\infty = \frac{2n+1}{\sqrt{2}}$: not unif power bdd

Hence the leap-frog/central difference scheme for the scalar wave equation is stable in the 2-norm $\Leftrightarrow |c|\lambda < 1$; this also holds for hyperbolic systems.

note

1. Fourier method $\Rightarrow ||u^{n+1}||_2^2 + ||u^n||_2^2 \leq \dfrac{16}{1 - c^2\lambda^2} \left(||u^1||_2^2 + ||u^0||_2^2\right)$

2. energy method $\Rightarrow ||u^{n+1}||_2^2 + ||u^n||_2^2 \leq \dfrac{1 + |c|\lambda}{1 - |c|\lambda} \left(||u^1||_2^2 + ||u^0||_2^2\right)$

proof

set $L_n = ||u^n||_2^2 + ||u^{n-1}||_2^2 + 2kc(u^n, D_0 u^{n-1})$, show $L_{n+1} = L_n$ for $n \geq 1$ ...

numerical wave speed for leap-frog scheme

$v_t + cv_x = 0$

$\dfrac{u_j^{n+1} - u_j^{n-1}}{2k} + cD_0 u_j^n = 0$, needs u_j^0 , u_j^1 to start

$$\begin{pmatrix} u_j^{n+1} \\ u_j^n \end{pmatrix} = \begin{pmatrix} -2ckD_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u_j^n \\ u_j^{n-1} \end{pmatrix}$$

$$\begin{pmatrix} \widehat{u}^{n+1}(\xi h) \\ \widehat{u}^n(\xi h) \end{pmatrix} = \begin{pmatrix} -2ic\lambda \sin \xi h & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \widehat{u}^n(\xi h) \\ \widehat{u}^{n-1}(\xi h) \end{pmatrix}$$

e-values of $G(\xi h)$: $\mu = -ic\lambda \sin \xi h \pm \sqrt{1 - c^2\lambda^2 \sin^2 \xi h}$

We showed that $G(\xi h)$ is unif power bdd $\Leftrightarrow |c|\lambda < 1$, but even when the scheme is stable, there is still a problem due to the extraneous root.

recall : numerical wave speed $= \tilde{c} = -\dfrac{\theta(\xi h)}{\lambda \xi h}$, $\theta(\xi h) = \arg \rho(\xi h)$

$\mu_1 = -ic\lambda \sin \xi h + \sqrt{1 - c^2\lambda^2 \sin^2 \xi h}$

$= -ic\lambda(\xi h - \frac{1}{6}(\xi h)^3 + \cdots) + 1 - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots$

$\tan \theta_1 = \dfrac{-c\lambda\xi h + \frac{1}{6}c\lambda(\xi h)^3 + \cdots}{1 - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots}$: same as LW

$\tilde{c}_1 = c(1 - \frac{1}{6}(1 - c^2\lambda^2)(\xi h)^2 + \cdots)$

$\mu_2 = -ic\lambda \sin \xi h - \sqrt{1 - c^2\lambda^2 \sin^2 \xi h}$

$= -ic\lambda(\xi h - \frac{1}{6}(\xi h)^3 + \cdots) - (1 - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots)$

$\tan \theta_2 = \dfrac{-c\lambda\xi h + \frac{1}{6}c\lambda(\xi h)^3 + \cdots}{-1 + \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots}$

$= (-c\lambda\xi h + \frac{1}{6}c\lambda(\xi h)^3 + \cdots)(-1 - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots)$

$\tilde{c}_2 = c(-1 + \frac{1}{6}(1 - c^2\lambda^2)(\xi h)^2 + \cdots)$: wrong direction

recall : the numerical solution has components of the form $u_j^n = \mu^n e^{ij\xi h}$

$\mu_1 = 1 - ic\lambda\xi h - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots = e^{-ic\lambda\xi h} + \cdots$

$u_j^n = \mu_1^n e^{ij\xi h} \approx e^{i\xi(x_j - \tilde{c}_1 t_n)} \approx$ genuine solution of PDE

$\mu_2 = (-1)(1 + ic\lambda\xi h - \frac{1}{2}c^2\lambda^2(\xi h)^2 + \cdots) = -e^{ic\lambda\xi h} + \cdots$

$u_j^n = \mu_2^n e^{ij\xi h} \approx (-1)^n e^{i\xi(x_j - \tilde{c}_2 t_n)}$: spurious solution, oscillates in time and travels in the wrong direction

lower order terms

scalar equation

$v_t + cv_x = bv \ , \ v(x,0) = f(x)$

The characteristics are still the lines $x - ct = \alpha$ in the xt-plane.

