CAAM 452 - Numerical soln of PDEs 1/14

A mathematical model (PDE) that arises from modelling an oil spill is given by $\partial_t c - \nabla\cdot D\nabla c + \beta\cdot\nabla c = F$, where $c(\vec{x},t)$ measures the concentration of oil in a body of water. The discrepancy between the solution to the model & reality is called "validation".

defn

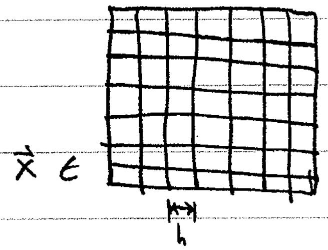
In many cases, no analytical soln can be obtained. We're left with attempting to solve the PDE numerically.

defn

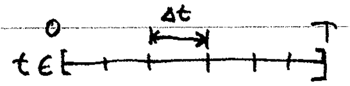
The study of "verification" captures the error in numerically solving the PDE.

How do we go about solving PDE's numerically?

Step 1: discretize both the domain & time interval



Choosing a good mesh is very difficult in real life if we'd like a more verified solution (resolution vs. comp. time)



Here, h & Δt don't have to be "uniform".

Step 2: pick a method for solving the PDE; this entails constructing a linear system

$$A\cdot U_{h,\Delta t} = \vec{b}$$ ← comes from F

↑ ↑ soln dep on h, Δt

comes from $\partial_t - \nabla\cdot D\nabla + \beta\cdot\nabla$, but depends on method.

Step 3: solve the system (in Matlab, $U_{h,\Delta t} = A\backslash\vec{b}$)

PETSC & TRILINOS applies conjugate gradient methods and Matlab uses the programmed method to implement $A\backslash\vec{b}$, e.g., if A is symmetric, L-U factorization

Step 4: post-processing, involving visualization or obtaining quantities of interest. If desired, and if the exact solution is known, we can test the method via $u_{exact} - U_{h,\Delta t}$.

Convergence of the numerical method means

$$\| u_{exact} - U_{h,\Delta t} \| \xrightarrow[\Delta t \to 0]{h \to 0} 0.$$

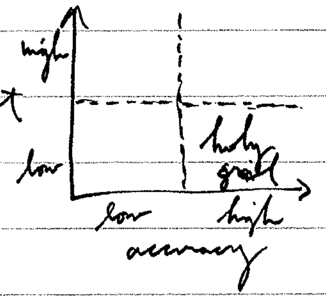
Order of convergence captures

$$\| error := (u_{exact} - U_{h,\Delta t}) \| \le C (h^r + \Delta t^q)$$

(so $r = q = 1$ → "first-order method"

and $r = q = 2$ → "second-order method")

Note that we are always concerned with cost vs. accuracy.



Research strives to improve accuracy, and then asks what is then the incurred cost.

Finite difference approximations

The main tool is the Taylor series expansion of u at a pt $\vec{x}$.

$$u(x+h) = u(x) + h \cdot u'(x) + h^2 \cdot \frac{u''(x)}{2!} + \cdots + \frac{h^k}{k!} \cdot u^{(k)}(x) + \underbrace{\frac{h^{k+1}}{(k+1)!} u^{(k+1)}(\xi)}$$

- there are many ways that we can express the error, but we can express it as $\frac{h^{k+1}}{(k+1)!} u^{(k+1)}(\xi)$ for some $\xi \in (x, x+h)$ when $u \in C^{k+1}((a,b))$ and $(x, x+h) \subseteq (a,b)$.
- $u \in C^{k+1}((a,b))$ is a big assumption!

① Finite difference for the 1st derivative.

Since $u'(x) := \lim_{\varepsilon \to 0} \frac{u(x+\varepsilon)-u(x)}{\varepsilon}$, we say, if h is small, then $u'(x) \simeq \frac{u(x+h)-u(x)}{h} =: D_+ u(x)$, the one-sided FD.

But why not $u'(x) \simeq \frac{u(x)-u(x-h)}{h} =: D_- u(x)$?

Hence, the less biased centered FD is given by

$$D_0 u(x) := \frac{u(x+h)-u(x-h)}{2h}$$

We say that D_+, D_-, and D_0 all approximate $u'(x)$, but what is the respective error? Well, OTOH

$$u(x+h) = u(x) + h \cdot u'(x) + \frac{h^2}{2!} u''(\xi)$$

$$\Rightarrow \frac{u(x+h)-u(x)}{h} - u'(x) = \frac{h}{2} u''(\xi).$$

Thus, error $= \mathcal{O}(h)$ (ie. $|\text{error}| \leq C \cdot h$)

In this case, we even know to take $C \geq \left|\frac{u''(\xi)}{2}\right|$ and can simply take $C = \sup_y \frac{1}{2} \cdot |u''(y)|$ ~~for all~~ $y \in (x, x+h)$.

Now, OTOH,

$$u(x-h) = u(x) - h \cdot u'(x) + \frac{h^2}{2!} \cdot u''(\xi)$$

$$\Rightarrow D_- u(x) - u'(x) = \frac{-h}{2} u''(\xi)$$

$\therefore$ D_+u and D_-u are 1st order FD approx of u'.

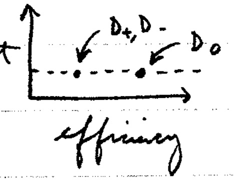
Observe, however, that

$$\begin{cases} u(x+h) = u(x) + h \cdot u'(x) + \frac{h^2}{2!} u''(x) + \frac{h^3}{3!} \cdot u'''(\xi_1) \\ u(x-h) = u(x) - h \cdot u'(x) + \frac{h^2}{2!} u''(x) - \frac{h^3}{3!} \cdot u'''(\xi_2) \end{cases}$$

$$\Rightarrow \frac{u(x+h)-u(x-h)}{2h} - u'(x) = \frac{h^2}{2 \cdot 3!} \cdot \left[u'''(\xi_1) + u'''(\xi_2)\right]$$

$\therefore$ $D_0 u$ is a 2nd order FD approximation of u', as $|D_0 u(x) - u'(x)| \leq \frac{1}{6} \sup_y |u'''(y)|$

Interestingly enough,



D_+, D_-, and D_0 have the same cost.

