Math 228B Tues 1/16/07 Week 1

Lec 1

Numerical solutions of PDE

what is a PDE? abstractly it's an equation of the form

$$F(u, Du, D^2u, \dots) = 0 \qquad Du = \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \vdots \\ \frac{\partial u}{\partial x_n} \end{pmatrix} = \text{gradient}$$

$$D^2u = \begin{pmatrix} \frac{\partial^2 u}{\partial x_1^2} & \cdots & \frac{\partial^2 u}{\partial x_1 \partial x_n} \\ \frac{\partial^2 u}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 u}{\partial x_n^2} \end{pmatrix} = \left(\frac{\partial^2 u}{\partial x_i \partial x_j} \right)_{i,j=1..n} = \text{Hessian matrix}$$

unlike ODE's, there is no general theory of PDE.

Each equation has its own special properties, and the behavior of solutions varies wildly from one PDE to the next.

2nd order, scalar, constant coefficients:

$$a u_{xx} + 2b u_{xy} + c u_{yy} + d u_x + e u_y + f u = g$$

a, b, c, d, e, f constants

$g(x,y)$ function

This equation may be written

$$P(\partial_x, \partial_y) u = g \qquad \text{or} \qquad Lu = g$$

where $P(\xi, \eta) = a\xi^2 + 2b\xi\eta + c\eta^2 + d\xi + e\eta + f$

is the symbol of the differential operator L.

polynomial with the coefficients of the operator the behavior of solutions of $Lu = g$ is largely determined by the algebraic properties of the polynomial $P(\xi, \eta)$, in fact by the discriminant:

$$b^2 - ac < 0 \qquad \text{hyperbolic}$$
$$b^2 - ac = 0 \qquad \text{parabolic}$$
$$b^2 - ac > 0 \qquad \text{elliptic}$$

note that only the principal terms (those of highest order) matter in this classification

what's special about $b^2 - ac$?

write $P(\xi, \eta) = (\xi \ \eta) \underbrace{\begin{pmatrix} a & b \\ b & c \end{pmatrix}}_{A} \begin{pmatrix} \xi \\ \eta \end{pmatrix} + (d \ e) \begin{pmatrix} \xi \\ \eta \end{pmatrix} + f$

the eigenvalues of A are:

$$\det \begin{pmatrix} a - \lambda & b \\ b & c - \lambda \end{pmatrix} = (a - \lambda)(c - \lambda) - b^2$$
$$= \lambda^2 - (a + c)\lambda + ac - b^2 = 0$$

$$\lambda = \frac{a + c \pm \sqrt{(a+c)^2 + 4(b^2 - ac)}}{2}$$ ← sign of discriminant determines whether evals have same or opposite sign

$$= \frac{a + c \pm \sqrt{(a-c)^2 + 4b^2}}{2}$$ ← they're always real (since A is symmetric)

when $b^2 - ac = 0$, at least one eigenvalue is zero (parabolic case)

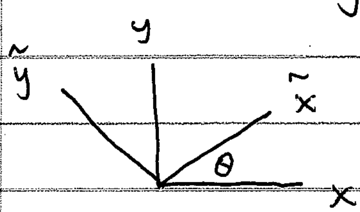
Change of variables

since A is symmetric, we can diagonalize it:

$$A = U \Lambda U^{-1}, \quad \Lambda = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}, \quad U \text{ orthogonal } (U^{-1} = U^T)$$

$$U = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}, \quad \begin{matrix} c = \cos\theta \\ s = \sin\theta \end{matrix} \quad \leftarrow \text{different } c$$

now let's try rotating the coordinate system by θ:



$$\left.\begin{matrix} \tilde{x} = cx + sy \\ \tilde{y} = -sx + cy \end{matrix}\right\} = U^{-1}\begin{pmatrix} x \\ y \end{pmatrix}$$

By the chain rule, we have

$$\left.\begin{matrix} \dfrac{\partial}{\partial x} = \dfrac{\partial \tilde{x}}{\partial x}\dfrac{\partial}{\partial \tilde{x}} + \dfrac{\partial \tilde{y}}{\partial x}\dfrac{\partial}{\partial \tilde{y}} = c\dfrac{\partial}{\partial \tilde{x}} - s\dfrac{\partial}{\partial \tilde{y}} \\ \dfrac{\partial}{\partial y} = \dfrac{\partial \tilde{x}}{\partial y}\dfrac{\partial}{\partial \tilde{x}} + \dfrac{\partial \tilde{y}}{\partial y}\dfrac{\partial}{\partial \tilde{y}} = s\dfrac{\partial}{\partial \tilde{x}} + c\dfrac{\partial}{\partial \tilde{y}} \end{matrix}\right\} = U\begin{pmatrix} \dfrac{\partial}{\partial \tilde{x}} \\ \dfrac{\partial}{\partial \tilde{y}} \end{pmatrix}$$

so our differential operator looks like

$$P\left(\begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix}\right) = P\left(U\begin{pmatrix} \partial_{\tilde{x}} \\ \partial_{\tilde{y}} \end{pmatrix}\right)$$

$$= (\partial_{\tilde{x}} \ \partial_{\tilde{y}}) \underbrace{U^T\begin{pmatrix} a & b \\ b & c \end{pmatrix}U}_{\begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}}\begin{pmatrix} \partial_{\tilde{x}} \\ \partial_{\tilde{y}} \end{pmatrix} + \underbrace{(d \ e)U}_{(\tilde{d} \ \tilde{e})}\begin{pmatrix} \partial_{\tilde{x}} \\ \partial_{\tilde{y}} \end{pmatrix} + f$$

or $P(\partial_x, \partial_y) = \lambda_1 \partial_{\tilde{x}}^2 + \lambda_2 \partial_{\tilde{y}}^2 + \tilde{d}\,\partial_{\tilde{x}} + \tilde{e}\,\partial_{\tilde{y}} + f$

$= \tilde{P}(\partial_{\tilde{x}}, \partial_{\tilde{y}})$

so we might as well assume that $b=0$, $a=\lambda_1$, $c=\lambda_2$ in the first place...

The words hyperbolic, parabolic, elliptic come from the graphs of the equation $P(\xi,\eta)=0$

it's really all about the eigenvalues of A...

prototypes:

wave equation: $u_{tt} - u_{xx} = 0$ } hyperbolic

telegraph equation: $u_{tt} + d u_t - u_{xx} = 0$

one way wave eqn: $u_t + a u_x = 0$ } also called hyperbolic (solutions are wave-like)

inviscid Burgers: $u_t + u u_x = 0$ (non-linear)

also called transport equation

heat equation, diffusion equation: $u_t = u_{xx}$ parabolic

Schrödinger equation: $-i u_t = u_{xx}$ ← totally different properties

Laplace equation $u_{xx} + u_{yy} = 0$ } elliptic

Poisson " $u_{xx} + u_{yy} = f(x,y)$

systems { linear elasticity $\mu \Delta u + (\lambda+\mu)\nabla(\nabla\cdot u) = f$

Stokes $-\mu\Delta u + \nabla p = f$, $\nabla\cdot u = 0$

elliptic equations can be made parabolic or hyperbolic
by adding time dependence

unsteady Stokes: $\begin{cases} u_t - \mu\Delta u + \nabla p = f \\ \nabla\cdot u = 0 \end{cases}$
(parabolic)

3d wave: $u_{tt} - \Delta u = 0$
(hyperbolic)

beam equation: $u_t + u_{xxxx} = 0$
(parabolic)

vibrations in elastic medium: $\rho u_{tt} = \mu\Delta u + (\lambda+\mu)\nabla(\nabla\cdot u)$
(hyperbolic)

There are many interesting non-linear PDE

incompressible Navier-Stokes: $\begin{cases} u_t + u\cdot\nabla u = -\nabla p + \mu\Delta u \\ \nabla\cdot u = 0 \end{cases}$

sometimes behaves like elliptic, parabolic or hyperbolic

Eikonal: $|\nabla u| = 1$ (first arrival time of a signal)

Burgers eqn: $u_t + uu_x = u_{xx}$ ← a 1d version of Navier Stokes

Korteweg-deVries (KdV): $u_t + uu_x + u_{xxx} = 0$ ← has soliton solutions

traffic equation: $cu_t - [\sigma(x)u_x]_x = 0$ ← has shocks each type of equation has special features that must be understood and incorporated into the numerical method.

if the solution has shocks, the numerical method must handle discontinuities,

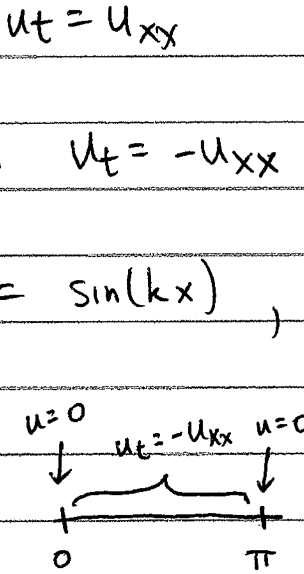
boundary conditions are often the most difficult part of solving numerical PDE

Next time: finite difference methods for the heat equation.

<u>in class exercise:</u>

(1) solve the
- heat equation $u_t = u_{xx}$
- backward heat eqn $u_t = -u_{xx}$

with initial condition $u(x,0) = \sin(kx)$, $k \in \mathbb{R}$
using separation of variables.



(2) how long does the solution of the backward heat equation

with $\begin{cases} b/c\text{'s}: & u(0,t) = u(\pi,t) = 0 \\ i/c\text{'s}: & u(x,0) = (x)(\pi - x) = \pi x - x^2 \end{cases}$

exist? hint: $\int_0^\pi x(\pi - x)\sin(nx)\,dx = \begin{cases} 0 & n \text{ even} \\ 4/n^3 & n \text{ odd} \end{cases}$

Last time:

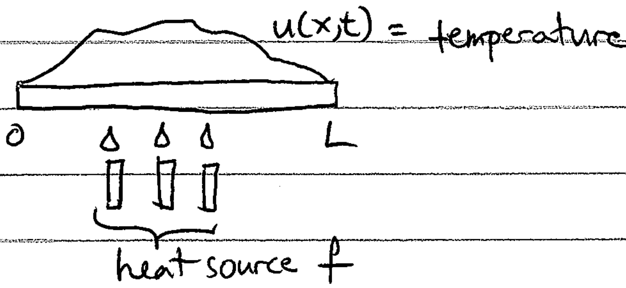
classification of PDE (hyperbolic, parabolic, elliptic)
change of variables
zoo of famous PDE's (no general theory, numerical methods must be tailored to the PDE you're solving)

Today: $u_t = u_{xx}$ 1d heat equation

setup (2 options)

1. rod of finite length $0 \le x \le L$
2. infinite domain $-\infty \le x \le \infty$

case 1:



If you include the heat source, equation is $u_t - u_{xx} = f$
Let's assume $f=0$.

initial conditions: $u(x,0) = g(x)$
boundary conditions: $u(0) = u(L) = 0$

g is the initial temperature distribution (given).
we want to find $u(x,t)$ for $t>0$, $0 \le x \le L$

To solve this equation analytically, use separation of variables:

first look for special solutions of the form

$$u(x,t) = X(x)\,T(t)$$

$$u_t = u_{xx} \Rightarrow X(x)T'(t) = X''(x)T(t)$$

$$\Rightarrow \frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} = C$$

must be a constant

$$\left.\begin{array}{l} X'' = CX \\ X(0) = X(L) = 0 \end{array}\right\} \Rightarrow X(x) = \sin\frac{k\pi x}{L}, \quad C = -\left(\frac{k\pi}{L}\right)^2$$

$$T' = CT \Rightarrow T(t) = e^{Ct}$$

result: $u(x,t) = e^{-\left(\frac{k\pi}{L}\right)^2 t}\sin\frac{k\pi x}{L}$ satisfies $u_t = u_{xx}$

Now use a Fourier sine series to represent the initial condition:

$$g(x) = \sum_{k=1}^{\infty} C_k \sin\frac{k\pi x}{L}, \qquad C_k = \frac{2}{L}\int_0^L g(x)\sin\frac{k\pi x}{L}\,dx$$

Finally, use superposition to obtain the exact soln:

$$u(x,t) = \sum_{k=1}^{\infty} C_k e^{-\left(\frac{k\pi}{L}\right)^2 t}\sin\frac{k\pi x}{L}$$

For the backward heat equation, the Fourier modes grow exponentially in time rather than decay

example: $L = \pi$, $g(x) = \pi x - x^2$

$$c_k = \frac{2}{\pi}\int_0^\pi g(x)\sin kx\,dx = \begin{cases} 0 & k \text{ even} \\ \dfrac{8}{\pi k^3} & k \text{ odd} \end{cases}$$

$$u_t = -u_{xx} \Rightarrow u(x,t) = \sum_{k \text{ odd}} \frac{8}{\pi k^3} e^{k^2 t} \sin kx$$

but for any $t>0$, $\dfrac{8}{\pi k^3} e^{k^2 t} \to \infty$ as $k \to \infty$

so the formula for u diverges for all $t>0$

(backward heat eqn. has no sol'n with this initial condition)

---

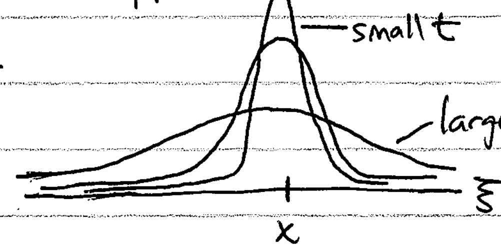
case 2: infinite domain

the boundary conditions $u(0) = u(L) = 0$ are now replaced by the requirement that <u>u remains bounded</u> as $x \to \pm\infty$

this problem may be solved using the <u>Fourier Transform</u> instead of the Fourier sine series above. (see Fritz John's PDE book) exact solution: $$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi$$

requirements on g: continuous and bounded

Note: $\frac{1}{\sqrt{4\pi t}} e^{-\frac{(x-\xi)^2}{4t}}$ is a gaussian centered at x which approaches a δ-function as $t \to 0$.



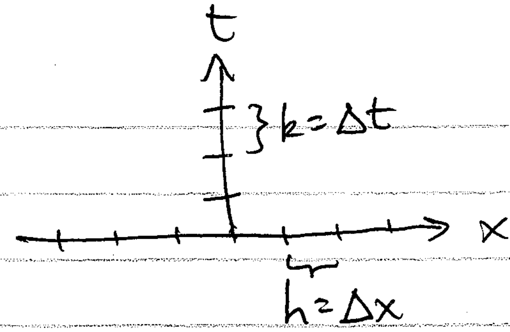
observations: ① the exact solution is a smoothed out version of the initial conditions (larger t = more smoothing)

② the value of u at x depends on all of $g(\xi)$ — information travels infinitely fast.

numerics: why ~~[illegible]~~ use finite differences when we know the exact solution?

1. have to compute the integrals somehow (probably numerically)
2. these exact sol'ns don't generalize to more complicated problems

discretization



notation $\quad u_j^n \approx u(jh, kn)$ ← exact solution

numerical solution ↗ $\qquad$ space ↑ $\quad$ time ↑

What is u_j^0 ? $\quad$ (there are many ways to do IC's!)

we could set $u_j^0 = g(jh, 0)$

or we could average g over some interval.

what is u_t ? $\qquad u_t \approx \dfrac{u(x, t+\Delta t) - u(x,t)}{\Delta t}$

$$= \frac{1}{k}\left[u_j^{n+1} - u_j^n \right]$$

what is u_{xx}? $\qquad u_x \approx \dfrac{u(x+\frac{h}{2}, t) - u(x - \frac{h}{2}, t)}{h}$

$$u_{xx} \approx \frac{\left[\dfrac{u(x+h,t) - u(x,t)}{h}\right] - \left[\dfrac{u(x,t) - u(x-h,t)}{h}\right]}{h}$$

$$\approx \frac{u(x+h,t) - 2\,u(x,t) + u(x-h,t)}{h^2}$$

scheme for $u_t = u_{xx}$:

$$\frac{1}{k}\left[u_j^{n+1} - u_j^n\right] = \frac{1}{h^2}\left[u_{j+1}^n - 2u_j^n + u_{j-1}^n\right]$$

$$u_j^{n+1} = u_j^n + \frac{k}{h^2}\left[u_{j+1}^n - 2u_j^n + u_{j-1}^n\right]$$

$$\boxed{u_j^{n+1} = \nu u_{j+1}^n + (1-2\nu)u_j^n + \nu u_{j-1}^n} \qquad \nu = \frac{k}{h^2}$$

This is a recipe. given the values $\{u_j^n\}_{j=-\infty}^{\infty}$

we use it to find $\{u_j^{n+1}\}_{j=-\infty}^{\infty}$

but remember the exact sol'n. It depends on all of $g(x)$!

let's try it:

$k = 1/4$

$h = 1/4$

$\nu = 4$

solution becomes oscillatory and blows up

t	$-4h$	$-3h$	$-2h$	$-h$	0	h	$2h$	$3h$	$4h$
$3k$	0	64	−336	780	−1015	780	−336	64	0
$2k$	0	0	16	−56	81	−56	16	0	0
k	0	0	0	4	−7	4	0	0	0
0	0	0	0	0	1	0	0	0	0

try again:

$k = 1/64$

$h = 1/4$

$\nu = 1/4$

solution decays and spreads out

t							
	0	1/16	4/16	6/16	4/16	1/16	0
	0	0	1/4	1/2	1/4	0	0
	0	0	0	1	0	0	0

the breakpoint for stability happens at $\nu = \frac{1}{2}$

$$u_j^{n+1} = \nu u_{j+1}^n + \underbrace{(1-2\nu)}_{\text{becomes negative for } \nu > \frac{1}{2}} u_j^n + \nu u_{j-1}^n$$

analysis in the max norm $\|u\|_\infty = \max_{-\infty < j < \infty} |u_j|$

when $\nu \leq \frac{1}{2}$ we have

$$|u_j^{n+1}| \underset{\substack{\text{triangle} \\ \text{inequality}}}{\leq} |\nu||u_{j+1}^n| + |1-2\nu||u_j^n| + |\nu||u_{j-1}^n|$$

$$\leq \underbrace{\left(|\nu| + |1-2\nu| + |\nu|\right)}_{1 \text{ since each is positive}} \|u^n\|_\infty$$

$$\therefore \|u^{n+1}\|_\infty = \max_j |u_j^{n+1}| \leq \|u^n\|_\infty$$

but when $\nu > \frac{1}{2}$ this argument doesn't work as

$$|\nu| + |1-2\nu| + |\nu| = \nu + (2\nu - 1) + \nu = 4\nu - 1 > 1$$

and indeed the initial condition $u^0 = \cdots 1 \ -1 \ 1 \ -1 \cdots$
leads to exponential growth:

$$u_j^0 = (-1)^j, \quad u_j^1 = -(4\nu-1)(-1)^j, \quad u_j^n = (-1)^{n+j}(4\nu-1)^n$$

so for this initial condition, if $\nu > \frac{1}{2}$, we have

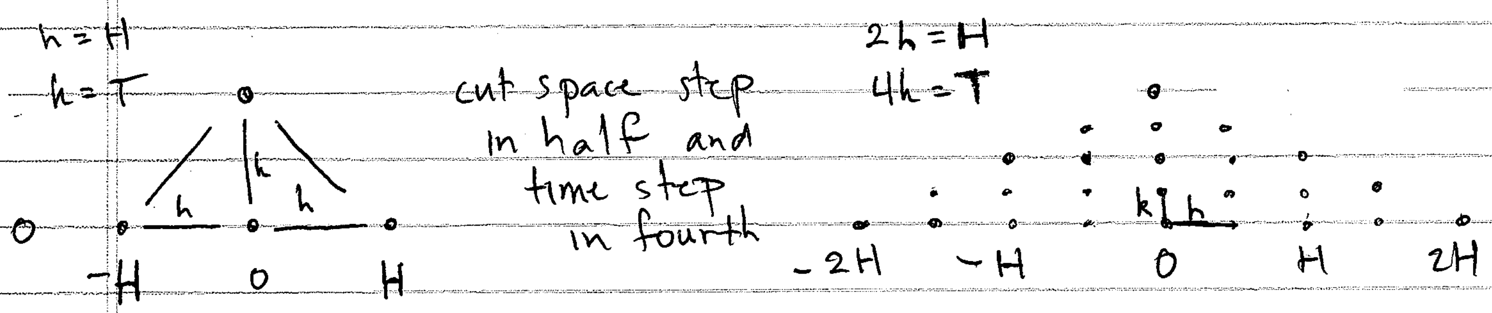
$$\|u^n\|_\infty = (4\nu - 1)^n \underbrace{\|u^0\|_\infty}_{1} \quad \text{exponential growth}$$

<u>def</u>: A method is stable if the solution at a fixed time $T = nk$ (i.e. n increases as k decreases) has norm bounded in terms of it's norm at time 0 <u>independent of the increments h and k</u>

our scheme is stable iff h and k satisfy the additional requirement

$$\frac{k}{h^2} \leq \frac{1}{2}$$

(timestep goes to zero faster than the space step. In the limit you actually see all the initial conditions just as the exact solution does)



Math 228B Lec 3

Last time

- 1d heat equation on finite domain (separation of variables)
- 1d " " " infinite domain (exact solution)
- forward in time, centered in space finite difference method
- preliminary definition of stability of a scheme

Today: error analysis of this scheme.

step 1: show scheme is consistent

step 2: show scheme is stable (do better job of defining stability)

step 3: show that these together imply convergence

Finite difference notation:

consider a function f defined on an evenly spaced grid $x_j = jh$

$$f_j = f(x_j)$$

define

$$D^+ f_j = \frac{f_{j+1} - f_j}{h} \qquad D^- f_j = \frac{f_j - f_{j-1}}{h}$$

$$D^0 f_j = \frac{f_{j+1} - f_{j-1}}{2h} \qquad D^+ D^- f_j = \frac{f_{j+1} - 2f_j + f_{j-1}}{h^2}$$

note that $f = \{f_j\}_{j=-\infty}^{\infty}$ is a sequence and so are D^+f, D^-f, etc.

(i.e. $D^+ f_j$ means $(D^+ f)_j$)

our scheme for $u_t = u_{xx}$ is $\underbrace{D_t^+}_{\substack{\text{operates on "n"} \\ \text{(forward in time)}}} u_j^n = \underbrace{D_x^+ D_x^-}_{\substack{\text{operates on "j"} \\ \text{(centered in space)}}} u_j^n$

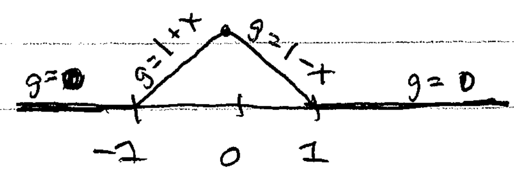
In ODE's, the solution of $y' = f(t,y)$ is guaranteed to exist for $0 \le t \le T$ and be k times continuously differentiable on this interval if

1. f is Lipschitz continuous on $[0,T] \times \mathbb{R}^d$
   (i.e. $\exists L$ s.t. $\|f(t,x) - f(t,y)\| \le L\|x-y\|$ for $0 \le t \le T$, $x,y \in \mathbb{R}^d$)
2. f is k times continuously differentiable.

But for PDE's this is not automatic

$\rightarrow$ high frequency modes can decay too rapidly for $u(x,t)$ to be differentiable at $t=0$.

example: if $g(x) =$



the solution $u(x,t)$ will have $u_t(0,t)$ blow up as $t \searrow 0$.

We will assume the initial condition $g(x)$ is nice enough that the exact solution $u(x,t)$ and a few of its derivatives (say $u_t, u_{tt}, u_x, u_{xx}, u_{xxx}, u_{xxxx}$) are bounded and continuous on the strip

$$-\infty < x < \infty, \quad 0 \le t \le T$$

Taylor's theorem with remainder:

if $f \in C^r[a,x]$ and $f \in C^{r+1}(a,x)$ then

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(r)}(a)}{r!}(x-a)^r + R_r(x)$$

where $R_r(x) = \int_a^x \frac{f^{(r+1)}(t)}{r!}(x-t)^r dt$ ← Cauchy form

$$= \frac{f^{(r+1)}(\xi)}{(r+1)!}(x-a)^{r+1}$$ ← Lagrange form (for some $\xi \in (a,x)$)

Thus we have

$$D^+ f_j = \frac{f(x_j+h) - f(x_j)}{h}$$ assume $f \in C^2[x_j, x_j+h]$

$$= \frac{f(x_j) + hf'(x_j) + \frac{h^2}{2} f''(x_j + \theta h) - f(x_j)}{h}$$

$$= f'(x_j) + \frac{h}{2} f''(x_j + \theta h) \quad \text{for some } \theta \in (0,1)$$

and

$$D^+D^- f_j = \frac{f(x_j+h) - 2f(x_j) + f(x_j-h)}{h^2}$$ assume $f \in C^4[x_j-h, x_j+h]$

$$= \frac{1}{h^2}\left\{ (1-2+1)f(x_j) + [h+(-h)]f'(x_j) + \left[\frac{h^2}{2} + \frac{(-h)^2}{2}\right]f''(x_j) + \left[\frac{h^3}{6} + \frac{(-h)^3}{6}\right]f'''(x_j) + \frac{h^4}{24} f^{(4)}(x_j+\theta_1 h) + \frac{h^4}{24} f^{(4)}(x_j - \theta_2 h) \right\}$$

$$= f''(x_j) + \frac{h^2}{12}\left[\frac{f^{(4)}(x_j+\theta_1 h) + f^{(4)}(x_j-\theta_2 h)}{2}\right] \quad \text{for some } \theta_1, \theta_2 \in (0,1)$$

Now define the truncation error to be what's left over when you plug the exact solution into the scheme:

$$\tau_j^n = D_t^+ u(x_j, t_n) - D^+ D^- u(x_j, t_n)$$

$$= u_t(x_j, t_n) + \frac{k}{2} u_{tt}(x_j, t_n + \theta k)$$

$$- u_{xx}(x_j, t_n) - \frac{h^2}{12}\left[\frac{u_{xxxx}(x_j + \theta_1 h, t_n) + u_{xxxx}(x_j - \theta_2 h, t_n)}{2}\right]$$

so if M is a bound on u_{tt}, u_{xxxx} on the strip $\begin{matrix} -\infty < x < \infty \\ 0 \le t \le T \end{matrix}$ we have

$$|\tau_j^n| \le \left(\frac{k}{2} + \frac{h^2}{12}\right) M$$

If we carry the expansions one step further and take M to be a bound on u_{ttt}, u_{xxxxxx} we get

$$\tau_j^n = \frac{k}{2} u_{tt}(x_j, t_n) - \frac{h^2}{12} u_{xxxx}(x_j, t_n) + \varepsilon_j^n$$

with $$|\varepsilon_j^n| \le \left(\frac{k^2}{6} + \frac{h^4}{360}\right) M$$

but $u_t = u_{xx} \Rightarrow u_{tt} = u_{txx} = u_{xxxx}$ (we're dealing with the exact sol'n here, after all)

so the leading term in τ_j^n is killed if $\frac{k}{2} = \frac{h^2}{12}$

or, recalling that $\nu = \frac{k}{h^2}$, if $\nu = 1/6$

<u>result:</u> $$\tau_j^n = \begin{cases} O(k) & \nu \ne 1/6 \\ O(k^2) & \nu = 1/6 \end{cases}$$ ← holding ν constant while letting $k, h \to 0$

def: A scheme is consistent if $\tau_j^n \to 0$ as $k, h \to 0$

a slightly stronger statement in our case is that the scheme is first order in time and second order in space unless $\nu = 1/6$, in which case it's 2nd order time, 4th order space.

stronger to also specify rate of convergence (the order of the method)

Said differently:

a scheme is consistent if the exact solution of the PDE is an approximate solution of the scheme

it is convergent if the exact solution of the scheme is an approximate solution of the PDE

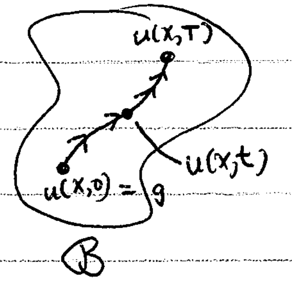
Lax Richtmyer equivalence theorem: A consistent finite difference scheme for a well-posed initial value problem is convergent iff it is stable.

The setting of the Lax-Richtmyer paper is very general:

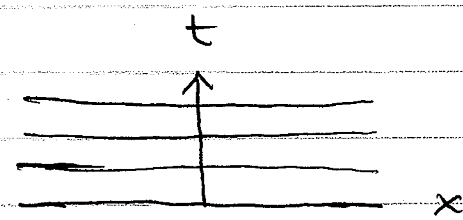
$$u_t = Au \quad (0 \le t \le T) \qquad \text{ODE in a } \underbrace{\text{Banach space}}_{\text{complete normed linear space}}$$
$$u(0) = g$$

In our case the Banach space $\mathcal{B}$ that $u(x,t)$ evolves in is $BC(\mathbb{R})$, the space of bounded continuous functions on $\mathbb{R}$ with norm

$$\|g\| = \max_{-\infty < x < \infty} |g(x)|$$



each point in this space is a function of x (with t frozen)



a time slice of the solution $u(x,t)$ here is a function of x and can be thought of as a point on the solution curve in the space $\mathcal{B}$

other Banach spaces also work nicely (e.g. $L^2(\mathbb{R})$ or $L^1(\mathbb{R})$)

The operator A in our case is the second derivative operator $Au = u_{xx}$

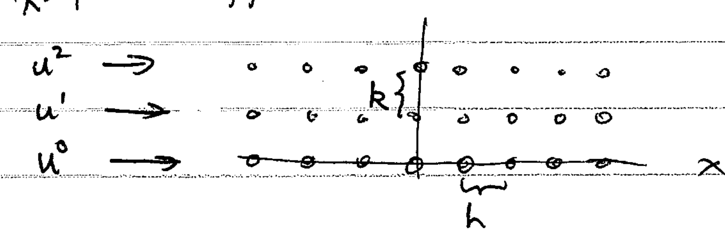
A is not defined for all functions in our space $\mathcal{B}$, but the assumption that the equation "$u_t = Au$, $u(0) = g$" is well posed means the solution ① exists, ② is unique, and ③ depends continuously on the ~~[illegible]~~ initial data g.

in particular, $u(t)$ belongs to the domain of A for $0 \le t \le T$

extra assumption about the initial data

Next we set up a grid and define our scheme as an operator from one discrete time slice to the next

$\mathcal{B}_h$ consists of bounded sequences $\{f_j\}_{j=-\infty}^{\infty}$



$$u^{n+1} = \underbrace{B(k,h)}_{\text{bounded linear operator from } \mathcal{B}_h \text{ to } \mathcal{B}_h} u^n$$

in our case, $B(k,h)u_j = \nu u_{j+1} + (1-2\nu)u_j + \nu u_{j-1}$, $\nu = \frac{k}{h^2}$

Finally, we choose a refinement path relating h to k.
in our case we'll consider ν fixed and set $h = \sqrt{k/\nu}$.
Now there is only one parameter controlling convergence, namely the timestep k.

Lax & Richtmyer give the operator $B(k, \sqrt{k/\nu})$ a new name $C(k)$

scheme: $u^{n+1} = C(k)u^n$

or $u^n = C(k)^n u^0$ ← you've applied the scheme n times starting with the initial condition u^0

<u>def</u>: A scheme is <u>stable</u> if for some $\varepsilon > 0$ the operators

$$C(k)^n \qquad \begin{array}{l} 0 < k \leq \varepsilon \\ 0 \leq nk \leq T \end{array}$$

are uniformly bounded.

This means there is a constant K indep. of k and n such that

$$\|C(k)^n\| \leq K \qquad \begin{array}{l} 0 < k \leq \varepsilon \\ 0 \leq nk \leq T \end{array}$$

(I'll talk more about norms next time. The norm of $C(k)^n$ is the smallest number $\|C(k)^n\|$ s.t. $\|C(k)^n u\| \leq \|C(k)^n\| \cdot \|u\| \quad \forall u \in \mathcal{B}_h$)

In our case $C(k)u_j = \nu u_{j+1} + (1-2\nu)u_j + \nu u_{j-1}$ doesn't depend on k, and we showed last time that

$$\|C(k)\| \le 1 \quad \text{if} \quad \nu \le \tfrac{1}{2}$$

when $\|\cdot\|$ is the infinity norm $\|u\| = \max_{-\infty<j<\infty} |u_j|$

so this scheme is definitely stable.

More generally we can have $\|C(k)\| \le 1 + K_1 k$ for any constant K_1 and the scheme will still be stable.

This is because

$$\|C(k)^n\| \le \|C(k)\|^n \le (1+K_1k)^n$$

$$\le \left(1 + K_1 k + \frac{(K_1k)^2}{2!} + \cdots\right)^n$$

$$= \left(e^{K_1k}\right)^n = e^{K_1(kn)} \le \underbrace{e^{K_1T}}_{K}$$

Now let's prove convergence.

define the error: $e_j^n = u_j^n - u(jh, kn)$

scheme: $u_j^{n+1} = u_j^n + kD_x^+D_x^- u_j^n$

definition of truncation error

exact: $u(jh,(n+1)k) = u(jh,nk) + kD_x^+D_x^- u(jh,nk) + k\tau_j^n$

subtract: $e_j^{n+1} = \underbrace{e_j^n + kD_x^+D_x^- e_j^n}_{C(k)e_j^n} + k\tau_j^n$

recursion for the error

now iterate backwards

$$e_j^n = C(h)\,e_j^{n-1} + k\tau_j^{n-1}$$

$$= C(h)\left[C(h)e_j^{n-2} + k\tau_j^{n-2}\right] + k\tau_j^{n-1}$$

$$\vdots$$

$$= C(h)^n e_j^0 + C(h)^{n-1} k\tau_j^0 + \cdots + C(h)k\tau_j^{n-2} + k\tau_j^{n-1}$$

take norms, use triangle inequality, use $\|C(h)^\ell\| \le K$ for $0 \le \ell \le n$

$$\|e^n\| \le K \underbrace{\|e^0\|}_{0} + K \underbrace{\left[k\|\tau^0\| + k\|\tau^1\| + \cdots + k\|\tau^{n-1}\|\right]}_{\le\, nk \max\limits_{0\le \ell < n} \|\tau^\ell\|}$$

but $nk \le T$, $K=1$ and each $\|\tau^\ell\|$ is bounded by $\left(\frac{k}{2} + \frac{h^2}{12}\right)M$

$$\therefore \quad \|e^n\| \le \begin{cases} TM\left(\frac{k}{2} + \frac{h^2}{12}\right) & \nu \ne 1/6 \\ TM\left(\frac{k^2}{6} + \frac{h^4}{360}\right) & \nu = 1/6 \end{cases}$$

true for all n satisfying $0 \le nk \le T$

$$\therefore \quad \max_n \|e^n\| = \max_{-\infty < j < \infty}\ \max_{0 \le nk \le T} |e_j^n|$$

so the maximum value of the error on the grid goes to zero as $k, h \to 0$ with $\nu = \frac{k}{h^2} \le \frac{1}{2}$ held fixed

Math 228B Lec 4

Last time

finite difference notation D^+, D^-, D^0, D^+D^-

truncation error (definition and bound for heat equation)

consistency, stability, and convergence

setup for Lax-Richtmyer paper

Today: ① crash course in functional analysis

② finish convergence proof for our scheme for $u_t = u_{xx}$

~~③ alternative norms to the max norm~~

functional analysis

core of this subject is figuring out how to do linear algebra in infinite dimensions. Once this is understood, you can go on to study non-linear problems, but we won't be so ambitious

a vector space V is a collection of objects that you can add together and multiply by scalars:

$$f_1, f_2 \in V \Rightarrow \alpha f_1 + \beta f_2 \in V$$

scalars (in $\mathbb{R}$ or $\mathbb{C}$)

a norm is a rule that assigns a real number $\|f\|$ to every element of the space such that

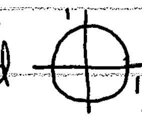
1. $\|f\| \geq 0 \quad \forall f \in V$ and $\|f\| = 0$ iff $f = 0$
2. $\|\alpha f\| = |\alpha| \cdot \|f\|$ homogeneity
3. $\|f_1 + f_2\| \leq \|f_1\| + \|f_2\|$ triangle inequality

A normed space is an example of a metric space where the metric (distance) is given by

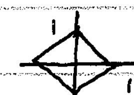
$$d(f,g) = \|f-g\|$$

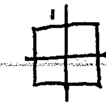
Examples

1. $\mathbb{R}$, $\|x\| = |x|$ absolute value

2. $\mathbb{R}^n$, $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ 2-norm or Euclidean distance (unit balls are round )

(the absolute values are only needed in $\mathbb{C}^n$)

3. $\mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ 1 norm, Manhattan norm (unit balls are diamonds )

4. $\mathbb{R}^n$, $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$ ∞ norm, max norm (unit balls are cubes )

5. $L^2(0,1)$ = "square integrable functions on $(0,1)$"

$$\|f\|_2 = \sqrt{\int_0^1 |f(x)|^2 \, dx}$$

← again the absolute value is only necessary if f takes on complex values

6. $C[0,1]$ = "continuous functions on $[0,1]$" ← it's important that this interval is closed

$$\|f\|_\infty = \max_{0 \le x \le 1} |f(x)|$$

all of these spaces have complex versions (where the set of scalars is $\mathbb{C}$ rather than $\mathbb{R}$)

so $C[a,b]$ can mean $\{f: [a,b] \to \mathbb{R} \mid f \text{ is continuous}\}$

or $\{f: [a,b] \to \mathbb{C} \mid f \text{ is continuous}\}$

depending on the context. We will usually work over $\mathbb{R}$ for simplicity <u>except</u> when the Fourier transform is involved, in which case we're forced to use complex numbers.

norms allow us to measure distances between points in our space. We need them to talk about errors in our numerical solutions.