$\frac{d}{dt}v(\alpha + ct, t) = v_x(\alpha + ct, t) \cdot c + v_t(\alpha + ct, t) = bv(\alpha + ct, t)$

$\Rightarrow v(\alpha + ct, t) = e^{bt}v(\alpha, 0) \Rightarrow v(x,t) = e^{bt}f(x - ct) \quad$ check $\ \ldots$

$||v(\cdot, t)||_2 = e^{bt}||f||_2 \ :$ energy of exact solution changes in time

for example consider $u_j^{n+1} = u_j^n - kcD_0u_j^n + \frac{1}{2}k^2c^2D_+D_-u_j^n + kbu_j^n$

$\rho(\xi h) = \rho_0(\xi h) + kb \ , \ \rho_0(\xi h) \ :$ LW

recall $:\ |c|\lambda \leq 1 \Rightarrow |\rho_0(\xi h)| \leq 1 \Rightarrow |\rho(\xi h)| \leq 1 + k|b|$

$\rho^n(\xi h) \leq (1 + k|b|)^n \leq \left(e^{k|b|}\right)^n = e^{|b|t} \ , \ t = nk$

$\widehat{u}^n(\xi h) = \rho(\xi h)^n \widehat{u}^0(\xi h) \Rightarrow ||u^n||_2 \leq e^{|b|t}||u^0||_2 \ :$ stability

system

$v_t + Av_x = Bv \ , \ A = TDT^{-1}$

$u_j^{n+1} = u_j^n - kAD_0u_j^n + \frac{1}{2}k^2A^2D_+D_-u_j^n + kBu_j^n$

$G(\xi h) = G_0(\xi h) + kB \ , \ G_0(\xi h) \ :$ LW

goal $:$ show that $G(\xi h)$ is uniformly power bounded

recall $:\ \max_i |c_i|\lambda \leq 1 \Rightarrow ||G_0^n(\xi h)||_2 \leq ||T||_2 \cdot ||T^{-1}||_2 = g \ ,$ set $||B||_2 = b$

$||G^n||_2 = ||(G_0 + kB)^n||_2 \leq ||G_0 + kB||_2^n \leq (||G_0||_2 + k||B||_2)^n = (g + kb)^n$

$= (g(1 + kb/g))^n \leq g^n e^{(b/g)t} \ :$ ok if $g \leq 1 \ ,$ not ok if $g > 1$

case 1 $:\ G_0B = BG_0$

$G^n = (G_0 + kB)^n = \sum_{i=0}^{n} \binom{n}{i} G_0^{n-i}(kB)^i$

$||G^n||_2 \leq \sum_{i=0}^{n} \binom{n}{i} ||G_0^{n-i}||_2 \cdot ||(kB)^i||_2 \leq g \sum_{i=0}^{n} \binom{n}{i} (kb)^i \leq g(1 + kb)^n \leq ge^{bt} \quad$ ok

case 2 $:\ G_0B \neq BG_0$

$(G_0 + kB)^2 = G_0^2 + k(G_0B + BG_0) + k^2B^2$

$(G_0+kB)^3 = G_0^3 + k(G_0^2B + G_0BG_0 + BG_0^2) + k^2(G_0B^2 + BG_0B + B^2G_0) + k^3B^3$

$(G_0 + kB)^n = G_0^n + k \cdot \binom{n}{1}$ terms with one B and $(n-1)\,G_0$

$+ k^2 \cdot \binom{n}{2}$ terms with two B and $(n-2)\,G_0 + \cdots$

$||G^n||_2 \le ||(G_0 + kB)^n||_2 \le g + k\binom{n}{1}g^2b + k^2\binom{n}{2}g^3b^2 + \cdots$

$= g\sum_{i=0}^{n}\binom{n}{i}(kgb)^i = g(1+kgb)^n \le ge^{gbt}$ ok

summary : if $G_0(\xi h)$ is unif power bdd, then so is $G(\xi h) = G_0(\xi h) + kB$

further theoretical results

von Neumann condition : $\max_{\xi h} \rho(G(\xi h)) \le 1 + kb$

$\uparrow$

spectral radius (largest magnitude of an e-value)

claim

1. The vN condition is necessary for stability in the 2-norm.
2. If $p = 1$, the vN condition is necessary and sufficient for stability in the 2-norm.
3. If $p \ge 2$, the vN condition in not sufficient for stability in the 2-norm.

note

The Kreiss matrix theorem gives several conditions on $G(\xi h)$ which are necessary and sufficient for stability in the 2-norm.

Lax equivalence theorem

Consider a linear constant coefficient PDE of the form $v_t = A_0v + A_1v_x + A_2v_{xx}$ (for example) and a finite-difference scheme of the form $u^{n+1} = Qu^n$. If the initial value problem for the PDE is well-posed and the difference scheme is consistent, then stability is necessary and sufficient for convergence.