We didn't get anything for free, though!
In order to apply D_0, we assume/need regularity of $u \in C^3$!

② Second order and higher order derivative FD approx.

Centered finite difference

$$D_2 u(x) = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2}$$

(computational time ~ 3 evaluations ⇒ 2 evals for D_+, etc)

We can show that

$$D_2 u(x) - u''(x) = \frac{h^2}{12} u^{(4)}(x) + O(h^4)$$

∴ $D_2 u$ is of 2nd order accuracy.

Well, $D_+ D_- u$ is given by

$$D_+\left(\frac{u(x) - u(x-h)}{h}\right) = \frac{1}{h}\left[D_+ u(x) - D_+ u(x-h)\right].$$

$$= \frac{1}{h}\left[\frac{u(x+h) - u(x)}{h} - \frac{u((x-h)+h) - u(x-h)}{h}\right]$$

$$= \frac{u(x+h) - 2u(x) + u(x-h)}{h} = D_2 u.$$

It is too naive to think

(1st order FD approx) • (1st order FD approx)

will be 2nd order.

For example, what is the order of $D_+ D_2 u$?
We "expect" 3rd order, but we only get 1st order of u".

CAAM 452 1/16

There is a general method to express the n^{th} order finite diff approximation of a function. Suppose we are given n points, $x_1, \dots, x_n$, and a function's values at those points, $u(x_1), \dots, u(x_n)$; how can we approximate $u^{(k)}(x)$ (for a given value of x)? Assume that $n \geq k+1$ and $\max_{1 \le i \le n} |x - x_i| \le Ch$ for some $C \geq 0$. Can we find $c_1, \dots, c_n$ so that $u^{(k)}(x) \approx c_1 \cdot u(x_1) + \dots + c_n \cdot u(x_n)$?

ex/ $u'(x) \approx c_1 \cdot u(x_1) + c_2 \cdot u(x_2) + c_3 \cdot u(x_3)$. $\vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

$x_1 = \cancel{x}$, $x_2 = x - h$, $x_3 = x - 2h$. $\downarrow$

We'd like to obtain a linear system $A\vec{c} = \vec{b}$.

Well, $u(x-h) = u(x) - h \cdot u'(x) + \frac{h^2}{2} u''(x) - \frac{h}{6} u'''(x) + \cdots$

and $u(x-2h) = u(x) - 2h \cdot u'(x) + 2 \cdot h^2 u''(x) - \frac{8h^3}{6} \cdot u'''(x) + \cdots$

$$\Rightarrow c_1 \cdot u(x) + c_2 \cdot u(x-h) + c_3 \cdot u(x-2h) = \underbrace{(c_1 + c_2 + c_3) \cdot u(x) - h \cdot (c_2 + 2c_3) \cdot u'(x) + h^2 \left(\tfrac{1}{2} c_2 + 2c_3\right) u''(x) - \tfrac{1}{6}(c_2 + 8c_3) h^3 u'''(x) + O(h^4)}_{\text{want} \approx u'(x)}$$

$$\therefore \begin{cases} c_1 + c_2 + c_3 = 0 \\ -h(c_2 + 2c_3) = 1 \\ h^2\left(\frac{1}{2}c_2 + 2c_3\right) = 0 \end{cases} \rightsquigarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & -h & -2h \\ 0 & \frac{h^2}{2} & 2h^2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$\therefore c_1 = \frac{3}{2h}$, $c_2 = \frac{-2}{h}$, $c_3 = \frac{1}{2h}$

Such c_1, c_2, c_3 would yield the RHS $= u'(x) - \frac{1}{6}(c_2 + 8c_3) h^3 u'''(x) + O(h^4)$

where $\frac{1}{2}(c_2 + 8c_3) h^3 u'''(x) + O(h^4) = O(h^2)$.

This yields a 2^{nd} order approx.

The method of undetermined coefficients

Where $x_i = x - ih$, we expand:

$$u(x_i) = u(x) + (x_i - x)\cdot u'(x) + \cdots + \frac{1}{j!}(x_i - x)^j \cdot u^{(j)}(x) + \cdots$$

$$\Rightarrow \sum_{i=1}^{n} c_i u(x_i) = \left(\sum_{i=1}^{n} c_i\right) u(x) + u'(x)\cdot\left(\sum_{i=1}^{n} c_i (x_i - x)\right) + \cdots$$

$$+ \cdots + u^{(j)}(x)\cdot \frac{1}{j!}\left(\sum_{i=1}^{n} c_i (x_i - x)^j\right) + \cdots$$

We want this to equal $u^{(k)}(x) + O(h^p)$ for as great a p as possible. We expect the method to be of order $p \geq n-k$.

$$\therefore \frac{1}{j!}\left(\sum_{i=1}^{n} c_i (x_i - x)^j\right) = \begin{cases} 0 & \text{if } j = k \\ 1 & \text{if } j \neq k \end{cases} \quad \text{for } \underset{\text{to get a square } A!}{0 \leq j \leq n-1}$$

Set $\xi_i = x_i - x$. Then:

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \xi_1 & \xi_2 & \cdots & \xi_n \\ \xi_1^2 & \xi_2^2 & \cdots & \xi_n^2 \\ \vdots & \vdots & \ddots & \\ \xi_1^{n-1} & \xi_2^{n-1} & \cdots & \xi_n^{n-1} \end{bmatrix} \longleftarrow \text{this is the Vandermonde matrix!}$$

ex/ $u'(h) \approx \dfrac{-u(0) + u(h)}{h}$ when $\begin{matrix} x_1 = 0 \\ x_2 = 0 + h \end{matrix}$ $x = h$

i.e. $c_1 = -1$, $c_2 = 1$

ex/ $u'(h) \approx \dfrac{-u(0) + u(2h)}{2h}$ when $\begin{matrix} x_1 = 0 \\ x_2 = 0 + h \\ x_3 = 0 + 2h \end{matrix}$ $x = h$

ex/ $u'(0) \approx \frac{3}{2h} u(0) + \frac{2}{h} u(h) - \frac{1}{2h} u(2h)$ when $\begin{matrix} x_1 = 0 \\ x_2 = h \\ x_3 = 2h \end{matrix}$ $x = 0$

costs more than, even though both are $O(h^2)$

Note! $D_0 u(x) = \dfrac{u(x-h) - u(x+h)}{2h}$

But our domain might not include $-h$! So, sometimes we need the above.

Let's use this to solve PDE's!

II Two-point Boundary value problem (Dirichlet)

$$\text{Consider} \begin{cases} -u''(x) = f(x) & 0 < x < 1 \\ u(0) = \alpha \\ u(1) = \beta \end{cases}$$

Let $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ and take $h = \frac{1}{N+1}$.
(So, $x_j = j \cdot h$.) This yields a uniform partition.
Note that this demands, in the interior,

$$-u''(x_j) = f(x_j) \quad \text{for } 1 \le j \le N$$

FDA of $u''(x)$ then says

$$u''(x_j) \approx \frac{u(x_j + h) - 2u(x_j) + u(x_j - h)}{h^2}$$

We'll adopt the convention: $\vec{U}$ will denote the numerical approx.

$$\therefore \quad \frac{-U_{j+1}^h + 2U_j^h - U_{j-1}^h}{h^2} = f(x_j) \qquad 1 \le j \le N$$

$$U_0^h = \alpha$$
$$U_{N+1}^h = \beta$$

where U_{j+1}^h ← denotes dep on h
← denotes approx val of $u(x_{j+1} = (j+1) \cdot h)$.

$$A^h \vec{U}^h = \vec{b}^h \text{ is then}$$

$$A^h = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & - & - & - & 0 \\ -1 & 2 & -1 & - & - & - & 0 \\ 0 & -1 & 2 & -1 & - & & 0 \\ \vdots & \vdots & -1 & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & 2 & -1 & 0 \\ \vdots & \vdots & \vdots & & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \qquad \vec{b}^h = \begin{bmatrix} \left(f(x_0) + \frac{\alpha}{h^2}\right) \\ f(x_1) \\ \vdots \\ f(x_N) \\ \left(f(x_{N+1}) + \frac{\beta}{h^2}\right) \end{bmatrix}$$

rather, record $\frac{1}{h^2}$ ← in A

CAAM 452 1/21

Recall: FD in 1-D has a grid domain (for ex) with nodes x_i, and the PDE is evaluated at node x_i; we replaced derivatives by FD approx.

ex// $\begin{cases} -u'' = f \\ u(0) = \alpha \\ u(1) = \beta \end{cases}$ $\rightarrow$ U_i^h for $0 \leq i \leq N+1$

↑ use the centered FD approx, yielding

$$\begin{cases} -\dfrac{U_{i+1}^h - 2U_i^h + U_{i-1}^h}{h^2} = f(x_i) \quad \text{for } 1 \leq i \leq N \\ U_0^h = u(0) = \alpha \\ U_{N+1}^h = u(1) = \beta \end{cases}$$

This produced the matrix eqn

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \vec{U}^h = \vec{b} \quad \text{where } \vec{b} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{bmatrix} + \begin{bmatrix} \alpha/h^2 \\ 0 \\ \vdots \\ 0 \\ \beta/h^2 \end{bmatrix}$$

This lets us approximate $u(x_1), \ldots, u(x_N)$.
How do we interpolate b/w these points? Also, how can we guarantee convergence of the approx soln?

Recall: error is $e_i^h := u(x_i) - \overline{U}_i^h$

We say that the FD method converges if $e_i^h \to 0$ as $h \to 0$.
Since $h = \frac{1}{N+1}$, we'd like $e_i^h \to 0$ as we take more pts.

defn <u>local truncation error at node x_i</u>, denoted τ_i^h, is given by, in this example,

$$\tau_i^h := -\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i+1})}{h^2} - f(x_i) .$$

By utilizing Taylor exp of $u(x_i+1)$, $u(x_{i+1})$, we get

$$\tau_i^h = -u''(x_i) - \frac{h^2}{12} u^{(4)}(\xi_i) - f(x_i),$$

where $\xi \in [x_i, x_{i+1})$.

Observe that u satisfies the ODE, so $\tau_i^h = \frac{-h^2}{12} u^{(4)}(\xi)$.

Where $\tau : \begin{pmatrix} \tau_1^h \\ \tau_2^h \\ \vdots \\ \tau_N^h \end{pmatrix}$, we have $\|\tau\|_\infty \leq \frac{h^2}{12} \cdot \max_{x \in [0,1]} |u^{(4)}(x)|$

defn Since $\lim_{h \to 0} \|\tau\|_\infty = 0$, we call ~~the~~ FD method <u>consistent</u>.

Note: $$① \frac{U_{i+1} - 2U + U_{i-1}}{h^2} \quad \text{and} \quad ② \frac{u(x_{i+1}) - 2u(x_i) - u(x_{i-1})}{h^2}$$

$$\|$$

$$f(x_i) \qquad\qquad f(x_i) + \tau_i$$

$e_i = u(x_i) - U_i$, so

$$② - ① : \frac{-1}{h^2}(e_{i+1} - 2e_i + e_{i-1}) = \tau_i$$

$$\therefore \frac{1}{h^2} \cdot \begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} = A \text{ (invertible!)}$$

and $A \cdot \vec{e} = \vec{\tau}$.