<u>convergence</u>: A sequence of points $f_1, f_2, \ldots \in V$ converges to $f \in V$

(written $f_n \to f$) if $\|f_n - f\| \to 0$ as $n \to \infty$

alternative notation:

$$\lim_{n\to\infty} f_n = f \qquad \text{if} \qquad \lim_{n\to\infty} \underbrace{\|f_n - f\|}_{\text{a sequence of real numbers}} = 0$$

All of these statements mean the same thing:

for any $\varepsilon > 0$ $\exists N$ s.t. if $n \geq N$ then $\|f_n - f\| < \varepsilon$

Think of ε as a tolerance given to you by the customer and you have to be sure that eventually all the remaining terms in your sequence are within that tolerance.

A <u>Cauchy sequence</u> $f_1, f_2, \ldots$ is a sequence in which the terms eventually stay arbitrarily close to <u>each other</u> :

$$\forall \varepsilon > 0 \quad \exists N \text{ s.t. } \forall n, m \geq N, \quad \|f_n - f_m\| < \varepsilon$$

It's easy to show that every convergent sequence is Cauchy (try it!)
A space is said to be <u>complete</u> if the reverse is also true, i.e every Cauchy sequence converges to an element of the space.

A complete space has no holes. $\mathbb{R}$ is complete
$\mathbb{Q}$ = "set of rational numbers" is not

A <u>Banach space</u> is a complete normed vector space

A <u>Hilbert space</u> is a Banach space where the norm comes from an inner product $\|f\| = \sqrt{(f,f)}$.

examples $\begin{cases} \mathbb{C}^n \text{ with the inner product } (x,y) = x^T \bar{y} \\ L^2(0,1) \text{ with " " } (f,g) = \int_0^1 f(x)\overline{g(x)}\,dx \end{cases}$

complex conjugation

An <u>inner product</u> is a rule that assigns a scalar (f,g) to every pair of points in the space such that:

1. $(\alpha f + \beta g, h) = \alpha(f,h) + \beta(g,h)$ bilinearity
2. $(f,g) = \overline{(g,f)}$ conjugate symmetry
3. $(f,f) > 0$ if $f \neq 0$ positive definiteness

in particular (f,f) is real

it follows from 1 and 2 that $\begin{cases} (0,f) = 0 \\ (f, \alpha g + \beta h) = \bar{\alpha}(f,g) + \bar{\beta}(f,h) \end{cases}$

Banach spaces and Hilbert spaces are the basic arena in which we do numerical analysis. Typically, the elements in these spaces are functions (solutions of PDE's or numerical approximations of these solutions), and we want bounds on the norms of the error.

In linear algebra, linear tranformations are very important, and can be represented by matrices. In infinite dimensions, matrices play a lesser role and we work with the transformation directly.

linear operator: $A : X \to Y$, $A(x+y) = Ax + Ay$
$A(\alpha x) = \alpha Ax$

Banach spaces

linear functional: $f : X \to \mathbb{C}$ $f(x+y) = f(x) + f(y)$
$f(\alpha x) = \alpha f(x)$

special name for the case when the target space is $\mathbb{R}$ or $\mathbb{C}$
(the objects in the space are often functions, so a functional is a "function of functions")

An operator is <u>bounded</u> if there is a constant C s.t.

$$\|Ax\| \leq C\|x\| \quad \forall x \in X$$

The smallest constant C that works is the norm of the operator.

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

sup means supremum, least upper bound. (think of it as the maximum value)

To show that $\|A\| = C$ (e.g. in homework),

① first show that $\|Ax\| \le C\|x\|$ $\forall x$.

② Then show (if you can) that some choice of $x_0 \neq 0$ yields

$$\|Ax_0\| = C\|x_0\|$$

This is always possible in finite dimensions, but in infinite dimensions there may not be a maximizer. (that's why we write sup instead of max). Instead, it suffices that

② If $K < C$ then $\exists x_0$ s.t. $\|Ax_0\| > K\|x_0\|$

in other words, you show that ① C works and ② no smaller choice works

Example: consider the operator $B: \ell^\infty \to \ell^\infty$ ← space of bounded sequences with $\|u\| = \sup_{-\infty<j<\infty} |u_j|$

given by

$$Bu_j = \nu u_{j+1} + (1-2\nu)u_j + \nu u_{j-1}$$

Suppose $0 \le \nu \le 1/2$. I claim $\|B\| = 1$.

step 1: for any j, $|Bu_j| \le |\nu|\,|u_{j+1}| + |1-2\nu|\,|u_j| + |\nu|\,|u_{j-1}|$

$$\le (\nu + 1 - 2\nu + \nu)\|u\| = \|u\|$$

so $\|Bu\| \le \|u\|$ ($C=1$ works)

step 2: The sequence $u_j^0 = 1$ for all j satisfies

$$Bu_j^0 = \nu + (1-2\nu) + \nu = 1$$

so $\|Bu^0\| = 1 = \|u^0\|$ (can't do better than $C=1$)

The norm notation for operators is used because the space of bounded operators $A: X \to Y$ is a Banach space with this norm ($A+B$ is the operator $(A+B)x = Ax + Bx$)

<u>exercise</u>: show that ① $\|A+B\| \le \|A\| + \|B\|$

② If $Y = X$, then $\|AB\| \le \|A\| \cdot \|B\|$

③ $\|A^n\| \le \|A\|^n$

Let's get back to our convergence proof following Lax/Richtmyer.

PDE $\quad u_t = u_{xx}$

$u(x,0) = g(x)$

numerics $\quad D_t^+ u_j^n = D_x^+ D_x^- u_j^n, \quad u_j^0 = g(jh)$

or $\quad u_j^{n+1} = Bu_j^n = \nu u_{j+1}^n + (1-2\nu)u_j^n + \nu u_{j-1}^n$

the exact solution $\quad u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi$

may be thought of as a time dependent curve through the Banach space $\mathcal{B} = BC(\mathbb{R})$ ← bounded, continuous functions on $\mathbb{R}$ (other spaces also work well)

the numerical solution $\quad u_j^n = B^n u_j^0$

may be thought of as repeated iterations of a bounded operator B on the Banach space $\mathcal{B}_h = \ell^\infty$ = "space of bounded sequences"

in general B depends on k and h, but after specifying a refinement path $(h = \sqrt{k/\nu})$ it depends only on k.

(for this refinement path it's a constant function of k)

If we chose a different refinement path, say $h = k$, we would have $B(k)u_j = \frac{1}{k}u_{j+1} + (1 - \frac{2}{k})u_j + \frac{1}{h}u_{j-1}$ $\left(\nu = \frac{k}{h^2} = \frac{1}{k}\right)$

(I can't stand using C as an operator since it's such a good letter for "a large constant", so today we'll use $B(k)$ to represent what I called $C(k)$ last time)

A scheme is <u>stable</u> if $\exists K, \varepsilon$ independent of n, k s.t.

$$\|B(k)^n\| \leq K \quad \text{for} \quad \begin{array}{l} 0 < k \leq \varepsilon \\ 0 \leq nk \leq T \end{array}$$

When $\nu \leq \frac{1}{2}$ is fixed, we have $\|B\| = 1$ so our scheme is stable ($K = 1$, ε arbitrary)

<u>proof of convergence</u>:

define the <u>error</u>: $e_j^n = u_j^n - u(jh, nk)$

<u>scheme</u>: $u_j^{n+1} = u_j^n + kD_x^+ D_x^- u_j^n$

def. of trunc. error

<u>exact</u>: $u(jh, (n+1)k) = u(jh, nk) + kD_x^+ D_x^- u(jh, nk) + k\tau_j^n$

<u>subtract</u>: $e_j^{n+1} = \underbrace{e_j^n + kD_x^+ D_x^- e_j^n}_{Be_j^n} + k\tau_j^n$

now iterate backwards

$$
\begin{aligned}
e_j^n &= B e_j^{n-1} + k\tau_j^{n-1} \\
&= B\left[B e_j^{n-2} + k\tau_j^{n-2}\right] + k\tau_j^{n-1} \\
&\vdots \\
&= B^n e_j^0 + B^{n-1} k\tau_j^0 + \cdots + B k\tau_j^{n-2} + k\tau_j^{n-1}
\end{aligned}
$$

take norms, use triangle inequality, use $\|B^\ell\| \le K$ for $0 \le \ell \le n$:

$$\|e^n\| \le K\|e^0\| + K\underbrace{\left[k\|\tau^0\| + k\|\tau^1\| + \cdots + k\|\tau^{n-1}\|\right]}_{\le nk \max\limits_{0 \le \ell \le n-1} \|\tau^\ell\|}$$

But $\|e^0\| = 0$, $K = 1$,

and each $\|\tau^\ell\|$ is bounded by $\begin{cases} \left(\frac{k}{2} + \frac{h^2}{12}\right)M_1 & \nu \ne 1/6 \\ \left(\frac{k^2}{6} + \frac{h^4}{360}\right)M_2 & \nu = 1/6 \end{cases}$

$$\therefore \|e^n\| \le \begin{cases} \left(\frac{1}{2} + \frac{1}{12\nu}\right)TM_1 k & \nu \ne 1/6 \\ \underbrace{\left(\frac{1}{6} + \frac{1}{360\nu^2}\right)}_{4/15} TM_2 k^2 & \nu = 1/6 \end{cases}$$

true for all n satisfying $0 \le nk \le T$

$M_1 =$ bound on $|u_{xxxx}|$ over the strip $-\infty < x < \infty$, $0 \le t \le T$

$M_2 =$ bound on $|\partial_x^6 u|$ over this strip

$$\therefore \max_n \|e^n\| = \max_{-\infty < j < \infty} \max_{0 \le nk \le T} |e_j^n|$$

so the maximum value of the error on the grid goes to zero as $k, h \to 0$ with $\nu = \frac{k}{h^2} \le \frac{1}{2}$ held fixed.

Math 228B Lec 5

Last time

norms, Banach spaces, linear operators

convergence of $D_t^+ u = D_x^+ D_x^- u$ in max norm (a bit rushed...)

Today: analysis in the 1 norm

~~energy estimates~~

~~Fourier analysis (for 2 norm estimates)~~

---

so far we've measured our errors using the max norm.
Today we'll explore alternatives to this choice.

1. The heat equation does not lead to growth of the 1 norm:

$$\begin{aligned} u_t &= u_{xx} \\ u(x,0) &= g(x) \end{aligned} \quad \rightarrow \quad u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi$$

so $$|u(x,t)| \leq \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} |g(\xi)|\, d\xi$$

(equality if $g(x) \geq 0$ for all x)

$$\therefore \int_{-\infty}^{\infty} |u(x,t)|\, dx \leq \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} |g(\xi)|\, d\xi\, dx$$

always legal to change order of integration when integrand is positive

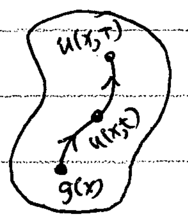
$$= \int_{-\infty}^{\infty} \underbrace{\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi t}} e^{-\frac{(x-\xi)^2}{4t}}\, dx \right)}_{1} |g(\xi)|\, d\xi$$

result: for all positive times t,

$$\int_{-\infty}^{\infty} |u(x,t)|\,dx \;\le\; \int_{-\infty}^{\infty} |g(x)|\,dx$$

and if $g(x) \ge 0 \;\; \forall x$, this is an equality rather than an inequality.
(if we solve on a finite interval with Dirichlet B.C.'s, it's an inequality again)

In our "evolution in a Banach space" picture, we have



$\mathcal{B}$

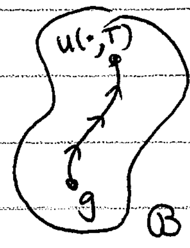
$\mathcal{B} = L^1(\mathbb{R})$ = "integrable functions on $\mathbb{R}$"

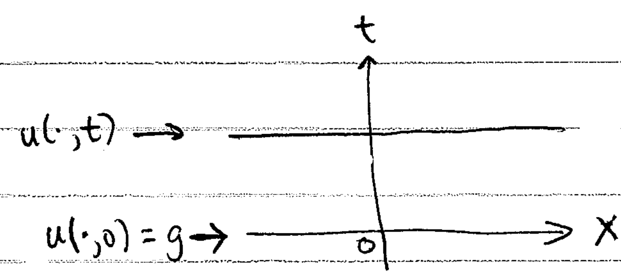
$\|g\| = \int_{-\infty}^{\infty} |g(x)|\,dx$ ← norm in $\mathcal{B}$

the solution $u(x,t)$ of $\left\{ \begin{array}{l} u_t = u_{xx} \\ u(x,0) = g(x) \end{array} \right\}$ satisfies

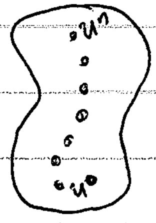
$$\|u(\cdot,t)\| \le \|g\| \qquad (t \ge 0)$$

the dot notation indicates that we're thinking of u as a function of its first argument only (with t given and fixed).
So $u(\cdot,t)$ is the timeslice of the solution at time t:





Next we want to put a norm on our grid that "looks like the 1 norm"



$\mathcal{B}_h$

we choose $\mathcal{B}_h = \ell^1$ = "summable sequences"

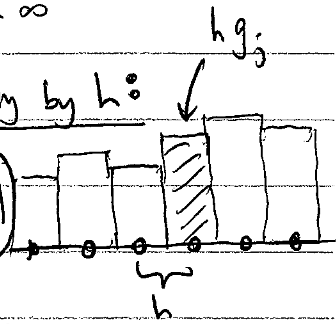
$$\|g\| = h\sum_{j=-\infty}^{\infty} |g_j|$$

In our max-norm analysis, the norms looked the same in $\mathcal{B}$ and $\mathcal{B}_h$:

$\mathcal{B}$ $\quad \|g\|_\infty = \sup_{-\infty \leq x < \infty} |g(x)|$, $\quad \mathcal{B}_h$ $\quad \|g\|_{\infty,h} = \sup_{-\infty < j < \infty} |g_j|$

But in the 1 norm analysis, we need to multiply by h:

$\mathcal{B}$ $\quad \|g\|_1 = \int_{-\infty}^{\infty} |g(x)|\,dx$, $\quad \mathcal{B}_h$ $\quad \|g\|_{1,h} = h\sum_{j=-\infty}^{\infty} |g_j|$



↑ trapezoidal rule of integration

Our scheme is the same as before:

$$u^{n+1} = Bu^n, \qquad Bu_j = \nu u_{j+1} + (1-2\nu)u_j + \nu u_{j-1}$$

Let's prove that $\|B\| = 1$ in $\mathcal{B}_h$ as long as $\nu \leq \frac{1}{2}$:

<u>step 1</u>: Check that $\|Bu\| \leq \|u\|$ for all $u \in \mathcal{B}_h$.

<u>proof</u>:

$$|Bu_j| \leq |\nu|\,|u_{j+1}| + |1-2\nu|\,|u_j| + |\nu|\,|u_{j-1}|$$

$$h\sum_j |Bu_j| \leq h|\nu|\sum_j |u_{j+1}| + h|1-2\nu|\sum_j |u_j| + h|\nu|\sum_j |u_{j-1}|$$

$$= (|\nu| + |1-2\nu| + |\nu|)\, h\sum_j |u_j| = h\sum_j |u_j|$$

$\sum_j |u_{j+1}| = \sum_j |u_j|$ $\qquad\qquad$ $|\nu| \leq \frac{1}{2}$

$$\therefore \|Bu\|_{1,h} \leq \|u\|_{1,h}$$

step 2: Check that 1 is the best possible bound.

Let $u_j^0 = \begin{cases} 1 & j=0 \\ 0 & j \neq 0 \end{cases}$

Then $Bu_j^0 = \begin{cases} \nu & j = \pm 1 \\ 1-2\nu & j=0 \\ 0 & |j| \geq 2 \end{cases}$

$$|\nu| \leq \tfrac{1}{2}$$

and so $h\sum_j |Bu_j^0| = h\left(|\nu| + |1-2\nu| + |\nu|\right) = h = h\sum_j |u_j^0|$

or $\|Bu^0\| = \|u^0\|$ in the discrete 1 norm.

Next we'll assume $g(x)$ is smooth enough that

$$\|\tau^n\|_{1,h} \leq \begin{cases} Ch^2 & \nu \neq 1/6 \\ Ch^4 & \nu = 1/6 \end{cases}$$

we'll talk more about this in a minute...

The error analysis now proceeds exactly as before.