Now, $\vec{e} = A^{-1}\vec{\tau} \Rightarrow \|\vec{e}\| = \|A^{-1}\| \cdot \|\tau\|$ w/ $\|B\| := \sup_{\vec{y} \neq \vec{0}} \frac{\|By\|}{\|y\|}$.

defn If the FD method given by $A^h \cdot u^h = b^h$ is s.t.

① A^h is invertible for small h <u>and</u>

② $\exists C$ (indep of h) st $\|(A^h)^{-1}\| \leq C$,

then we say it is <u>stable</u>.

If the FD method is stable <u>and</u> consistent, then the error satisfies

$\|\vec{e}\| \to 0$ as $h \to 0$ and the FD method is convergent!

Here, $\|\tau\| = O(h^2)$. Thus, if we can show that the method is stable, then the order of convergence is 2, ~~as~~ $\|\vec{e}\| = O(h^2)$.

To prove consistency, we use Taylor series (easy).
To prove stability, it is harder.

If A is symmetric, then $\|A\|_2 = \max_{\lambda \in \sigma(A)} |\lambda| = \max_{\mu \in \sigma(A)} |\mu| = \frac{1}{\min_{\mu \in \sigma(A)} |\mu|}$
where $\lambda \in \sigma(A) \Leftrightarrow \exists \vec{x} \neq \vec{0}$ s.t. $A\vec{x} = \lambda \vec{x}$.
and $\|A\|_2 := \sup_{\vec{x} \neq \vec{0}} \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2}$, $\|x\|_2 = \sqrt{\sum x_i^2}$.

ex/ $A^h = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}$ & $\sigma(A) \ni \lambda_k = \frac{2}{h^2}\left(1 - \cos(k\pi h)\right)$

with $k = 1, \ldots, N$ and $k\pi h = \frac{k\pi}{N+1}$ for $1 \leq k \leq N$,
we have $\lambda_1 = \frac{2}{h^2}\left(1 - \cos(\pi h)\right) \simeq \frac{2}{h^2}\left(\frac{(\pi h)^2}{2} + O(h^4)\right)$
$\therefore \lambda_1 = \pi + O(h^2)$. Can uniformly bd λ_k by some C.
$\therefore \|A^{-1}\|_2 \leq \frac{1}{\pi^2}$.

To depict rate of convergence, plot: (this is heuristic!)

h	$\|error\|$	rate
$1/10$	A_1	
$1/20$	A_2	$\log(A_1/A_2)/\log(2)$
$1/40$	A_3	$\log(A_2/A_3)/\log(2)$
$1/80$	A_4	$\vdots$
$\vdots$		

why? $\|e_h\| \approx Ch^p$ ~~to~~ go from $h=\frac{1}{10}$ to $\frac{h}{2}=\frac{1}{20}$,

$$\Rightarrow \|e_{\frac{h}{2}}\| \approx C\cdot\frac{h^p}{2^p}$$

$$\therefore \log\left(\frac{\|e_h\|}{\|e_{\frac{h}{2}}\|}\right) \approx \log\left(\frac{Ch^p}{C\frac{h^p}{2^p}}\right) = \log(2^p)$$

$$\therefore p \approx \log\left(\frac{\|e_h\|}{\|e_{h/2}\|}\right) \Big/ \log(2)$$

(Returning to $u''=f$ example)

We had $\|e^h\| \le C\cdot h^2 \cdot \|u^{(4)}\|_\infty$.

So, it's good to test: if we know the exact soln to have $u^{(4)}=0$, then the error needs to be $\|e^h\|=0$.

CAAM 452 1/23

Recall: $-u'' = f$

$u(0) = \alpha \quad \rightsquigarrow \quad A^h U^h = b^h$ where $A^h = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & -1 & \\ & -1 & \ddots & \ddots \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{bmatrix}$

$u(1) = \beta$

But we can also use $\tilde{A}^h \cdot \tilde{U}^h = \tilde{b}^h$ where

$$\tilde{A}^h = \frac{1}{h^2}\begin{bmatrix} h^2 & 0 & 0 & \cdots & & 0 \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & -1 & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 & -1 \\ & & \cdots & 0 & 0 & h^2 \end{bmatrix} \quad \text{and} \quad \tilde{b}^h = \begin{bmatrix} \alpha \\ f(x_1) \\ \vdots \\ f(x_n) \\ \beta \end{bmatrix}$$

2 Neumann problem

$-u'' = f$ in $(0,1)$

$u'(0) = \sigma \leftarrow$ Neumann BC

$u(1) = \beta \leftarrow$ Dirichlet BC

Since we have a Neumann BC at $x=0$, we can't use a centered FDA for $u'(x)$; instead, we'll use a right sided FDA:

$$u'(0) \simeq \frac{u(x_1) - u(x_0)}{h} \rightsquigarrow \frac{U_1 - U_0}{h} = \sigma$$

In the interior, we still have

$$-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} = f(x_i) \quad \text{for } 1 \leq i \leq N.$$

The other boundary condition yields $U_{N+1} = \beta$.

$$\therefore A^h = \frac{1}{h^2}\begin{bmatrix} -h & h & 0 & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & -1 & 2 & -1 \\ & & & & h^2 \end{bmatrix} \quad \text{and} \quad b^h = \begin{bmatrix} \sigma \\ f(x_1) \\ \vdots \\ f'(x_N) \\ \beta \end{bmatrix}$$

This method will see error = $O(h)$, since the Neumann boundary condition is $O(h)$ even though the rest are $O(h^2)$. The Dirichlet condition imposes $u(1) = U_{N+1} = 0$.