The error $e_j^n = u_j^n - u(jh, kn)$ satisfies the recursion

$$e_j^{n+1} = e_j^n + kD_x^+ D_x^- e_j^n - k\tau_j^n = Be_j^n - k\tau_j^n$$

so that

$$e_j^n = B\underbrace{\left[Be_j^{n-2} - k\tau_j^{n-2}\right]}_{e_j^{n-1}} - k\tau_j^{n-1}$$

$$\vdots$$

$$= B^n e_j^0 - B^{n-1}k\tau_j^0 - \cdots - Bk\tau_j^{n-2} - k\tau_j^{n-1}$$

Finally, since $\|B^\ell\| \le \|B\|^\ell = 1$ for $0 \le \ell \le n$, we have

$$\|e^n\| \le \|B^n\| \cdot \underbrace{\|e^0\|}_{0} + k\|B^{n-1}\|\cdot\|\tau^0\| + \cdots + k\|B\|\,\|\tau^{n-2}\| + k\|\tau^{n-1}\|$$

$$\le k\left[\|\tau^0\| + \cdots + \|\tau^{n-1}\|\right] \le kn \cdot \begin{cases} Ch^2 & \nu \ne 1/6 \\ Ch^4 & \nu = 1/6 \end{cases}$$

But this time $\|e^n\| = h\sum_{j=-\infty}^{\infty} |e_j^n|$

conclusion:

$$\max_{0 \le nk \le T} h\sum_{j=-\infty}^{\infty} |e_j^n| \le \begin{cases} CTh^2 & \nu \ne 1/6 \\ CTh^4 & \nu = 1/6 \end{cases}$$

or

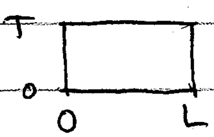
$$k\sum_{n=0}^{T/k} h\sum_j |e_j^n| \le \begin{cases} CT^2h^2 & \nu \ne 1/6 \\ CT^2h^4 & \nu = 1/6 \end{cases}$$

here we summed over n and multiplied by $k = \Delta t$, which introduces another factor of T in the bound

$$k[C + C + \cdots + C] \le nkC \le TC$$

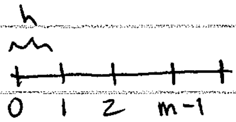
---

how reasonable was our assumption that $\|\tau^n\| \le \begin{cases} Ch^2 & \nu \ne 1/6 \\ Ch^4 & \nu = 1/6 \end{cases}$?

On a finite domain $0 \le x \le L$, our previous assumption that "g is nice enough that the exact solution $u(x,t)$ has 4 (or 6 if $\nu = 1/6$) continuous, bounded derivatives $\partial_x^\ell u$, $0 \le \ell \le 4$ or 6 on the rectangle 

" does the trick

This works because $\quad h\sum_{j=1}^{m-1}|\tau_j^n| \le h(m-1)M \le LM$



$L = mh$

$0 \le x \le L$

M is a bound on $|\tau_j^n|$ (from the max norm analysis)

But on the whole real line a uniform bound on $|\tau_j^n|$ by M does not give a bound on $\|\tau^n\|_{1,h}$ (since $L=\infty$)

Let's go back to our truncation error analysis and try to bound $\|\tau^n\|_{1,h}$ directly. This time we'll use the Cauchy form of Taylor's theorem with remainder:

$$f(x+h) = f(x) + hf'(x) + \cdots + \frac{h^r}{r!}f^{(r)}(x) + R_r(x,h)$$

$$R_r(x,h) = \int_0^h \frac{f^{(r+1)}(x+\xi)}{r!}(h-\xi)^r\,d\xi = \frac{h^{r+1}}{r!}\int_0^1 f^{(r+1)}(x+\theta h)(1-\theta)^r\,d\theta$$

plugging the exact solution into the scheme and simplifying:

$$\tau_j^n = \frac{u(x_j,t_n+k)-u(x_j,t_n)}{k} - \frac{u(x_j+h,t_n)-2u(x_j,t_n)+u(x_j-h,t_n)}{h^2}$$

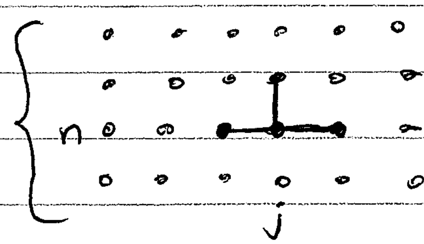
$$= \underbrace{u_t(x_j,t_n) - u_{xx}(x_j,t_n)}_{=0} + k\int_0^1 u_{tt}(x_j,t_n+\theta k)(1-\theta)\,d\theta - \frac{h^2}{6}\int_0^1 u_{xxxx}(x_j+\theta h,t_n)(1-\theta)^3\,d\theta - \frac{h^2}{6}\int_0^1 u_{xxxx}(x_j-\theta h,t_n)(1-\theta)^3\,d\theta$$

an exact formula $\rightarrow$ $\tau_j^n = k\int_0^1 u_{tt}(x_j,t_n+\theta k)(1-\theta)\,d\theta - \frac{h^2}{6}\int_0^1 u_{xxxx}(x_j+\theta h,t_n)(1-\theta)^3\,d\theta - \frac{h^2}{6}\int_0^1 u_{xxxx}(x_j-\theta h,t_n)(1-\theta)^3\,d\theta$

now we use $u_{tt} = u_{xxxx}$ and take absolute values to obtain

$$|\tau_j^n| \le \begin{array}{l} k\int_0^1 |u_{xxxx}(x_j, t_n + \theta k)|(1-\theta)\, d\theta \\ + \frac{h^2}{6}\int_0^1 |u_{xxxx}(x_j + \theta h, t_n)|\,(1-\theta)^3\, d\theta \\ + \frac{h^2}{6}\int_0^1 |u_{xxxx}(x_j - \theta h, t_n)|\,(1-\theta)^3\, d\theta \end{array}$$

an integral of u_{xxxx} over the lines shown, where u is the exact solution.



Next we look at our favorite formula $u(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi$

and differentiate under the integral sign:

$$u_{xxxx}(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} \frac{\partial^4}{\partial x^4} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} \left[\frac{\partial^4}{\partial \xi^4} e^{-\frac{(x-\xi)^2}{4t}}\right] g(\xi)\, d\xi$$

$$\underset{\text{integrate by parts}}{=} \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g_{xxxx}(\xi)\, d\xi$$

so u_{xxxx} is just the solution of the heat equation with initial conditions g_{xxxx}. As a result, we have the bound

$$|u_{xxxx}(x,t)| \le \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} |g_{xxxx}(\xi)|\, d\xi$$

Let's write $\tilde{u} = u_{xxxx}$ and $\tilde{g} = g_{xxxx}$ to avoid all those x's.

can change the order of summation and integration since integrand is positive

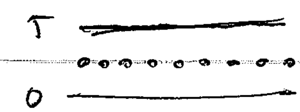
Note that

$$\|\tau^n\|_{1,h} = h\sum_j |\tau_j^n| \le \begin{cases} k\int_0^1 \left(h\sum_j |\tilde{u}(x_j, t_n+\theta k)|\right)(1-\theta)\,d\theta \\ + \frac{h^2}{6}\int_0^1 \left(h\sum_j |\tilde{u}(x_j+\theta h, t_n)|\right)(1-\theta)^3\,d\theta \\ + \frac{h^2}{6}\int_0^1 \left(h\sum_j |\tilde{u}(x_j-\theta h, t_n)|\right)(1-\theta)^3\,d\theta \end{cases}$$

$$\le C\left[k\int_0^1 (1-\theta)\,d\theta + \frac{h^2}{6}\int_0^1 (1-\theta)^3\,d\theta + \frac{h^2}{6}\int_0^1 (1-\theta)^3\,d\theta\right]$$

$$= \left(\frac{k}{2} + \frac{h^2}{12}\right)C$$

worst discrete integral of $|u_{xxxx}|$ in the strip



where $C \ge \max\limits_{\substack{0\le x\le h \\ 0\le t\le T \\ 0\le h\le 1}} h\sum_j |\tilde{u}(x+jh, t)|$

$0 \le h \le 1$ ← arbitrary upper limit on h.

Finally, we note that

$$h\sum_j |\tilde{u}(x+jh,t)| \le h\sum_j \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{(x-\xi+jh)^2}{4t}} |\tilde{g}(\xi)|\,d\xi$$

$$= h\sum_j \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{4t}} |\tilde{g}(x+jh+y)|\,dy$$

$y = \xi - x - jh$

$$= \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{4t}} \left(h\sum_j |\tilde{g}(x+jh+y)|\right) dy$$

$$\le C\left(\frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-y^2/4t}\,dy\right) = C$$

where $C = \max\limits_{\substack{0\le h\le 1 \\ 0\le x\le h}} h\sum_j |g_{xxxx}(x+jh)|$ ← worst discrete integral of $|g_{xxxx}|$

In particular, if $g \in C^4(\mathbb{R})$ and $\exists M$ s.t. $|g^{(\ell)}(x)| \le \frac{M}{1+x^2}$ $\quad x\in\mathbb{R}$, $\ell = 0,1,2,3,4$

then the discrete 1-norm of the truncation error is $O(h^2)$ as required.

Math 228B Lec 6

Last time: analysis of $D_t^+ u = D_x^+ D_x^- u$ in the 1 norm

Today: energy estimates

Fourier analysis of a scheme (analysis in the 2-norm)

Last time we saw that the solution of $u_t = u_{xx}$, $u(x,0) = g(x)$ satisfies

$$\|u(\cdot,t)\|_1 \le \|g\|_1 \quad \text{for } t \ge 0, \qquad \text{i.e.} \int_{-\infty}^{\infty} |u(x,t)|\,dx \le \int_{-\infty}^{\infty} |g(x)|\,dx$$

it's also true that without absolute values,

$$\int_{-\infty}^{\infty} u(x,t)\,dx = \int_{-\infty}^{\infty} g(x)\,dx \qquad (t \ge 0)$$

proof: integrate the representation formula and change order of integration

– or – differentiate the integral: $\frac{d}{dt}\int_{-\infty}^{\infty} u(x,t)\,dx = \int_{-\infty}^{\infty} u_{xx}\,dx = 0$

This remains true on a finite domain with insulating boundary conditions:

$$\frac{d}{dt}\int_0^L u(x,t)\,dx = \int_0^L u_t(x,t)\,dx = \int_0^L u_{xx}(x,t)\,dx = u_x(\cdot,t)\Big|_0^L = 0$$

Let's see what happens if we differentiate the 2-norm:

$$\frac{d}{dt}\int_0^L u^2\,dx = \int_0^L 2u_t u\,dx = \int_0^L 2uu_{xx}\,dx$$

$$= \underbrace{2uu_x\Big|_0^L}_{\text{assume either } u=0 \text{ or } u_x=0 \text{ at each end}} - 2\int_0^L u_x^2\,dx < 0$$

"energy" decreases in time

For the infinite domain, the same is true as long as

$$u(x,t)\,u_x(x,t) \to 0 \quad \text{as } x \to \pm\infty \quad \text{for fixed } t.$$

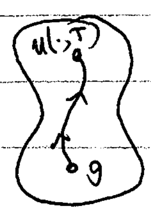
(this is guaranteed if $g(x) \to 0$ as $x \to \pm\infty$)
(g is square integrable and)

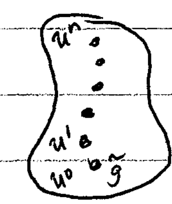
so the exact solution satisfies

$$\|u(\cdot,t)\|_2 \le \|g\|_2 \qquad (t \ge 0)$$

$$\text{where} \quad \|g\|_2 = \sqrt{\int_{-\infty}^{\infty} |g(x)|^2\,dx} \quad \leftarrow L^2 \text{ norm}$$

Our ODE in a Banach space picture looks like





$\mathcal{B} = L^2(\mathbb{R})$ = "square integrable functions"

$$\|g\|_2 = \sqrt{\int_{-\infty}^{\infty} |g(x)|^2\,dx}$$

$\mathcal{B}_h = \ell^2$ = "square summable sequences"

$$\|\tilde{g}\|_{2,h} = \sqrt{h \sum_{j=-\infty}^{\infty} |\tilde{g}_j|^2}$$

$$u_t = u_{xx} \qquad u(\cdot,0) = g$$

$$u^{n+1} = Bu^n \qquad u^0 = \tilde{g} \quad \leftarrow \tilde{g}_j = g(jh)$$

The absolute values in the integrands are there because we are about to consider complex valued functions (due to the Fourier transform)

So what's the norm of our operator $\mathcal{B}$ acting on $\mathcal{B}_h$?

In finite dimensions, the 2-norm of a matrix A is the hardest of the three to compute:

1-norm: $\|A\|_1 =$ "max absolute column sum" $= \max_j \sum_i |A_{ij}|$

∞-norm: $\|A\|_\infty =$ "max absolute row sum" $= \max_i \sum_j |A_{ij}|$

2-norm: $\|A\|_2 =$ largest singular value σ_1

The singular value decomposition of an $n \times n$ matrix looks like

$$A = USV^T, \quad U^TU = I, \; V^TV = I, \; S = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$$

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$

columns of U and V are orthogonal

note that $(U^TU)U^T = (I)U^T$
$U^T(UU^T) = U^T$
so $UU^T = I$ as well

The key feature of an orthogonal matrix is that it preserves norms:

$$\|Ux\|_2^2 = (Ux)^T(Ux) = x^TU^TUx = x^Tx = \|x\|_2^2$$

So A and S have the same 2-norms:

$$\|Ax\| = \|U^TAx\| = \|SV^Tx\| \leq \|S\| \cdot \|V^Tx\| = \|S\| \cdot \|x\|$$

$$\Rightarrow \|A\| \leq \|S\|$$

$$\|Sx\| = \|USx\| = \|AVx\| \leq \|A\| \cdot \|Vx\| = \|A\| \cdot \|x\|$$

$$\Rightarrow \|S\| \leq \|A\|$$

But the 2-norm of a diagonal matrix is the largest absolute value of its entries:

① $\|Sx\|_2^2 = \sum_{i=1}^{n} (\sigma_i x_i)^2 \le \sigma_1^2 \sum_{i=1}^{n} x_i^2 = \sigma_1^2 \|x\|_2^2$

$\uparrow$ $\sigma_i^2 \le \sigma_1^2$ for $i = 1 \dots n$

② $\|Se_1\| = \|\sigma_1 e_1\| = \sigma_1$

So $\|A\| = \|S\| = \sigma_1$

---

if A has complex entries, we use the Hermitian transpose instead

$$A = USV^H, \quad U^H U = I, \; V^H V = I, \; S = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

$(U^H)_{ij} = \bar{U}_{ji}$ ← complex conjugate

$\underbrace{\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_n \ge 0}_{\text{still real}}$

and we still obtain $\|A\| = \|S\| = \sigma_1$

In general, the SVD is hard to compute (take Math 221 to find out how)

If $A = A^H$, then the singular values are the absolute values of the eigenvalues

$$A = U \Lambda U^H = USV^H, \quad \sigma_i = |\lambda_i|$$
$$V(:,i) = \mathrm{sign}(\lambda_i) U(:,i)$$

so the 2-norm of A is the magnitude of the largest eigenvalue (if $A = A^H$)

Now let's get back to our scheme $u^{n+1} = Bu^n$. On a finite interval, B looks like

$$B = \underbrace{\begin{pmatrix} 1-2\nu & \nu & & 0 \\ \nu & 1-2\nu & \nu & \\ & \nu & 1-2\nu & \ddots & \nu \\ 0 & & \ddots & \nu & 1-2\nu \end{pmatrix}}_{J-1 \text{ rows and columns}}$$

if $|\nu| \le \frac{1}{2}$ then:

$\|B\|_1 = |\nu| + |1-2\nu| + |\nu| = 1$ ← sum along column

$\|B\|_\infty =$ same $= 1$ ← sum along row

$\|B\|_2 = ?$

Since B is symmetric, we need to find its largest eigenvalue.

Note that $B = (1-2\nu)\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} + \nu \begin{pmatrix} 0 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 0 \end{pmatrix} = (1-2\nu)I + \nu E$

So it suffices to find the eigenvalues of E

For the continuous problem $u_t = u_{xx}$, the eigenvalues and eigenfunctions of the operator $Au = u_{xx}$ are $Au = \lambda u$, $u = \sin\frac{n\pi x}{L}$, $\lambda = -\left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, 3 \ldots$

By blind luck, these eigenfunctions also work for B and E:

$$U_{j\ell} = \sin\frac{j\ell\pi}{J} \qquad j = 1, 2, \ldots, J-1$$

$$\begin{array}{ccccccc} 0 & & h & & & & L \\ 0 & 1 & 2 & \cdots & & J-1 & J \end{array}$$

$$(EU)_{j\ell} = \sum_m E_{jm} U_{m\ell} = \sin\frac{(j-1)\ell\pi}{J} + \sin\frac{(j+1)\ell\pi}{J}$$

works even when $j=1$ and $j=J-1$ since $\sin(0)=0$, $\sin(\ell\pi)=0$

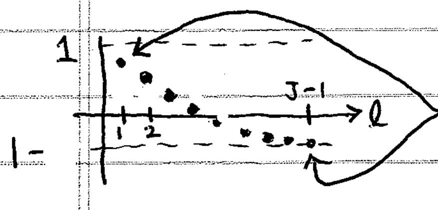
but $\sin x + \sin y = 2\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)$ so

$$(EU)_{j\ell} = 2\sin\frac{j\ell\pi}{J}\cos\frac{\ell\pi}{J} = \underbrace{2\cos\frac{\ell\pi}{J}}_{\lambda_\ell} U_{j\ell}$$

$$EU = U\Lambda$$

columns of U are the eigenvectors of E

$\therefore$ The eigenvalues of B are $\underbrace{(1-2\nu) + 2\nu\cos\frac{\ell\pi}{J}}_{1-4\nu\left(\frac{1-\cos\frac{\ell\pi}{J}}{2}\right) = 1-4\nu\sin^2\left(\frac{\ell\pi}{2J}\right)}$ $\ell = 1, \ldots, J-1$



$\|B\|_2$ is whichever of these is larger in magnitude (so $\|B\|_2 < 1$ if $\nu \le \frac{1}{2}$)

Above we used the usual 2-norm in $\mathbb{R}^{J-1}$, $\|x\|_2^2 = \sum_{j=1}^{J-1} x_j^2$

we would have gotten the same answer $\|B\|_{2,h} = \max_{1 \le \ell \le J-1} \left|1 - 4\nu \sin^2\left(\frac{\ell\pi}{2J}\right)\right|$

using $\|x\|_{2,h}^2 = h \sum_{j=1}^{J-1} x_j^2$ instead.

Now consider the case of an infinite domain. We need to find a way to "diagonalize" our operator B to compute its 2-norm. The tool for doing this is the Fourier series. Normally you think of Fourier series as a way to represent a function $f(x)$ defined on the interval $-\pi \le x < \pi$ via

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}, \qquad c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) e^{-inx}\,dx$$

<u>Theorem</u>: If $f \in L^2(-\pi,\pi)$ (i.e. f is square integrable) then the sequence of numbers $c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) e^{-inx}\,dx$, $-\infty < n < \infty$

belongs to ℓ^2 (i.e. $\sum_n |c_n|^2 < \infty$) and

① $$f = \lim_{N\to\infty} \sum_{n=-N}^{N} c_n e^{inx}$$ (i.e. this limit exists in the Hilbert space $L^2(-\pi,\pi)$)

② $$\sum_{n=-\infty}^{\infty} |c_n|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2\,dx$$ (Parseval's identity)

We're going to turn this idea around and represent sequences by the function that has that sequence as its Fourier coefficients:

<u>Theorem</u>: if $c \in \ell^2$, the limit in ① exists and the resulting function $f \in L^2(-\pi,\pi)$ satisfies ②.

Now let's compute the norm of $Bu_j = \nu u_{j+1} + (1-2\nu)u_j + \nu u_{j-1}$

Let $\hat{u}(\xi) = \sum_j u_j e^{ij\xi}$ $\left(\begin{array}{c} u_j \leftrightarrow c_n \\ \hat{u}(\xi) \leftrightarrow f(x) \end{array}\right)$

Then
$$\widehat{Bu}(\xi) = \sum_j Bu_j e^{ij\xi}$$
$$= \sum_j \left[\nu u_{j+1} e^{ij\xi} + (1-2\nu)u_j e^{ij\xi} + \nu u_{j-1} e^{ij\xi}\right]$$

change dummy indices in 1st and 3rd sums

$$= \sum_j \left[\nu u_j e^{i(j-1)\xi} + (1-2\nu)u_j e^{ij\xi} + \nu u_j e^{i(j+1)\xi}\right]$$
$$= \left(\nu e^{-i\xi} + (1-2\nu) + \nu e^{i\xi}\right)\sum_j u_j e^{ij\xi}$$
$$= \left(1 - 2\nu + 2\nu\cos\xi\right)\hat{u}(\xi)$$
$$= \underbrace{\left[1 - 4\nu\sin^2(\xi/2)\right]}_{G(\xi)}\hat{u}(\xi)$$

$G(\xi)$ ← amplification factor

So applying B to a sequence is the same as multiplying its Fourier series by $G(\xi)$. The amplification factor plays the same role here that the singular value matrix $S = \begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_n\end{pmatrix}$ played for matrices.

<u>Claim</u>: $\|B\|_{2,h} = \max_{-\pi \le \xi \le \pi} |G(\xi)|$

proof next time.

Math 228B Lec 7

Last time: analysis in the 2-norm < finite interval: SVD / infinite domain: Fourier series

Today: finish stability analysis in 2-norm

more about the amplification factor

how to fix the broken $\nu = 1/6$ in the homework

Clarification: if you include the constants, the 3-d heat equation looks like

$$\rho C \frac{\partial u}{\partial t} - \nabla \cdot (k \nabla u) = f \quad \leftarrow \text{heat source } \left(\frac{cal}{cm^3 \cdot s}\right)$$

$$u|_{t=0} = g \quad \leftarrow \text{initial conditions}$$

u = temperature (K)

C = specific heat $\left(\frac{cal}{g \cdot K}\right)$

k = thermal conductivity $\left(\frac{cal}{cm\ s\ K}\right)$ $\leftarrow$ $\underbrace{J = -k\nabla u}_{\text{Fourier's law}}$

ρ = density $\left(\frac{g}{cm^3}\right)$

flux $\left(\frac{cal}{cm^2\ s}\right)$ $\downarrow$ J

so the right thing to call energy is $\iiint \rho C u \, dV$

or in 1d without constants: $\int_0^L u\,dx$

we saw that $\boxed{\text{insulating B.C.'s} \Rightarrow \frac{d}{dt}\int_0^L u\,dx = 0}$ energy conservation

For most other equations, the energy is the integral of the square of something.

(Poisson equation, elasticity, wave equation, Maxwell equations, Stokes eqs., etc.)

Last time we saw that the 2-norm of a matrix is the largest singular value of the matrix

$$A = USV^H, \qquad \|A\|_2 = \|S\|_2 = \sigma_1 \qquad \left\{ \begin{array}{l} U^H U = I \\ V^H V = I \\ S = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \\ \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \end{array} \right.$$

The key idea was that the orthogonal (or unitary in the complex case) matrices U & V do not change the 2-norms of vectors or ~~operators~~ matrices.

The same is true in infinite dimensions. The mapping between square integrable functions on $(-\pi,\pi)$ and their fourier coefficents preserves the 2-norm (up to a factor of $\frac{h}{2\pi}$):

ℓ^2 (square summable sequences) $\qquad\qquad$ $L^2(-\pi,\pi)$ (square integrable functions)

$$\{u_j\}_{j=-\infty}^{\infty} \xrightarrow{\mathcal{F}} \hat{u}(\xi) = \sum_j u_j e^{-ij\xi}$$

$$\check{f}_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\xi) e^{ij\xi}\, d\xi \xleftarrow{\mathcal{F}^{-1}} f(\xi)$$

The point is that this mapping is 1-1, onto, and $\overbrace{\text{isometric}}^{\text{norm preserving}}$ (up to the $\frac{h}{2\pi}$ factor)

$$\int_{-\pi}^{\pi} |\hat{u}(\xi)|^2 d\xi = \frac{2\pi}{h}\left(h\sum_j |u_j|^2\right) \longleftarrow \|\mathcal{F}\| = \sqrt{\frac{2\pi}{h}}$$

and

$$h\sum_j |\check{f}_j|^2 = \frac{h}{2\pi}\int_{-\pi}^{\pi} |f(\xi)|^2 d\xi \longleftarrow \|\mathcal{F}^{-1}\| = \sqrt{\frac{h}{2\pi}}$$

we showed that our scheme $Bu_j = \nu u_{j+1} + (1-2\nu)u_j + \nu u_{j-1}$

maps a sequence $\{u_j\}_{j=-\infty}^{\infty}$ with Fourier series $\hat{u}(\xi) = \sum_j u_j e^{-ij\xi}$

to the sequence $\{Bu_j\}_{j=-\infty}^{\infty}$ with $\widehat{Bu}(\xi) = \underbrace{\left[1 - 4\nu \sin^2\left(\frac{\xi}{2}\right)\right]}_{G(\xi)} \hat{u}(\xi)$

$G(\xi)$ ← amplification factor

so now we have two ways of applying B:

$$\begin{array}{ccc} \mathcal{B}_h & \xrightarrow{\mathcal{F}} & L^2(-\pi,\pi) \\ \downarrow B & & \downarrow \mathcal{G} \\ \mathcal{B}_h & \xleftarrow{\mathcal{F}^{-1}} & L^2(-\pi,\pi) \end{array}$$

$\mathcal{G}f(\xi) = G(\xi)f(\xi)$

$\mathcal{G}$ is the operator "take $f(\xi)$ and multiply it by $G(\xi)$"

$B = \mathcal{F}^{-1}\mathcal{G}\mathcal{F}$ ← just like our SVD $A = USV^H$

multiplying by a diagonal matrix is similar to multiplying by a function: $(Sx)_i = \sigma_i x_i$

each component gets multiplied by something, but there's no mixing of the components.

$\therefore \|B\| \le \|\mathcal{F}^{-1}\| \cdot \|\mathcal{G}\| \cdot \|\mathcal{F}\|$

$= \sqrt{\frac{h}{2\pi}}\,\|\mathcal{G}\|\,\sqrt{\frac{2\pi}{h}} = \|\mathcal{G}\|$

and $\mathcal{G} = \mathcal{F}\,B\,\mathcal{F}^{-1}$

so $\|\mathcal{G}\| \le \|\mathcal{F}\| \cdot \|B\| \cdot \|\mathcal{F}^{-1}\| = \|B\|$

$\boxed{\text{conclusion: } \|B\| = \|\mathcal{G}\|}$ (just like $\|A\| = \|S\| = \sigma_1$)

Our amplification factors $G(\xi)$ will always be continuous functions on the interval $-\pi \le \xi \le \pi$.

Claim: $\|\mathcal{G}\| = \max\limits_{-\pi \le \xi \le \pi} |G(\xi)|$ call the RHS C for now.

<u>proof</u>: <u>step 1</u>: show $\|\mathcal{G}f\| \le C\|f\|$ for all f

$$\|\mathcal{G}f\|^2 = \int_{-\pi}^{\pi} |\mathcal{G}f(\xi)|^2 d\xi = \int_{-\pi}^{\pi} |G(\xi) f(\xi)|^2 d\xi$$

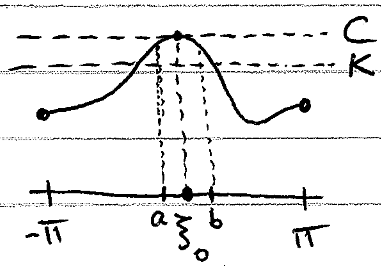
$$\le C^2 \int_{-\pi}^{\pi} |f(\xi)|^2 d\xi = C^2 \|f\|^2$$

key step: $|G(\xi)|^2 \le C^2$ for every $\xi \in [-\pi, \pi]$

<u>step 2</u>: show that if $K < C$ then $\overset{\text{exists}}{\exists} f$ s.t. $\|\mathcal{G}f\| > K\|f\|$

(i.e. no smaller constant than C will work)

<u>idea</u>: $|G(\xi)|$ is a continuous function, so it



achieves its maximum value at some point $\xi_0 \in [-\pi, \pi]$ and there's a neighborhood $[a, b]$ containing ξ_0 so that $|G(\xi)| > K$ for $a \le \xi \le b$.

now define $f(\xi) = \begin{cases} 0 & \xi < a \\ 1 & a \le \xi \le b \\ 0 & b < \xi \end{cases}$. Then

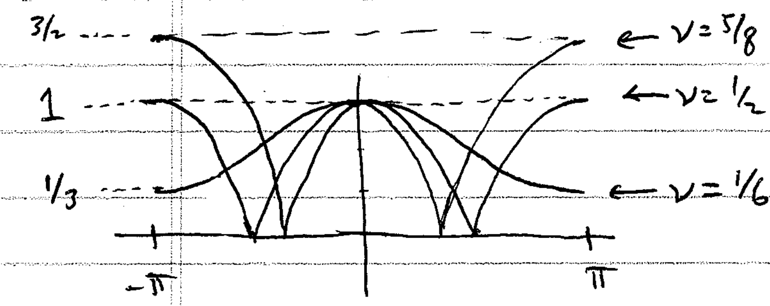
$$\|\mathcal{G}f\|^2 = \int_{-\pi}^{\pi} |G(\xi) f(\xi)|^2 d\xi = \int_a^b |G(\xi)|^2 d\xi > \int_a^b K^2 d\xi = K^2 (b-a)$$

and $\|f\|^2 = \int_{-\pi}^{\pi} |f(\xi)|^2 d\xi = \int_a^b 1^2 d\xi = b - a$ so $\|\mathcal{G}f\| > K\|f\|$ as claimed.

conclusion: the 2-norm of a finite difference scheme is the maximum absolute value of the amplification factor $G(\xi)$.

for our scheme we found that $G(\xi) = 1 - 4\nu \sin^2(\xi/2)$

$|G(\xi)| = |1 - 4\nu \sin^2(\xi/2)|$



as expected, the transition from $\|B\| > 1$ to $\|B\| = 1$ happens when $\nu = 1/2$.

The rest of the convergence proof is the same as before:

assume g is nice enough that $\|\tau^n\|_{2,h} \le \begin{cases} Ch^2 & \nu \ne 1/6 \\ Ch^4 & \nu = 1/6 \end{cases}$

the error $e_j^n = u_j^n - u(jh, nk)$ satisfies $e^{n+1} = Be^n - k\tau^n$

backward iteration gives $\displaystyle\max_{0 \le nk \le T} \sqrt{h \sum_j |e_j^n|^2} \le \begin{cases} CTh^2 & \nu \ne 1/6 \\ CTh^4 & \nu = 1/6 \end{cases}$

the condition $g \in C^4(\mathbb{R})$ and $\exists M$ s.t. $|g^{(l)}(x)| \le \frac{M}{1+x^2}$ $\quad x \in \mathbb{R}$, $l = 0,1,2,3,4$

is sufficient to ensure $\|\tau^n\|_{2,h} \le Ch^2$

and $g \in C^6(\mathbb{R})$, $|g^{(l)}(x)| \le \frac{M}{1+x^2}$ $\quad 0 \le l \le 6$ ensures $\|\tau^n\|_{2,h} \le Ch^4$

In the homework, you'll find that $\nu = 1/6$ is no longer magic for the scheme

$$D_t^+ u = D_x^+ D_x^- u + 10 D_x^0 u$$

Let's try to figure out why and see what we can do about it.

Taylor expansions:

$$D_t^+ u = \frac{u(x_j, t_n + k) - u(x_j, t_n)}{k} = u_t + \frac{k}{2} u_{tt} + \cdots$$

$$D_x^+ D_x^- u = \qquad = u_{xx} + \frac{h^2}{12} u_{xxxx} + \cdots$$

$$D_x^0 u = \frac{u(x_j + h, t_n) - u(x_j - h, t_n)}{2h} = u_x + \frac{h^2}{6} u_{xxx} + \cdots$$

$$\tau_j^n = \frac{k}{2} u_{tt} - \frac{h^2}{12} u_{xxxx} - 10 \frac{h^2}{6} u_{xxx} + O(h^2 + h^4)$$

exact soln satisfies $u_t = u_{xx} + 10 u_x$

$$\text{so} \quad u_{tt} = u_{txx} + 10 u_{tx}$$

$$= u_{xxxx} + 10 u_{xxx} + 10(u_{xxx} + 10 u_{xx})$$

$$= u_{xxxx} + 20 u_{xxx} + 100 u_{xx}$$

$$\therefore \tau_j^n = \underbrace{\left(\frac{k}{2} - \frac{h^2}{12}\right)}_{0 \text{ if } \nu = 1/6} u_{xxxx} + \underbrace{\left(10k - \frac{10}{6} h^2\right)}_{0 \text{ if } \nu = 1/6} u_{xxx} + \underbrace{50k}_{\text{not zero!}} u_{xx}$$

so actually we want

$$\tau_j^n = \frac{k}{2} u_{tt} - \frac{h^2}{12} u_{xxxx} - 10 \frac{h^2}{6} u_{xxx} - \frac{50}{6} h^2 u_{xx} + \cdots$$

but we know how to approximate u_{xx} (just use $D_x^+ D_x^- u$)
so a better scheme would be

$$D_t^+ u = \left(1+\frac{50}{6}h^2\right) D_x^+ D_x^- u + 10\, D_x^0 u$$

try it... it works! (gives $O(k^2+h^4)$ errors)

more general schemes

consider the shift operator $S u_j = u_{j+1}$ on ℓ^2

it looks like an infinite matrix with 1's on the superdiagonal $\begin{pmatrix} \ddots & \ddots & & \\ & 0 & 1 & \\ & & 0 & 1 \\ & & & \ddots \end{pmatrix}$

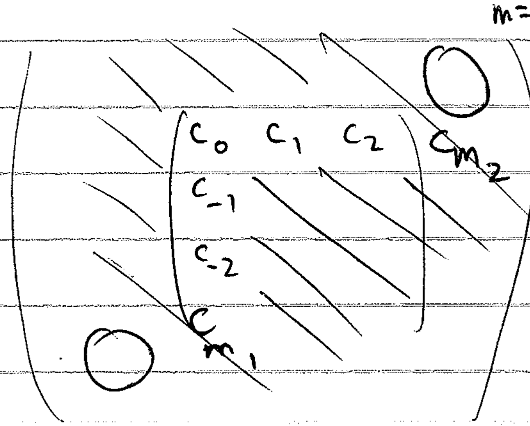
it's inverse S^{-1} has 1's on the subdiagonal $S^{-1} u_j = u_{j-1}$

(the finite dimensional version of $S = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$ is not invertible)

our scheme $u^{n+1} = B u^n$ has $B = \nu S^1 + (1-2\nu)\overbrace{S^0}^{I} + \nu S^{-1}$

and we can write $D_x^+ u = \frac{S - I}{h} u$

A general finite difference scheme looks like $B = \sum_{m=m_1}^{m_2} c_m S^m$



constant along diagonals

Its amplification factor may be computed as

$$\widehat{Bu}(\xi) = \sum_j Bu_j e^{-ij\xi} \qquad \sum_m = \sum_{m=m_1}^{m_2}$$

$$= \sum_j \left(\sum_m c_m u_{j+m} \right) e^{-ij\xi}$$

$$= \sum_m \sum_j c_m u_{j+m} e^{-ij\xi} \qquad \ell = j+m$$

$$= \sum_m \sum_\ell c_m u_\ell e^{-i(\ell-m)\xi}$$

$$= \underbrace{\sum_m c_m e^{im\xi}}_{G(\xi)} \underbrace{\sum_\ell u_\ell e^{-i\ell\xi}}_{\hat{u}(\xi)}$$

you can think of $w_j = e^{ij\xi}$ as an infinite column vector indexed by j with ξ as a fixed parameter

Then $$Bw_j = \sum_m c_m w_{j+m} = \sum_m c_m e^{i(j+m)\xi} = \left(\sum_m c_m e^{im\xi} \right) \underbrace{e^{ij\xi}}_{w_j}$$

or $Bw = G(\xi) w$

so $G(\xi)$ is the eigenvalue associated with the eigenvector w

only problem is, w is not square summable, so $w \notin \mathbb{B}_h$

(the operator $\mathbb{B}$ doesn't have any eigenvalues or eigenvectors)
since none of the candidate eigenvectors are "legal"

Math 228 B Lec 8

Last time: von Neumann stability analysis

$$B = \mathcal{F}^{-1} \mathcal{G} \mathcal{F}, \quad \|B\|_{2,h} = \|\mathcal{G}\|_2 = \|G\|_\infty$$

fixing the broken $\nu = 1/6$ scheme for $u_t = u_{xx} + 10u_x$

Today: amplification factors for arbitrary schemes

bounds on the finite dimensional versions of B in terms of $G(\xi)$

~~implicit methods~~

General explicit finite difference schemes

shift operator: $S u_j = u_{j+1}$ (maps ℓ^2 to ℓ^2)

inverse: $S^{-1} u_j = u_{j-1}$

(the finite dimensional version $S = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}$ is not invertible)

but the circulant version $\longleftrightarrow$ $S = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ is invertible

---

we can write our previous operators in terms of S:

$$D_x^+ u = \frac{S - I}{h} u, \quad D_x^- u = \frac{I - S^{-1}}{h} u, \quad D_x^+ D_x^- u = \frac{S - 2I + S^{-1}}{h^2} u$$

our favorite scheme: $u^{n+1} = Bu^n$, $\quad B = \nu S^1 + (1-2\nu) S^0 + \nu S^{-1}$

a general scheme: $B = \sum_{m=m_1}^{m_2} c_m S^m$

constant along diagonals (Toeplitz)

infinite matrix:

$$\begin{pmatrix} \ddots & & & & & & \\ & c_0 & c_1 & c_2 & c_{m_2} & 0 & 0 \\ & c_{-1} & \ddots & & & & \\ & c_{-2} & & \ddots & & & \\ & c_{m_1} & & & \ddots & & \\ & 0 & & & & \ddots & \\ & 0 & & & & & \ddots \end{pmatrix}$$

Its amplification factor may be computed as

$$\widehat{Bu}(\xi) = \sum_j Bu_j \, e^{-ij\xi} \qquad\qquad \sum_m = \sum_{m=m_1}^{m_2}$$

$$= \sum_j \left(\sum_m c_m u_{j+m} \right) e^{-ij\xi}$$

$$= \sum_m \sum_j c_m u_{j+m} e^{-ij\xi} \qquad \ell = j+m$$

$$= \sum_m \sum_\ell c_m u_\ell e^{-i(\ell-m)\xi}$$

$$= \underbrace{\sum_m c_m e^{im\xi}}_{G(\xi)} \; \underbrace{\sum_\ell u_\ell e^{-i\ell\xi}}_{\hat{u}(\xi)}$$

note that if B is symmetric, i.e. $m_1 = -m_2$ and $c_{-m} = c_m$ then $G(\xi)$ is real valued:

$$G(\xi) = \sum_m c_m e^{im\xi} = c_0 + \sum_{m=1}^{m_2} c_m \underbrace{\left(e^{im\xi} + e^{-im\xi}\right)}_{2\cos m\xi}$$

you can think of $w_j = e^{ij\xi}$ as an infinite column vector indexed by j with ξ as a fixed parameter

Then $$Bw_j = \sum_m c_m w_{j+m} = \sum_m c_m e^{i(j+m)\xi} = \left(\sum_m c_m e^{im\xi}\right) \underbrace{e^{ij\xi}}_{w_j}$$

or $Bw = G(\xi) w$

so $G(\xi)$ is the eigenvalue associated with the eigenvector w

only problem is, w is not square summable, so $w \notin \mathcal{B}_h$

(the operator B doesn't have any eigenvalues or eigenvectors) since none of the candidate eigenvectors are "legal"

The amplification factor allows us to compute the norm of B for the infinite domain problem. What does it tell us in the finite domain case? The answer depends on the boundary conditions.