How can we get this to be $O(h^2)$? Using Taylor's series,

$$u(x_1) = u(x_0) + h \cdot u'(x_0) + \frac{h^2}{2!} u''(x_0) + O(h^3)$$

We don't know $u''(x_0)$, but to try to get $O(h^2)$ convergence, we can assume $-u''(x_0) = f(x_0)$, and so

$$h \cdot u'(x_0) = u(x_1) - u(x_0) + \frac{h^2}{2!} f(x_0) + O(h^3).$$

$$\Rightarrow u'(x_0) \approx \frac{u(x_1) - u(x_0)}{h} + \frac{h}{2} f(x_0).$$

$$\therefore \frac{U_1 - U_0}{h} + \underbrace{\frac{h}{2} f(x_0)}_{\text{we have this perturbation now!}} = \sigma \text{ is the imposed condition.}$$

A^h is unchanged, but now,

$$b^h = \begin{bmatrix} \sigma - \frac{h}{2} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_N) \\ \beta \end{bmatrix}$$

Now, error = $O(h^2)$. The drawback is the assumption that $-u''(x_0) = f(x_0)$!

Alternatively, we can write down a 2nd order approx of $u'(0)$:

$$u'(x_0) \approx \frac{-3}{2h} u(x_0) + \frac{2}{h} u(x_1) - \frac{1}{h} u(x_2) + O(h^2)$$

by taking three points.

$$\therefore -\frac{3}{2h} U_0 + \frac{2}{h} U_1 - \frac{1}{2h} U_2 = \sigma.$$

Therefore we get

$$A=\frac{1}{h^2}\begin{bmatrix} -\frac{3}{2}h & 2h & -\frac{h}{2} & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & & h^2 \end{bmatrix}.$$

The error is $O(h^2)$.

3 Pure Neumann Boundary Conditions

$$-u''=f$$

$$u'(0)=\sigma_0$$

$$u'(1)=\sigma_1$$

Well, $-\int_0^1 u''(x)\,dx = \int_0^1 f(x)\,dx$.

$$-(u'(1)-u'(0))$$

$\Rightarrow\ -\sigma_1+\sigma_0=\int_0^1 f$ is a-priori satisfied!

This is called the compatibility criterion!

If f is a solution, then $v+C$ is ~~[illegible]~~ a soln as well, for any constant C.

We need an additional condition on u : $\int_0^1 u = 0$.

4 General boundary ~~[illegible]~~ value problem.

$$-\left(k(x)\cdot u'(x)\right)' + d(x)\cdot u(x) = f(x)$$

$0<a<b$ and $u(a)=\alpha, u(b)=\beta$.

Product rule says: ← as before!

$$-(k\cdot u')' = -k'\cdot u' - k\cdot u''.$$

We might try to apply FDA to $k \cdot u'$, using

$$u'(x_i) \simeq \frac{u(x_{i+1}) - u(x_{i-1})}{2h} \text{ is } O(h^2).$$

But, this expression needs $u \in C^2((0,1))$!

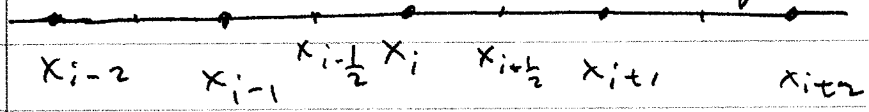
$k \cdot u'$ is the flux, or velocity, ...

Well, $k \cdot u'$ is continuous, but u' may be discontinuous

Denote $g := k \cdot u'$. The ODE becomes

$$-g'(x_i) + d(x_i)u(x_i) = f(x_i).$$

We use a trick: introduce a staggered mesh $\{x_{i+\frac{1}{2}}\}$



$$\Rightarrow g'(x_i) \simeq \frac{g(x_{i+\frac{1}{2}}) - g(x_{i-\frac{1}{2}})}{h}$$

where $g(x_{i+\frac{1}{2}}) = k(x_{i+\frac{1}{2}}) \cdot u'(x_{i+\frac{1}{2}})$.

Note that $u'(x_{i+\frac{1}{2}}) = \dfrac{u(x_i+1) - u(x_i)}{h}$

$$\therefore g'(x_i) = \frac{-1}{h^2}\left(k(x_{i+\frac{1}{2}}) \cdot (u(x_{i+1}) - u(x_i)) - k_{i-\frac{1}{2}}(u(x_i) - u(x_{i-1}))\right)$$

This method ensures that we have a tri-diagonal matrix!

The ODE becomes

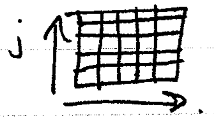
$$\frac{-1}{h^2} \cdot \left[k(x_{i+\frac{1}{2}}) \cdot (U_{i+1} - U_i) - k(x_{i-\frac{1}{2}}) \cdot (U_i - U_{i-1})\right]$$

$$+ d(x_i) \cdot U_i = f(x_i).$$

with error $O(h^2)$.

Elliptic problems in 2-D

$\Omega = (0,1)^2$



Where $x_i = x_0 + ih_x$ we'll assume $h_x = h_y$.
$y_j = y_0 + jh_y$.