Case 1: Dirichlet B.C.'s

$$\begin{array}{ll} j: & 0 \;\; 1 \;\; 2 \;\; \cdots \;\; J \\ x: & 0 \;\; h \;\; \cdots \;\; L \end{array}$$

Let's keep the letter B for the operator on ℓ^2 and use the letter A for the one on $\mathbb{R}^{J-1}$

so
$$A = \begin{pmatrix} c_0 & c_1 & \cdots & c_{m_2} & 0 & \cdots & 0 \\ c_{-1} & \ddots & & & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & 0 \\ c_{m_1} & & & \ddots & & & c_{m_2} \\ 0 & \ddots & & & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & c_1 \\ 0 & \cdots & 0 & c_{m_1} & \cdots & c_{-1} & c_0 \end{pmatrix}$$

since A is a proper submatrix of B, $\|A\|_{2,h} \le \|B\|_{2,h}$

reason: if $x \in \mathbb{R}^{J-1}$, define $u \in \ell^2$ via $u_j = \begin{cases} 0 & j \le 0 \\ x_j & 1 \le j \le J-1 \\ 0 & j \ge J \end{cases}$

Then $\|u\|^2_{2,h} = h \sum_{j=-\infty}^{\infty} |u_j|^2 = h \sum_{j=1}^{J-1} |x_j|^2 = \|x\|^2_{2,h}$

and Ax is a subvector of Bu:

$$Bu = \left(\begin{array}{c|c|c} B_{11} & B_{12} & O \\ \hline B_{21} & A & B_{23} \\ \hline O & B_{32} & B_{33} \end{array}\right) \begin{pmatrix} 0 \\ x \\ 0 \end{pmatrix} = \begin{pmatrix} B_{12}x \\ Ax \\ B_{32}x \end{pmatrix}$$

e.g. $B_{12} = \begin{pmatrix} \vdots & & \vdots \\ 0 & & 0 \\ 0 & & \vdots \\ c_{m_2} & & 0 \\ \vdots & \ddots & 0 \\ c_2 & & c_{m_2} \\ c_1 \; c_2 & \cdots & c_{m_2} \end{pmatrix}$

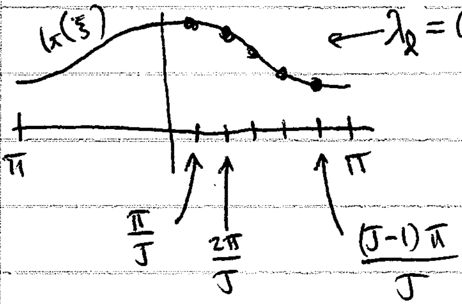
so $\|Bu\|^2_{2,h} = \|Ax\|^2_{2,h} + \overbrace{\|B_{12}x\|^2_{2,h} + \|B_{32}x\|^2_{2,h}}^{\text{non-negative}}$

$\therefore \|Ax\|_{2,h} \le \|Bu\|_{2,h} \le \|B\|_{2,h} \underbrace{\|u\|_{2,h}}_{\|x\|_{2,h}}$

$$A = \begin{array}{c} \overbrace{\qquad\qquad}^{J-1 \text{ columns}} \\ \begin{pmatrix} \alpha & \beta & & 0 \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ 0 & & \beta & \alpha \end{pmatrix} \end{array} \Bigg\} J-1 \text{ rows}$$

<u>case 1a:</u> B is symmetric and tri-diagonal

then not only is $\|A\| \le \|B\|$ but the eigenvalues of A are the values of G sampled at equal intervals:



$\leftarrow \lambda_\ell = G\left(\frac{\ell\pi}{J}\right) \quad 1 \le \ell \le J-1$

$x \in \mathbb{R}^{J-1}, \quad x_j = \sin\frac{j\pi\ell}{J} \quad 1 \le j \le J-1$

$w \in \ell^2, \quad w_j = e^{ij\xi} \quad -\infty < j < \infty, \ \xi = \frac{\ell\pi}{J}$

$x_j = \mathrm{Im}(w_j)$

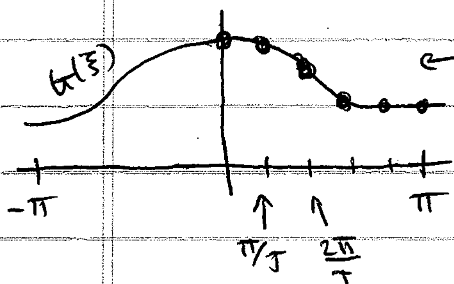
taking the imaginary part of $Bw = G(\xi)w$ and using $\mathrm{Im}(w_0) = \mathrm{Im}(w_J) = 0$ and $G(\xi)$ real

gives $Ax = G\left(\frac{\ell\pi}{J}\right)x$ as claimed.

<u>case 2a:</u> Neumann B.C.'s, B symmetric & tri diagonal

$$A = \underbrace{\begin{pmatrix} \alpha & 2\beta & & & \\ \beta & \alpha & \beta & & \\ & \beta & \alpha & \beta & \\ & & \ddots & \ddots & \ddots \\ & & \beta & \alpha & \beta \\ & & & 2\beta & \alpha \end{pmatrix}}_{(J+1)\times(J+1)}$$

This time $G(0)$ and $G(\pi)$ are also eigenvalues



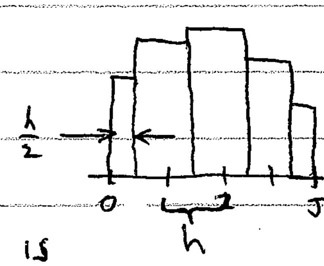
$\leftarrow \lambda_\ell = G\left(\frac{\ell\pi}{J}\right) \quad 0 \le \ell \le J$

$x_j = \cos\frac{j\pi\ell}{J} \quad 0 \le j \le J$

$w_j = e^{ij\xi} \quad -\infty < j < \infty, \ \xi = \frac{\ell\pi}{J}$

$x_j = \mathrm{Re}(w_j)$

$$\left.\begin{array}{c} Bw = G(\xi)w \\ \mathrm{Re}(w_{-1}) = \mathrm{Re}(w_1) \\ \mathrm{Re}(w_{J+1}) = \mathrm{Re}(w_{J-1}) \\ G(\xi) \text{ real} \end{array}\right\} \Rightarrow Ax = G(\xi)x \ \checkmark$$

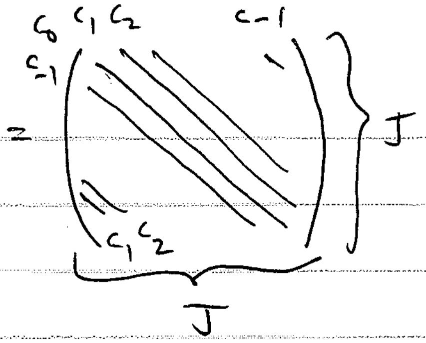


The correct discrete inner product is

$$(u,v) = \frac{h}{2}u_0\bar{v}_0 + \sum_{j=1}^{J-1} u_j\bar{v}_j + \frac{h}{2}u_J v_J$$

in this inner product, A is self-adjoint: $(Au, v) = (u, Av)$

case 3 : Periodic B.C.'s , B arbitrary

$A =$ 

these values represent the solution

$$\begin{array}{lccccc} j = & 0 & 1 & 2 & J-1 & J \\ x = & 0 & h & 2h & & L \end{array}$$

when computing e.g. $D_x^+ D_x^- u_j$ at $j=0$, "$j-1$" means $J-1$ instead of -1

$j = J-1$, "$j+1$" means 0 instead of J

This time the "illegal" eigenvectors $w_j = e^{ij\xi}$ of B are eigenvectors of A as long as ξ respects the periodic b.c.'s (no need to assume B is symmetric or tridiagonal)

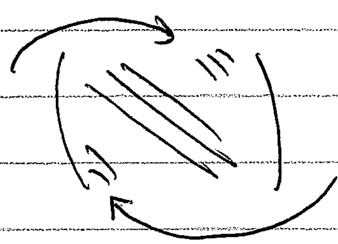
requirement : $e^{iJ\xi} = e^{i0\xi}$ or $\xi = \frac{2\pi l}{J}$

spacing is double what it was in the Dirichlet & Neumann cases

rows 0 through $J-1$ of B :

$$\left(\begin{array}{ccccc|ccccc|ccc} \cdots & 0 & c_{m_1} & c_{-2} & c_{-1} & c_0 & c_1 & \cdots & c_{m_2} & 0 \cdots 0 & 0 & \cdots 0 & 0 \cdots \\ \cdots & 0 & 0 & \ddots & c_{m_1} & c_{-1} & \ddots & & \ddots & c_{m_2} & 0 & & \\ & \vdots & \vdots & & 0 & c_{m_1} & & \ddots & & c_1 & c_{m_2} & \ddots & 0 \quad 0 \cdots \\ \cdots & 0 & 0 & \cdots & 0 & 0 \cdots & 0 & c_{m_1} & \cdots & c_0 & c_1 & \cdots & c_{m_2} \; 0 \cdots \end{array}\right)$$

A is the middle part of this matrix with the "wings" mapped back inside

$A =$ 

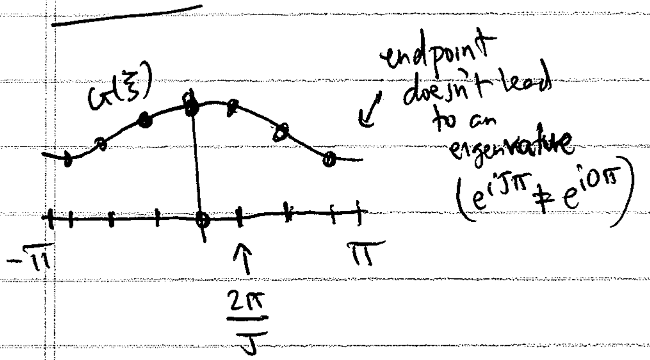
$$\left.\begin{array}{l} Bw = G(\xi)w \\ w_{j+J} = w_j \end{array}\right\} \Rightarrow \quad Ax = G(\xi)x$$

$w_j = e^{ij\xi} \qquad -\infty < j < \infty \qquad w \in \ell^2$

$x_j = w_j \qquad 0 \le j \le J-1 \qquad x \in \mathbb{C}^J$

result (periodic b.c.'s):

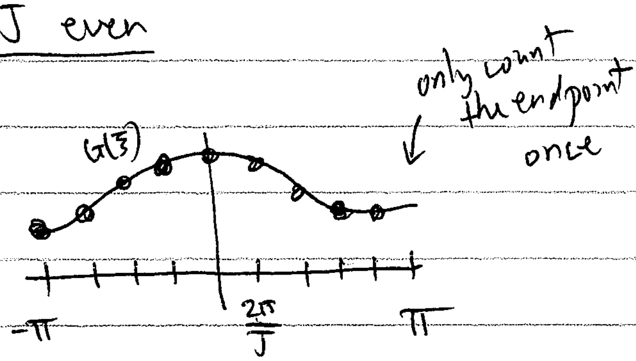
J odd



$$\lambda_\ell = G\left(\frac{2\pi \ell}{J}\right)$$

$$-\frac{J-1}{2} \le \ell \le \frac{J-1}{2}$$

J even



$$\lambda_\ell = G\left(\frac{2\pi \ell}{J}\right)$$

$$-\frac{J}{2} \le \ell \le \frac{J}{2} - 1$$

$G(\xi)$ is allowed to take on complex values in both plots.

Implicit schemes

the timestep restriction $\nu = \frac{k}{h^2} \le \frac{1}{2}$ makes the schemes we have studied so far rather impractical. Let's see what happens if we try

$$D_t^+ u_j^n = D_x^+ D_x^- u_j^{n+1} \longleftarrow \text{space derivatives done at } t_{n+1}$$

or

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

or

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

Math 228B Lec 9

Last time: amplification factors for arbitrary schemes
eigenvalues of finite dimensional versions of B in terms of $G(\xi)$

Today: implicit methods
higher dimensions
~~ADI (alternating direction implicit)~~

implicit schemes

the timestep restriction $\nu = \frac{k}{h^2} \leq \frac{1}{2}$ makes the schemes we have studied so far rather impractical. Let's see what happens if we try

$$D_t^+ u_j^n = D_x^+ D_x^- u_j^{n+1} \leftarrow \text{space derivative done at } t_{n+1}$$

or

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

or

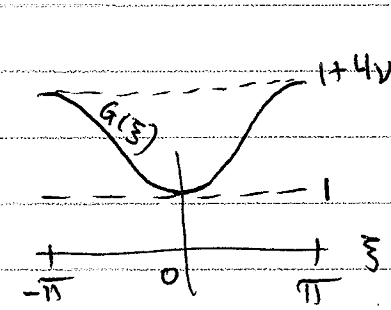
$$-\nu u_{j+1}^{n+1} + (1+2\nu) u_j^{n+1} - \nu u_{j-1}^{n+1} = u_j^n$$

we can write this as $\quad B u^{n+1} = u^n$

where $\quad B u_j = -\nu u_{j+1} + (1+2\nu) u_j - \nu u_{j-1}$

The amplification factor for B is

$$\begin{aligned} G(\xi) &= -\nu e^{i\xi} + (1+2\nu) - \nu e^{-i\xi} \\ &= 1 + 2\nu(1-\cos\xi) \\ &= 1 + 4\nu \sin^2(\xi/2) \end{aligned}$$



Since $G(\xi) \neq 0$ for $-\pi \le \xi \le \pi$,
the operator $\mathcal{G} : L^2(-\pi,\pi) \to L^2(-\pi,\pi)$ is invertible:

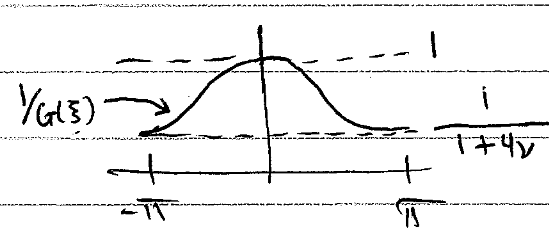
$\exists\, \mathcal{G}^{-1}$ s.t. $\mathcal{G}\mathcal{G}^{-1} f = f$, $\mathcal{G}^{-1}\mathcal{G} f = f$

$$\mathcal{G} f(\xi) = G(\xi) f(\xi)$$

$$\mathcal{G}^{-1} f(\xi) = \frac{1}{G(\xi)} f(\xi)$$

we know how to compute norms of multiplication operators already:

$$\|\mathcal{G}^{-1}\|_{L^2} = \left\|\frac{1}{G}\right\|_\infty = \max_{-\pi \le \xi \le \pi} \frac{1}{|G(\xi)|} = 1$$



and since $B = \mathcal{F}^{-1} \mathcal{G} \mathcal{F}$

we have $B^{-1} = \mathcal{F}^{-1} \mathcal{G}^{-1} \mathcal{F}$

so B is invertible and $\|B^{-1}\|_{2,h} = \|\mathcal{G}^{-1}\|_{L^2} = 1$

no matter what ν is. $\therefore$ $\boxed{u^{n+1} = B^{-1} u^n \text{ is unconditionally stable}}$

this allows us to choose a much more reasonable refinement path, e.g.

$$k = h \qquad \left(\text{or } \nu = \frac{1}{h} \text{ instead of a constant}\right)$$

remember, the requirement for stability was that

$$\exists K, \varepsilon \text{ s.t. } \|B(k)^n\| \le K \quad \text{for } \left\{\begin{matrix} 0 < k < \varepsilon \\ 0 \le nk \le T \end{matrix}\right\}$$

in our case $\varepsilon = 1$ and $K = 1$ works.

---

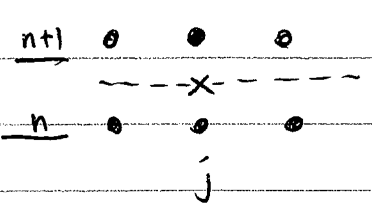
problem: our truncation error τ_j^n is still $O(k + h^2)$. This was fine when k was νh^2, but now it's unacceptable.

↑ fixed constant

<u>solution:</u> Crank-Nicolson scheme

$$D_t^+ u_j^n = \frac{1}{2}\left[D_x^+ D_x^- u_j^n + D_x^+ D_x^- u_j^{n+1} \right]$$

<u>stencil</u>



← approximate u_{xx} at midpoint $(x_j, t_n + \frac{k}{2})$

claim: $\tau_j^n = O(k^2 + h^2)$

proof: plug in the exact sol'n and do a Taylor expansion around the point $(x_j, t_{n+\frac{1}{2}})$

$$(*) \quad D_t^+ u_j^n = \frac{u(jh, nk+k) - u(jh, nk)}{k}$$

$$= \frac{\left[u + \frac{k}{2}u_t + \frac{1}{2}\left(\frac{k}{2}\right)^2 u_{tt} + \frac{1}{6}\left(\frac{k}{2}\right)^3 u_{ttt} + \cdots\right] - \left[u - \frac{k}{2}u_t + \frac{1}{2}\left(\frac{k}{2}\right)^2 u_{tt} - \frac{1}{6}\left(\frac{k}{2}\right)^3 u_{ttt} + \cdots\right]}{k}$$

$$= u_t(jh, nk+\tfrac{k}{2}) + \frac{k^2}{24} u_{ttt}(jh, nk+\tfrac{k}{2}) + O(k^4)$$

$$D_x^+ D_x^- u_j^n = u_{xx}(jh, nk) + \frac{h^2}{12} u_{xxxx}(jh, nk) + O(h^4)$$

$$D_x^+ D_x^- u_j^{n+1} = u_{xx}(jh, nk+k) + \frac{h^2}{12} u_{xxxx}(jh, nk+k) + O(h^4)$$

$$(**) \quad \frac{1}{2}(\cdot + \cdot) = u_{xx}(jh, nk+\tfrac{k}{2}) + \frac{k^2}{8} u_{xxtt}(jh, nk+\tfrac{k}{2}) + O(k^4)$$

$$+ \frac{h^2}{12}\left[u_{xxxx}(jh, nk+\tfrac{k}{2}) + \frac{k^2}{8} u_{xxxxtt}(jh, nk+\tfrac{k}{2}) + O(k^4)\right]$$

$$+ O(h^4)$$

conclusion: $\tau_j^n = (*) - (**)$

$$= \underbrace{(u_t - u_{xx})}_{0} + \frac{k^2}{24} u_{ttt} - \frac{h^2}{8} u_{xxtt}$$

$$- \frac{h^2}{12} u_{xxxx} + O(k^4 + h^4)$$

claim: Crank-Nicolson is unconditionally stable.

Proof: $D_t^+ u_j^n = \frac{1}{2}\left[D_x^+ D_x^- u_j^n + D_x^+ D_x^- u_j^{n+1} \right]$

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{1}{2h^2}\left[u_{j+1}^n - 2u_j^n + u_{j-1}^n + u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1} \right]$$

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

$$\circledast \quad \left(I - \frac{\nu}{2}B\right) u^{n+1} = \left(I + \frac{\nu}{2}B\right)u^n, \quad Bu_j = u_{j+1} - 2u_j + u_{j-1}$$

The amplification factor of B is
$$G(\xi) = e^{i\xi} - 2 + e^{-i\xi} = -2(1-\cos\xi) = -4\sin^2(\xi/2)$$

The mapping of an operator to its amplification factor is linear $\left(\mathcal{G} = \mathcal{F}B\mathcal{F}^{-1} \Rightarrow \alpha \mathcal{G}_1 + \beta \mathcal{G}_2 = \mathcal{F}(\alpha B_1 + \beta B_2)\mathcal{F}^{-1} \right)$ so if we take the Fourier transform of $\circledast$ we get

$$\left[1 - \frac{\nu}{2}\left(-4\sin^2\left(\frac{\xi}{2}\right)\right)\right] \hat{u}^{n+1}(\xi) = \left[1 + \frac{\nu}{2}\left(-4\sin^2\left(\frac{\xi}{2}\right)\right)\right]\hat{u}^n(\xi)$$

or
$$\hat{u}^{n+1}(\xi) = \underbrace{\frac{1 - 2\nu\sin^2(\xi/2)}{1 + 2\nu\sin^2(\xi/2)}}_{G_1(\xi)} \hat{u}^n(\xi)$$

for any choice of ξ, $G_1(\xi)$ has the form $\frac{1-a}{1+a}$ for some $a > 0$.

But $\left|\frac{1-a}{1+a}\right| = \sqrt{\frac{1-2a+a^2}{1+2a+a^2}} \le 1$ since $0 \le 1 - 2a + a^2 < 1 + 2a + a^2$

$\therefore \|B_1\|_{2,h} = \|\mathcal{G}_1\|_{L^2} = \|G_1\|_\infty \le 1$ where $u^{n+1} = B_1 u^n = (I - \frac{\nu}{2}B)^{-1}(I + \frac{\nu}{2}B)u^n$

The finite dimensional version of Crank-Nicoloson works the same, but instead of diagonalizing the operator with Fourier series, we use the discrete sine, cosine or Fourier transform

---

Dirichlet B.C.'s : $B = \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & 1 & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{pmatrix}$, $u^{n+1} = (I - \frac{\nu}{2}B)^{-1}(I + \frac{\nu}{2}B)u^n$

$$\begin{tikzcd} \mathbb{R}^{J-1} \arrow[r, "S"] \arrow[d, "B_1"'] & \mathbb{R}^{J-1} \arrow[d, "\Lambda"] \\ \mathbb{R}^{J-1} \arrow[r, "S"] & \mathbb{R}^{J-1} \end{tikzcd}$$

$\boxed{B_1 = S^{-1}\Lambda S}$, $\Lambda_{\ell\ell} = G_1\left(\frac{\pi \ell}{J}\right)$, $1 \le \ell \le J-1$

$(S^{-1})_{j\ell} = \sqrt{2} \sin \frac{j\pi\ell}{J}$ ← columns are eigenvectors of B_1

Neumann B.C.'s $B = \begin{pmatrix} -2 & 2 & & & \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & 1 \\ & & 1 & -2 & 1 \\ & & & 2 & -2 \end{pmatrix}$, $u^{n+1} = \underbrace{(I - \frac{\nu}{2}B)^{-1}(I + \frac{\nu}{2}B)}_{B_1}u^n$

$$\begin{tikzcd} \mathbb{R}^{J+1} \arrow[r, "C"] \arrow[d, "B_1"'] & \mathbb{R}^{J+1} \arrow[d, "\Lambda"] \\ \mathbb{R}^{J+1} \arrow[r, "C"] & \mathbb{R}^{J+1} \end{tikzcd}$$

$\boxed{B_1 = C^{-1}\Lambda C}$, $\Lambda_{\ell\ell} = G_1\left(\frac{\pi \ell}{J}\right)$, $0 \le \ell \le J$

$(C^{-1})_{j\ell} = c_\ell \cos \frac{j\pi\ell}{J}$, $c_\ell = \begin{cases} 1 & \ell = 0, J \\ \sqrt{2} & 1 \le \ell \le J-1 \end{cases}$

Periodic B.C.'s $B = \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & & \ddots & 1 \\ 1 & & & 1 & -2 \end{pmatrix}$ $u^{n+1} = \underbrace{(I - \frac{\nu}{2}B)^{-1}(I + \frac{\nu}{2}B)}_{B_1}u^n$

$$\begin{tikzcd} \mathbb{R}^{J} \arrow[r, "\mathcal{F}"] \arrow[d, "B_1"'] & \mathbb{R}^{J} \arrow[d, "\Lambda"] \\ \mathbb{R}^{J} \arrow[r, "\mathcal{F}"] & \mathbb{R}^{J} \end{tikzcd}$$

$\boxed{B_1 = \mathcal{F}^{-1}\Lambda \mathcal{F}}$, $\Lambda_{\ell\ell} = G_1\left(\frac{2\pi\ell}{J}\right)$ $0 \le \ell \le J-1$

$(\mathcal{F}^{-1})_{j\ell} = e^{\frac{2\pi i j \ell}{J}}$ $\mathcal{F}_{mj} = \frac{1}{J} e^{-\frac{2\pi i m j}{J}}$

---

all three operators $S, C, \mathcal{F}$ are isometries up to a constant factor, so B_1 and Λ have the same norm.

In particular, $I - \frac{\nu}{2}B$ is invertible since its eigenvalues are ≥ 1 (and therefore not zero)

To implement these schemes, you can either use the appropriate transform (fast sine, fast cosine, or fast fourier transform) and then iterate by multiplying by the diagonal matrix Λ, or you can solve a tridiagonal system.

$$u^{n+\frac{1}{2}} = \left(I + \tfrac{\nu}{2}B\right)u^n \quad \leftarrow \text{explicit half-step (easy)}$$

$$u^{n+1} = \left(I - \tfrac{\nu}{2}B\right)^{-1}u^{n+\frac{1}{2}} \quad \leftarrow \text{implicit half-step (solve tridiag. system)}$$

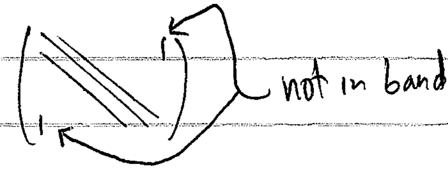
tridiagonal systems can be solved in $O(N)$ time $\left\{ \begin{array}{l} Ax = b \\ \uparrow \text{ N×N matrix} \\ (N \text{ is } J-1, J+1, \text{ or } J) \end{array}\right.$

It's a mistake to form the inverse, though, since A^{-1} is a dense matrix, so applying A^{-1} to b by matrix multiplication requires $O(N^2)$ flops

The LU factorization of a banded matrix is banded even if you use pivoting (but the band grows a little)

LAPACK: $\underbrace{\text{DGBTRF}}_{\text{factor}}$, $\underbrace{\text{DGBTRS}}_{\text{solve}}$ ← banded systems
(C, C++, Fortran)

This still won't handle the periodic case



<u>option 1</u>: don't pivot the last row in
(end up with arrow shaped matrices: $L = \begin{pmatrix} \diagdown & 0 \\ \underline{0} & \diagdown \end{pmatrix}$, $U = \begin{pmatrix} \diagdown & 0 \,| \\ 0 & \diagdown \end{pmatrix}$)
have to be mildly careful with numerical stability of factorizations, but for discretizations of PDE's it's often OK not to pivot (diagonal dominance)

<u>option 2</u>: use a sparse solver (e.g. colamd, symamd)
matlab has these solvers built in, so just designate your matrix $A = (I - \frac{\nu}{2}B)$ as <u>sparse</u> and solve with backslash: $x = A\backslash b$

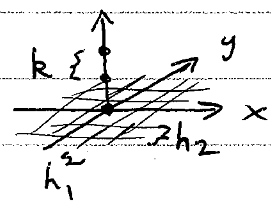
Higher dimensions $\quad u_t = \Delta u \quad$ (or $\nabla^2 u$ if you prefer)

in 2-d: $\quad u_t = u_{xx} + u_{yy}$

exact solution: $\quad u(x,y,t) = \dfrac{1}{4\pi t} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2 + (y-\eta)^2}{4t}} g(\xi,\eta)\, d\xi\, d\eta$

not square rooted now

numerics:



2-d discrete grids attached to each time t_n

we'll often write u_j^n but j means (j_1, j_2) now

discrete 2-norm:

$$\|u\|_{2,h}^2 = h_1 h_2 \sum_j |u_j|^2 \qquad \left(= h_1 h_2 \sum_{j=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} |u_{j,\ell}|^2 \right)$$

explicit scheme:

$$D_t^+ u_j^{n+1} = D_x^+ D_x^- u_j^n + D_y^+ D_y^- u_j^n$$

$$u_{j\ell}^{n+1} = \nu_1 u_{j-1,\ell}^n + (1 - 2\nu_1 - 2\nu_2) u_{j\ell}^n + \nu_1 u_{j+1,\ell}^n + \nu_2 u_{j,\ell+1}^n + \nu_2 u_{j,\ell-1}^n$$

$$\nu_1 = \frac{k}{h_1^2} \qquad \nu_2 = \frac{k}{h_2^2}$$

or $\quad u^{n+1} = B u^n$

it's easy to show that $\|B\|_\infty = 1$ and $\|B\|_{1,h} = 1$

iff $\nu_1 + \nu_2 \le \frac{1}{2}$

228B Lecture 10

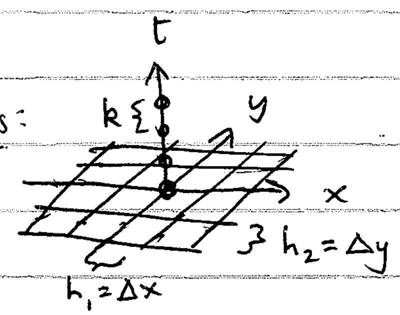
<u>Last time</u>: implicit methods
Crank-Nicolson ($O(k^2+h^2)$ and unconditional stability!)

<u>Today</u>: higher dimensions, ADI (alternating direction implicit)

<u>higher dimensions</u>: $u_t = \Delta u$ (or $u_t = \nabla^2 u$)

in 2-d: $u_t = u_{xx} + u_{yy}$, $u(x,y,0) = g(x,y)$

exact soln: $$u(x,y,t) = \frac{1}{4\pi t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2+(y-\eta)^2}{4t}} g(\xi,\eta)\, d\xi\, d\eta$$

<u>numerics</u>: 

2-d discrete grids attached to each time t_n

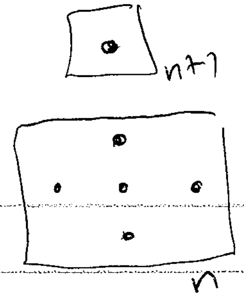
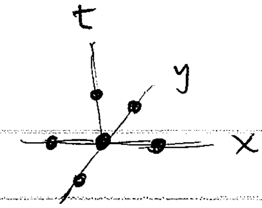
discrete 2-norm: $$\|u\|_{2,h}^2 = h_1 h_2 \sum_{j,l=-\infty}^{\infty} |u_{jl}|^2$$

explicit scheme: $D_t^+ u_{jl}^{n+1} = D_x^+ D_x^- u_{jl}^n + D_y^+ D_y^- u_{jl}^n$

$$u_{jl}^{n+1} = \nu_1 u_{j-1,l}^n + (1-2\nu_1-2\nu_2)u_{jl}^n + \nu_1 u_{j+1,l}^n + \nu_2 u_{j,l+1}^n + \nu_2 u_{j,l-1}^n$$

$\nu_1 = \frac{k}{h_1^2}$, $\nu_2 = \frac{k}{h_2^2}$

or $u^{n+1} = Bu^n$

The stencil for B looks like  or 

it's easy to check that $\|B\|_\infty = 1$ and $\|B\|_{1,h} = 1$

as long as $\nu_1 + \nu_2 \le \frac{1}{2}$

and if $\nu_1 + \nu_2 > \frac{1}{2}$ then $\|B\|_\infty > 1$ and $\|B\|_{1,h} > 1$

(same idea as 1d case. counterexample for $\|\cdot\|_\infty$:
$$\begin{matrix} 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 \end{matrix}$$
gets mapped to itself times $1-4(\nu_1+\nu_2)$

counterexample for $\|\cdot\|_1$:
$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix}$$
(gets mapped to
$$\begin{matrix} 0 & \nu_2 & 0 \\ \nu_1 & 1-2\nu_1-2\nu_2 & \nu_1 \\ 0 & \nu_2 & 0 \end{matrix}$$

if $\nu_1 + \nu_2 > \frac{1}{2}$ then

$$h_1 h_2 \left[2|\nu_1| + 2|\nu_2| + |1-2\nu_1-2\nu_2| \right]$$
$$= h_1 h_2 \left[4(\nu_1+\nu_2) - 1 \right] > h_1 h_2 [1]$$

$$\|Bu\|_{1,h} > 1 \, \|u\|_{1,h} \longleftarrow \text{with } u_{j\ell} = \begin{cases} 1 & j=0, \ell=0 \\ 0 & \text{o.w.} \end{cases}$$

To analyze the 2-norm, we use a higher dimensional Fourier series

$\ell^2(\mathbb{Z}\times\mathbb{Z})$ square summable doubly-indexed sequences $\qquad$ $L^2([-\pi,\pi]\times[-\pi,\pi])$ square integrable functions

$$\{u_{j\ell}\}_{j,\ell=-\infty}^{\infty} \xrightarrow{\mathcal{F}} \hat{u}(\xi,\eta) = \sum_{j,\ell} u_{j,\ell} e^{-i(j\xi+\ell\eta)}$$

$$f^\vee_{j\ell} = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\xi,\eta) e^{i(j\xi+\ell\eta)} \, d\xi \, d\eta \xleftarrow{\mathcal{F}^{-1}} f(\xi,\eta)$$

In 2-d, Parseval's identity is

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |\hat{u}(\xi,\eta)|^2\, d\xi\, d\eta = \frac{4\pi^2}{h_1 h_2}\left(h_1 h_2 \sum_{j\ell} |u_{j\ell}|^2\right) \quad \leftarrow \|\mathcal{F}\| = \frac{2\pi}{\sqrt{h_1 h_2}}$$

$$h_1 h_2 \sum_{j\ell} |\check{f}_{j\ell}|^2 = \frac{h_1 h_2}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |f(\xi,\eta)|^2\, d\xi\, d\eta \quad \leftarrow \|\mathcal{F}^{-1}\| = \frac{\sqrt{h_1 h_2}}{2\pi}$$

so $\mathcal{F}$ and $\mathcal{F}^{-1}$ are again isometries up to a scale factor and $\|\mathcal{F}\|\cdot\|\mathcal{F}^{-1}\| = 1$

Amplification factors can be computed just as before

$$\mathcal{B} = Bu_{j\ell} = \sum_{p,q} c_{pq} u_{j+p,\ell+q} \qquad \left(\sum_{p,q} = \sum_{p=p_1}^{p_2}\sum_{q=q_1}^{q_2} \text{ a compact stencil}\right)$$

coefficients of the stencil

$$\widehat{Bu}(\xi,\eta) = \sum_{j,\ell}\left(\sum_{p,q} c_{pq} u_{j+p,\ell+q}\right) e^{-i(j\xi+\ell\eta)}$$

$$= \sum_{p,q}\sum_{r,s} c_{pq} u_{rs} e^{-i((r-p)\xi+(s-q)\eta)}$$

$$= \underbrace{\sum_{p,q} c_{pq} e^{i(p\xi+q\eta)}}_{G(\xi,\eta)} \underbrace{\sum_{r,s} u_{rs} e^{-ir\xi} e^{-is\eta}}_{\hat{u}(\xi,\eta)}$$