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = g & \text{on } \partial\Omega \end{cases}$$ where $\Delta u = \partial_x^2 u + \partial_y^2 u$.

$$\Delta u(x_i,y_j) \approx -\frac{u(x_{i+1},y_j) - 2u(x_i,y_j) + u(x_{i-1},y_j)}{h^2} - \frac{u(x_i,y_{j+1}) - 2u(x_i,y_j) + u(x_i,y_{j-1})}{h^2}.$$

Write $U_{i,j} \approx u(x_i,y_j)$.

$$\therefore\ U_{i,j} = -\left(\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}\right) - \left(\frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2}\right)$$

$$= \frac{1}{h^2}\left(-U_{i+1,j} - U_{i,j+1} + 4U_{i,j} - U_{i-1,j} - U_{i,j-1}\right).$$

$\therefore\ U_{ij} \doteq f(x_i,y_j) \quad \forall\, 1 \le i \le N,\ \forall\, 1 \le j < N$

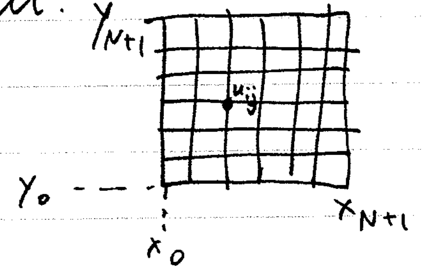
is our system, along with the boundary conditions.

$U_{0,j} = g(x_0, y_j)$

$U_{N+1,j} = g(x_{N+1}, y_j)$, etc.

CAAM 452 1/28

Recall:



$\Delta u = f$ in Ω

$u = g$ on $\partial\Omega$

$$\frac{1}{h^2}\left(-U_{i-1,j} - U_{i,j-1} + 4U_{ij} - U_{i+1,j} - U_{i,j+1}\right) = f(x_i, y_j) \quad \text{(interior pts)}$$

for $1 \le i \le N$, $1 \le j \le N$

$U_{i,\cancel{0}} = g(x_i, y_0)$ $\qquad u_{0,j} = g(x_0, y_j)$

$U_{i,\cancel{N+1}} = g(x_i, y_{N+1})$ $\qquad U_{N+1,j} = g(x_{N+1}, y_j)$

Matrix U is N^2 entries long,

so A is now $N^2 \times N^2$ if we don't include bdy.

The structure of the matrix depends on the ordering of the nodes; we adopt the "natural row-wise ordering".

So, we'd have $U_{ij} =: \vec{U}_{i+(j-1)\cdot N}$ and also

$$A = \frac{1}{h^2}\begin{bmatrix} T & -I & & & 0 \\ -I & T & -I & & \\ & -I & T & \ddots & \\ & & \ddots & T & -I \\ 0 & & & -I & T \end{bmatrix}_{(N\times N)\times(N\times N)}$$

is block-tridiagonal

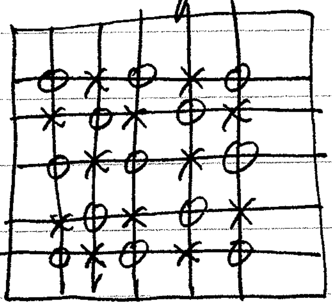
where $T = \begin{bmatrix} 4 & -1 & & 0 \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 4 \end{bmatrix}_{N\times N}$ and $I = I_{N\times N}$.

Note that the Dirichlet boundary condition makes b look like

$$b = \begin{pmatrix} \left.\begin{matrix} f(x_1,y_1)\\ f(x_2,y_1)\\ \vdots \\ f(x_N,y_1)\end{matrix}\right] N \\ \left.\begin{matrix} f(X_2,y_1)\\ \vdots \\ \vdots \end{matrix}\right] N \\ \vdots \\ f(x_N,y_N) \big] N \end{pmatrix} + \frac{1}{h^2}\begin{pmatrix} \left.\begin{matrix} g(1,0)\\ g(2,0)\\ g(N,0)\end{matrix}\right] N \\ | \\ 0 \\ | \end{pmatrix} + \frac{1}{h^2}\begin{pmatrix} \\ \\ \\ \end{pmatrix} + \frac{1}{h^2}\begin{pmatrix} \\ \\ \end{pmatrix} + \frac{1}{h^2}\begin{pmatrix} \\ \\ \end{pmatrix}$$

bottom bdy top bdy left bdy right bdy

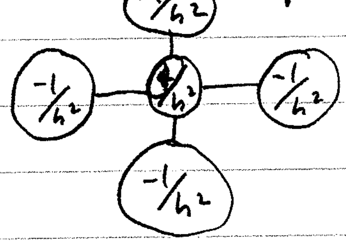
Alternatively, we can apply a "checkerboard ordering":



Where the X's & O's have the natural row-wise ordering.

Then, $A = \begin{bmatrix} D & H \\ H^T & D \end{bmatrix}$ where $D = \frac{4}{h^2} I_{\frac{N^2}{2} \times \frac{N^2}{2}}$

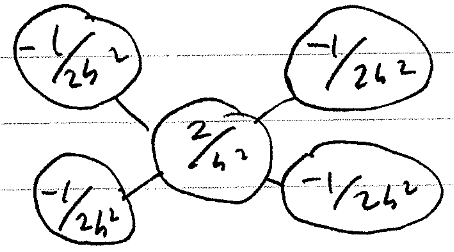
The "stencil" for this FDA scheme is given by



error = $O(h^2)$

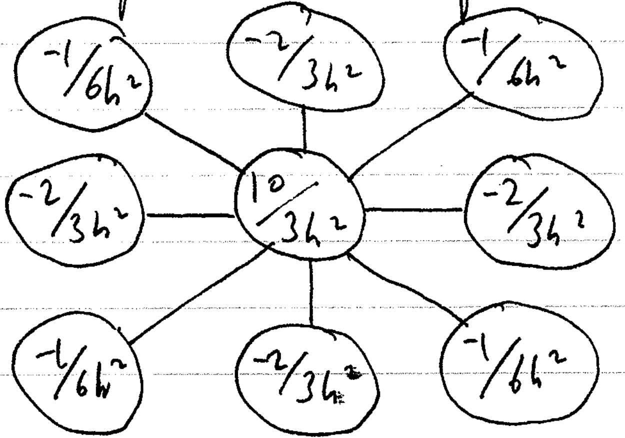
5 point stencils

An alternative could be



error = $O(h^2)$

A 9-point stencil for Δu is



this can be a 4th order method!

A becomes less sparse with a 9-point stencil, but we hope to get more accuracy.

Local truncation error for the first 5-pt stencil is

$$\tau_{ij} = \frac{1}{h^2}\left(-u(x_{i-1},y_j) - u(x_i,y_{j-1}) + 4u(x_i,y_j) - u(x_{i+1},y_j) - u(x_i,y_{j+1})\right) - f(x_i,y_j)$$

(This is just analyzing the Taylor series!)

Since $u(x_{i-1},y_j) = u(x_i - h, y_j)$

$$= u(x_i,y_j) - h\cdot\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} - \cdots, \text{ etc.}$$

Hence, $\tau_{ij} = \frac{-1}{12}h^2\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right) + O(h^4)$ is $O(h^2)$.

Local truncation error for the 9-pt stencil is

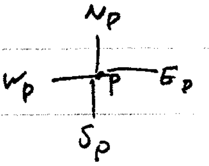
$$\tau_{ij} = \frac{-h^2}{12}\underbrace{\left(\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}\right)}_{\begin{array}{c}\| \\ (\partial_x^2 + \partial_y^2)\circ(\partial_x^2 + \partial_y^2)u\end{array}} + O(h^4) \text{ is } O(h^2).$$

So, if $f = 0$, then τ_{ij} is in fact $O(h^4)$!

The error is given by $e_{ij} = u(x_i, y_j) - u_{i,j}$.
With the same ordering, we have $A\vec{e} = \vec{\tau}$,
so $\vec{e} = A^{-1}\vec{\tau}$. Thus, if $\|A\| < \infty$, then $\|\vec{e}\| \searrow 0$ as $\|\vec{\tau}\| \searrow 0$.

To simplify implementation of the 5 pt FDA Δu scheme, consider $S_h :=$ set of nodes (interior & bdry).
For $P, Q \in S_h$, we can define

$$B(P,Q) = \begin{cases} 4/h^2 & \text{if } Q = P \\ -1/h^2 & \text{if } Q \in \{N_P, E_P, W_P, S_P\} \text{ where} \\ 0 & \text{else.} \end{cases}$$



Then the FDA is written as

$$\begin{cases} \sum_{Q \in S_h} B(P,Q)\,\hat{u}(Q) = f(P) & \text{for all } P \in S_h \setminus \partial\Omega. \\ \hat{u}(P) = g(P) & \text{for all } P \in \partial\Omega. \end{cases}$$

This interpretation lets us realize an operator L_h defined via $L_h \hat{V}(P) := \sum_{Q \in S_h} B(P,Q)\,\hat{V}(Q)$ for all $P \in S_h \setminus \partial\Omega$, for any $\hat{V}(P)$ a function on S_h.
Equivalently,

$$\begin{cases} L_h \hat{u}(P) = f(P) & \text{for all } P \in S_h \setminus \partial\Omega \\ \text{and } \hat{u}(P) = g(P) & \text{for all } P \in \partial\Omega. \end{cases}$$

Some results from PDE theory will help us to prove convergence.

Thm Maximum Principle

Where $L(u) := -\Delta u$ for all $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$.

If $Lu \le 0$, then $\max_{(x,y)\in\bar{\Omega}} u(x,y) = \max_{(x,y)\in\partial\Omega} u(x,y)$.

If $Lu \ge 0$, then $\min_{(x,y)\in\bar{\Omega}} u(x,y) = \min_{(x,y)\in\partial\Omega} u(x,y)$.

Theorem Assume Ω is contained in a strip $\{(x,y) \mid a \le x < b\}$.

Then we have $|e(P)| \le \frac{1}{2}\max(a^2, b^2) \cdot \max_{Q\in S_h\setminus\partial\Omega} |\tau(Q)|$.

error $e(P) = e_{ij}$ if $P = (x_i, y_j)$.

This theorem will give us convergence!

Proof Define $\phi(P) = \frac{1}{2}(\max(a^2, b^2) - x_P^2) \cdot \max_{Q\in S_h\setminus\partial\Omega} |\tau(Q)|$

Note that $\phi(P) \ge 0$, as $\Omega \subseteq (a,b) \times \mathbb{R}$.

Compute:

$$
\begin{aligned}
L_h e(P) &= \sum_Q B(P,Q)\, e(Q) \\
&= \sum_Q B(P,Q)\,[u(Q) - \bar{u}(Q)] \\
&= \sum_Q B(P,Q)\,u(Q) - \underbrace{\sum_Q B(P,Q)\,\bar{u}(Q)}_{f''(P)} \\
&= \Big(\sum_Q B(P,Q)\,u(Q)\Big) - f(P) \\
&= \tau(P) \le \max_Q |\tau(Q)| \le L_h(\phi(P))
\end{aligned}
$$

Lemma: we will show this.

$\therefore L_h(e(P)) \le L_h(\phi(P))$

$\Rightarrow L_h((e-\phi)(P)) \le 0$ for all $P \in S_h \setminus \partial\Omega$. Does the Maximum principle applys, even though we are only interested in the function at S_h? Well, $P \in \partial\Omega \Rightarrow e(P) = 0 \le \phi(P)$, i.e. $P \in \partial\Omega \Rightarrow (e-\phi)(P) \le 0$. This is ripe for something like Max. Principle. So, we need to develop a discrete MP.

CAAM 452 1/29

Recall: we were analyzing the convergence of the FD discrete operator

$$L_h \widehat{U}(P) = \sum_{Q \in S} B(P,Q) \widehat{U}(Q)$$

Defining $\widehat{W} = e - \phi$, we showed that

Lemma: MP!

$L_h \widehat{W}(P) \leq 0 \quad \forall P$ in the interior

$\widehat{W}(P) \leq 0 \ \forall P$ on the boundary

$\Rightarrow \widehat{W}(P) \leq 0$ for all P (i.e. in the interior as well)

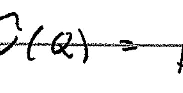
$$\Rightarrow e(P) \leq \phi(P) = \frac{1}{2}\left(\max(a^2,b^2) - x_P^2\right) \cdot \max_Q |\tau(Q)|$$
$$\leq \frac{1}{2}\max(a^2,b^2) \max_Q |\tau(Q)|.$$

We can also show $-e(P) \leq \frac{1}{2} \cdot \max(a^2,b^2) \cdot \max_Q |\tau(Q)|$.

$$\therefore |e(p)| \leq \frac{1}{2}\max(a^2,b^2) \cdot \max_Q |\tau(Q)|.$$

Proof (of Lemma)

Let $\widehat{W}(R) := \max_{P \in S_h} \widehat{W}(P)$, i.e. R attains a max of $\widehat{W}$. (Note: the mesh is finite. If $R \in \partial\Omega$, we are done, so suppose R is interior). Consider a path from R to the boundary via mesh points. We will show $\widehat{W}$ is constant (and ≤ 0) on that path. Denote the path $R, P_1, P_2, \ldots, P_h \in \partial\Omega$. Well, $\widehat{W}(R)$ is a max; suppose for a ~~X~~ that $\widehat{W}(R) \neq \widehat{W}(P_1)$. Then $\widehat{W}(P_1) \lneq \widehat{W}(R)$. By assumption, $\widehat{W}(R) \leq 0$, i.e. $\sum_{Q \in S_h} B(R,Q)\widehat{W}(Q) \leq 0$.

(Note that the  stencil sees $\sum_{Q \in S_h} B(P,Q) = 0$ for all P interior)

$$\sum_{Q \in S_h} B(R,Q)\widehat{W}(Q) = \underset{\searrow 0 = \frac{4}{h^2}}{B(R,R)} + \sum_{Q \neq R} B(R,Q)\widehat{W}(Q) \leq 0$$

$$\therefore \widehat{W}(R) \leq -\sum_{Q \neq R}\left(\frac{B(R,Q)}{B(R,R)}\widehat{W}(Q)\right) \leq -\sum_{Q \neq R}\frac{B(R,Q)}{B(R,R)}\widehat{W}(R).$$

since $\hat{W}(R)$ is maximal.

But, we have a strict inequality since $\hat{W}(P_1) \lneq \hat{W}(R)$!

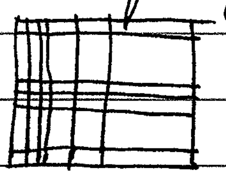
Note that $-\sum_{Q \neq R} \frac{B(R,Q)}{B(R,R)} = 1$. Hence, $\hat{W}(R) \lneq \hat{W}(R)$. $\times\!\!\times$

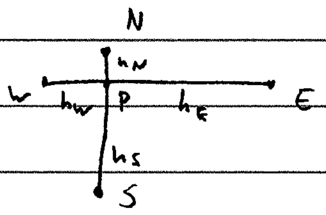
Repeat until we get to the boundary. $\square$

This MP holds so long as $\sum_Q B(P,Q) = 0$ for all P in the interior, more generally.

4 Generalizations of this FDA.

a) Non-uniform grid





$$\Delta u(P) \simeq \frac{2}{h_W h_E (h_E + h_W)} \cdot \left[h_W u(E) + h_E u(W) \right]$$
$$+ \frac{2}{h_S h_N (h_S + h_N)} \cdot \left[h_S u(E) + h_N u(W) \right]$$
$$- 2\left(\frac{1}{h_W h_E} + \frac{1}{h_S h_N} \right) \cdot u(P).$$

b) PDE w/ coefficients

$-\nabla \cdot K \nabla u$ w/ $K(x,y)$ a scalar

$$-\frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right)+c(x)\cdot u(x) = f(x) \text{ in } (0,1)$$

$$\left.\begin{aligned} u(0)&=0\\ u(1)&=0\end{aligned}\right]\text{ homogeneous Dirichlet boundary condition.}$$

The FEM (finite element method) soln is a cts piecewise polynomial of degree p. Let V_h = set of cts piecewise polynomials of degree p that vanish at the boundary.

Step 1) multiply ODE by $v \in V_h$ and integrate over domain

$$\int_0^1 \left(-(ku')'v + cuv\right)dx = \int_0^1 fv$$

Step 2) IBP where it makes sense

$$\int_0^1 ku'\cdot v' - \cancelto{0}{ku'v\Big|_0^1} + \int_0^1 cuv\,dx = \int_0^1 fv\,dx$$

$$\therefore \int_0^1 ku'\cdot v'$$

FEM: Want to find $u_h \in V_h$ st $\forall v_h \in V_h$, we have

$$\int_0^1 (k u_h' v_h' + c u_h v_h)\,dx = \int_0^1 f v_h\,dx.$$

How do we obtain a linear system from this?

Define $a_h : V_h \times V_h \to \mathbb{R}$

$$(u,v) \mapsto a_h(u,v) := \int_0^1 (k u'v' + cuv)$$

a bilinear form and

$$l : V_h \to \mathbb{R}$$

$$v \mapsto l(v) = \int_0^1 fv.$$

Hence, FEM amounts to finding $u_h \in V_h$ s.t. $\forall v_h \in V_h$, we have $a(\underset{\substack{\uparrow \\ (\text{trial func})}}{u_h}, \underset{\substack{\uparrow \\ (\text{test func})}}{v_h}) = l(v_h)$ ← variational form.