Let's split our explicit scheme $u^{n+1} = Bu^n$ into smaller pieces:

$$B = I + \nu_1 B_1 + \nu_2 B_2 ,$$

$$B_1 u_{j\ell} = u_{j-1,\ell} - 2u_{j\ell} + u_{j+1,\ell} \rightarrow G_1(\xi,\eta) = e^{-i\xi} - 2 + e^{i\xi}$$
$$= -4\sin^2(\xi/2)$$

$$B_2 u_{j\ell} = u_{j,\ell-1} - 2u_{j\ell} + u_{j,\ell+1} \rightarrow G_2(\xi,\eta) = e^{-i\eta} - 2 + e^{i\eta}$$
$$= -4\sin^2(\eta/2)$$

so $\quad G(\xi,\eta) = 1 - 4\nu_1 \sin^2(\xi/2) - 4\nu_2 \sin^2(\eta/2)$

worst case: $\xi = \pi$, $\eta = \pi$, $\quad G(\pi,\pi) = 1 - 4(\nu_1 + \nu_2)$

So $$\|G\|_\infty = \max_{\substack{-\pi \le \xi \le \pi \\ -\pi \le \eta \le \pi}} |G(\xi,\eta)| = \begin{cases} 1 & \nu_1 + \nu_2 \le \frac{1}{2} \\ > 1 & \nu_1 + \nu_2 > \frac{1}{2} \end{cases}$$

---

we can also do implicit methods in 2-d. $\quad (u^{n+1} = Bu^n)$

$$(I - \nu_1 B_1 - \nu_2 B_2) u^{n+1} = u^n \qquad \leftarrow \boxed{\text{Backward Euler}}$$

$$B = (\quad)^{-1}, \quad G(\xi,\eta) = \frac{1}{1 + 4\nu_1 \sin^2(\xi/2) + 4\nu_2 \sin^2(\eta/2)}$$

$\|G\|_\infty \le 1$ unconditionally stable, but $O(k + h_1^2 + h_2^2)$

$$\left(I - \frac{\nu_1}{2} B_1 - \frac{\nu_2}{2} B_2\right) u^{n+1} = \left(I + \frac{\nu_1}{2} B_1 + \frac{\nu_2}{2} B_2\right) u^n \qquad \boxed{\text{Crank-Nicolson}}$$

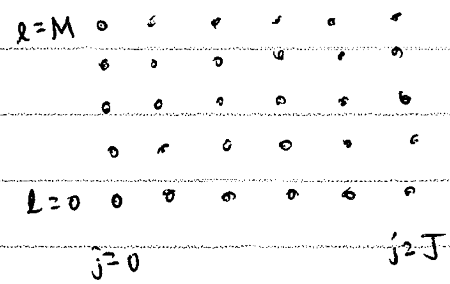
$$B = (\quad)^{-1}(\quad)$$

unconditionally stable and $O(k^2 + h_1^2 + h_2^2)$

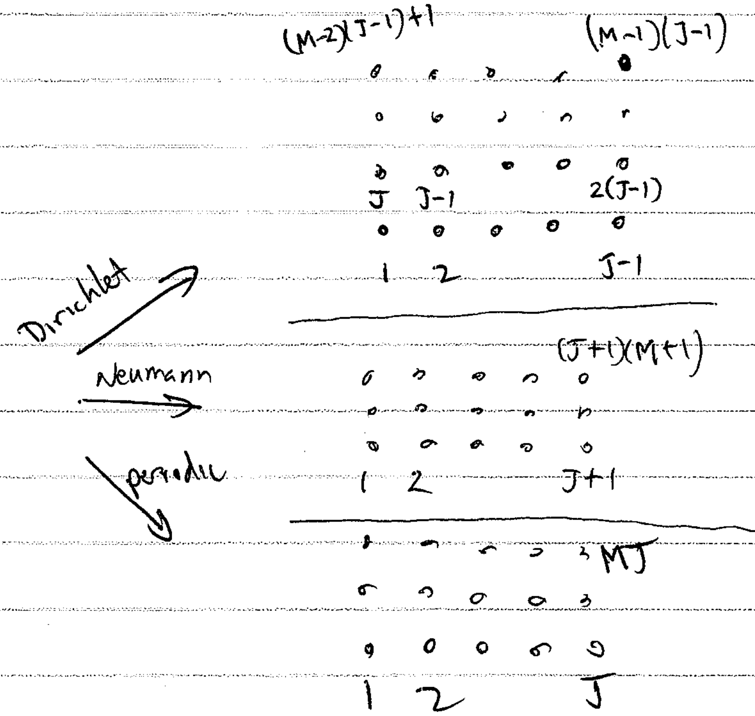
$$G(\xi,\eta) = \frac{1-a}{1+a}, \quad a = 2\nu_1 \sin^2(\xi/2) + 2\nu_2 \sin^2(\eta/2)$$

The problem is that these matrices are not tightly banded.

grid:



Natural numbering of nodes:



matrix representation of

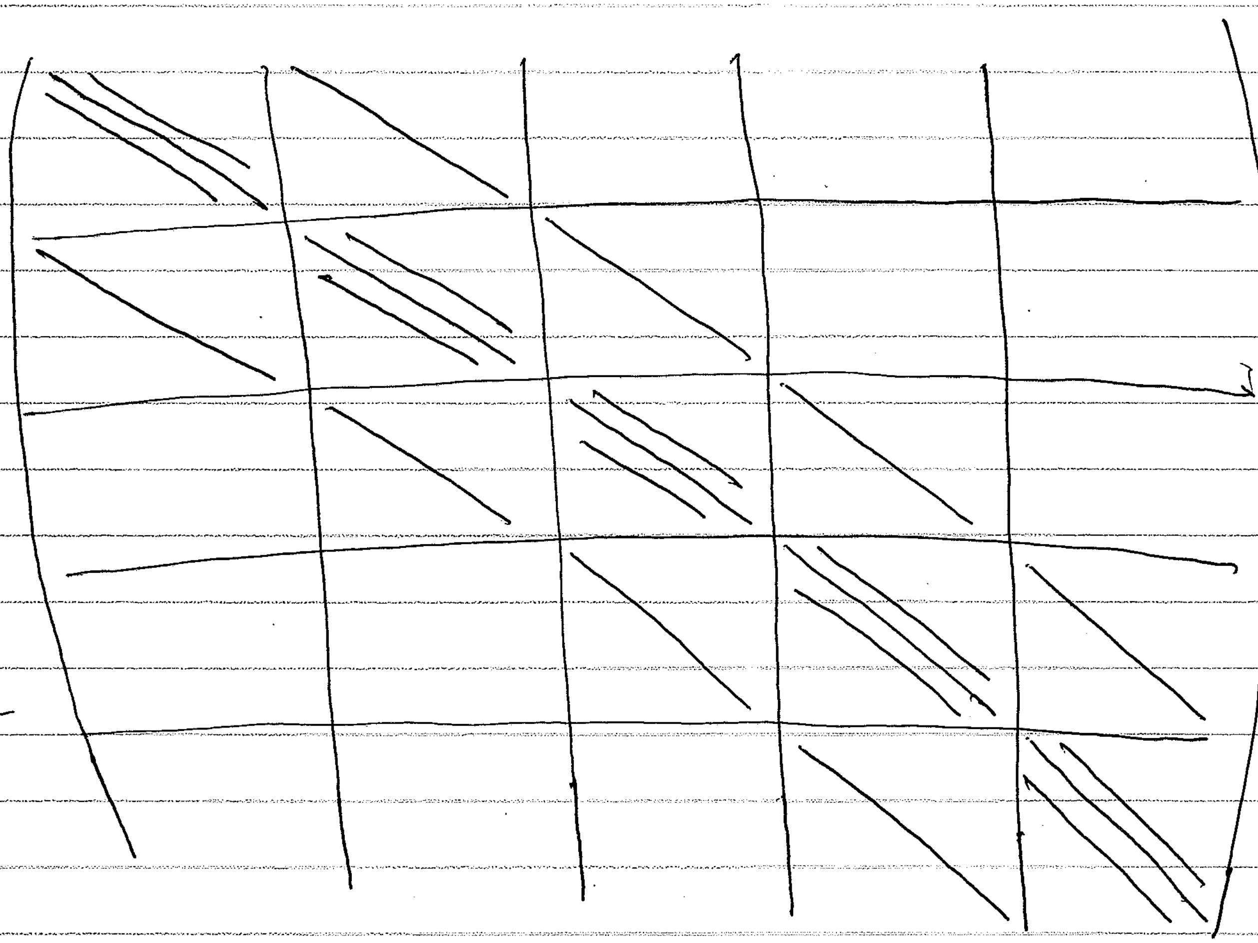
$$I - \frac{\nu_1}{2} B_1 - \frac{\nu_2}{2} B_2$$

looks like (in Dirichlet case):

sub-blocks are $(J-1)\times(J-1)$

the big block matrix is $(M-1)\times(M-1)$

bandwidth is $\approx J$



There are very effective numerical methods for solving linear system like this (multigrid, fast sine transform) but today will talk about an approach known as

ADI (alternating direction implicit)

it's also frequently referred to as "operator splitting"

<u>ADI scheme:</u>

$$\left(I - \frac{\nu_1}{2}B_1\right)\left(I - \frac{\nu_2}{2}B_2\right)u^{n+1} = \left(I + \frac{\nu_1}{2}B_1\right)\left(I + \frac{\nu_2}{2}B_2\right)u^n$$

multiply it out:

$$\left(I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}B_1B_2\right)u^{n+1} = \left(I + \frac{\nu_1}{2}B_1 + \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}B_1B_2\right)u^n$$

(if these terms weren't present, this would be the Crank-Nicolson scheme)

truncation error:

$$\tau^n_{ADI} = \frac{1}{k}\left[\left(I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}B_1B_2\right)u^{n+1} - \left(I + \cdots\right)u^n\right]$$

$$= \tau^n_{C.N.} + \frac{\nu_1\nu_2}{4}B_1B_2\left(\frac{u^{n+1}-u^n}{k}\right)$$

$$= \tau^n_{C.N.} + \frac{k^2}{4}D_x^+D_x^-D_y^+D_y^-D_t^+u^n$$

$$= \frac{k^2}{24}u_{ttt} - \frac{k^2}{8}u_{xxtt} - \frac{k^2}{8}u_{yytt} - \frac{h_1^2}{12}u_{xxxx} - \frac{h_2^2}{12}u_{yyyy} + \frac{k^2}{4}u_{xxyyt}$$

$$= -\frac{1}{12}\left(k^2u_{ttt} + h_1^2u_{xxxx} + h_2^2u_{yyyy} - 3k^2u_{xxyyt}\right) + O(k^4 + h_1^4 + h_2^4)$$

so the additional terms don't do any essential harm.

But now the linear systems we have to solve are tri-diagonal 1d systems

$$u^{n+\frac{1}{2}} = \left(I + \frac{\nu_1}{2}B_1\right)\overbrace{\left(I + \frac{\nu_2}{2}B_2\right)u^n}^{u^{n+1/4}} \quad \leftarrow \text{two explicit steps}$$

$$u^{n+3/4} = \left(I - \frac{\nu_1}{2}B_1\right)^{-1}u^{n+1/2} \quad \leftarrow \text{a bunch of 1d tridiagonal systems in the x-direction}$$

$$x^{n+1} = \left(I - \frac{\nu_2}{2}B_2\right)^{-1}u^{n+\frac{3}{4}} \quad \leftarrow \text{same story in y-direction}$$

The four operations can be done in any order since B_1 and B_2 commute

<u>stencils:</u>

$$u: \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$B_1u: \begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad B_2u: \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$B_2B_1u \quad \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \quad B_1B_2u$$

A stencil is like a column of a matrix. it tells you what the operator does to an elementary unit vector.

1d: $\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \leftarrow$ jth slot

2d:
$$\begin{array}{cc} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} & \begin{array}{c} \\ \leftarrow l \\ \\ \\ \end{array} \\ \begin{array}{cccc} & \uparrow & & \\ & j & & \end{array} & \end{array}$$

$$AB = BA \Rightarrow (I+A)B = B(I+A)$$

$$AB = BA \Rightarrow BA^{-1} = A^{-1}B$$

conclusion: $\left(I - \frac{\nu_1}{2}B_1\right)^{-1}, \left(I - \frac{\nu_2}{2}B_2\right)^{-1}$

$$\left(I + \frac{\nu_1}{2}B_1\right), \left(I + \frac{\nu_2}{2}B_2\right)$$

all commute with each other.

Our book distinguishes between the various orders of applying the 1-d operators due to difficulties with non-zero boundary conditions. This makes no sense to me. I'll explain the "right way" to deal with b.c.'s next time.

---

remark about truncation errors:

our schemes today were all of the form $Au^{n+1} = Bu^n$

I defined $\tau^n = \frac{1}{k}\left[Au^{n+1} - Bu^n\right]$.

To fit in the Lax-Richtmyer convergence proof setup, we really should define

$$\tau^n = \frac{1}{k}\left[u^{n+1} - A^{-1}Bu^n\right]$$

so $\tau^n_{\text{correct def}} = A^{-1}\,\tau^n_{\text{what I did}}$

and $$\|\tau^n_{\text{correct def}}\| \le \underbrace{\|A^{-1}\|}_{\substack{1 \text{ in all the cases of} \\ \text{interest so far}}} \cdot \|\tau^n_{\text{what I did}}\|$$

$$= O(k^2 + h_1^2 + h_2^2)$$

so it's fine to work with the definitions I used.

228B Lec 11

Last time: ① philosophy of Crank-Nicolson (discretize space first, get an ODE (e.g. $u_t = D_x^+ D_x^- u$), use your favorite scheme for stiff equations to solve the ODE (e.g. the trapezoidal rule)

② 2-d heat equation

③ 1-norm, ∞-norm analysis almost identical to the 1d case

④ 2-norm requires 2d Fourier analysis & amplification factors

⑤ Crank-Nicolson still gives an $O(k^2+h^2)$ unconditionally stable method, but the matrix you have to invert is not tightly banded

Today ① discussion of truncation errors for implicit methods

② ADI methods

③ non-zero heat source

④ nonzero boundary conditions

---

truncation errors: our schemes have always been of the form $Au^{n+1} = Bu^n$.

when talking about truncation errors for the Backward-Euler and Crank-Nicolson methods, we defined

$$\tau^n = \frac{1}{k}\left[Au^{n+1} - Bu^n\right].$$

To fit in the Lax-Richtmyer framework, we really should use

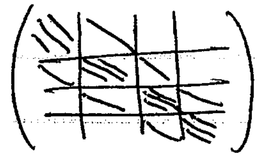
$$\tau^n = \frac{1}{k}\left[u^{n+1} - A^{-1}Bu^n\right]$$

so $\tau^n_{correct} = A^{-1}\tau^n_{convenient}$ and $\|\tau^n_{correct}\| \le \underbrace{\|A^{-1}\|}_{\substack{\text{often equal to 1} \\ \text{but actually any constant} \\ \text{is fine here.}}} \cdot \|\tau^n_{convenient}\|$, where $\|\tau^n_{convenient}\|$ is e.g.: $O(k^2+h_1^2+h_2^2)$

Any time A^{-1} is bounded, it's fine to work with the more convenient definition.

Crank-Nicolson in 2-d: $\overbrace{\left(I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2\right)}^{A} u^{n+1} = \overbrace{\left(I + \frac{\nu_1}{2}B_1 + \frac{\nu_2}{2}B_2\right)}^{B} u^n$

$$B_1 u_{j,\ell} = u_{j-1,\ell} - 2u_{j\ell} + u_{j+1,\ell}$$
$$B_2 u_{j\ell} = u_{j,\ell-1} - 2u_{j\ell} + u_{j,\ell+1}$$



sparsity structure of A

pros: $\tau^n = O(k^2 + h_1^2 + h_2^2)$, unconditionally stable

con: A is not tightly banded, Expensive to solve using Gaussian elimination

ADI scheme:

$$\left(I - \frac{\nu_1}{2}B_1\right)\left(I - \frac{\nu_2}{2}B_2\right)u^{n+1} = \left(I + \frac{\nu_1}{2}B_1\right)\left(I + \frac{\nu_2}{2}B_2\right)u^n$$

multiply it out:

$$\left(I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2 + \underbrace{\frac{\nu_1\nu_2}{4}B_1B_2}\right)u^{n+1} = \left(I + \frac{\nu_1}{2}B_1 + \frac{\nu_2}{2}B_2 + \underbrace{\frac{\nu_1\nu_2}{4}B_1B_2}\right)u^n$$

(if these terms weren't present, this would be the Crank-Nicolson scheme)

truncation error:

$$\tau^n_{ADI} = \frac{1}{k}\left[\left(I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}B_1B_2\right)u^{n+1} - \left(I + \cdots\right)u^n\right]$$

$$= \tau^n_{C.N.} + \frac{\nu_1\nu_2}{4}B_1B_2\left(\frac{u^{n+1}-u^n}{k}\right)$$

$$= \tau^n_{C.N.} + \frac{k^2}{4}D_x^+D_x^-D_y^+D_y^-D_t^+u^n$$

$$= \frac{k^2}{24}u_{ttt} - \frac{k^2}{8}u_{xxtt} - \frac{k^2}{8}u_{yytt} - \frac{h_1^2}{12}u_{xxxx} - \frac{h_2^2}{12}u_{yyyy} + \frac{k^2}{4}u_{xxyyt}$$

$$= -\frac{1}{12}\left(k^2 u_{ttt} + h_1^2 u_{xxxx} + h_2^2 u_{yyyy} - 3k^2 u_{xxyyt}\right) + O(k^4 + h_1^4 + h_2^4)$$

so the additional terms don't do any essential harm.

But now the linear systems we have to solve are tri-diagonal 1d systems

$$u^{n+\frac{1}{2}} = \left(I + \frac{\nu_1}{2} B_1\right) \overbrace{\left(I + \frac{\nu_2}{2} B_2\right) u^n}^{u^{n+1/4}}$$ ← two explicit steps

$$u^{n+3/4} = \left(I - \frac{\nu_1}{2} B_1\right)^{-1} u^{n+1/2}$$ ← a bunch of 1d tridiagonal systems in the x-direction

$$x^{n+1} = \left(I - \frac{\nu_2}{2} B_2\right)^{-1} u^{n+\frac{3}{4}}$$ ← same story in y-direction

The four operations can be done in any order since B_1 and B_2 commute

<u>stencils:</u>

$$u: \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$B_1 u: \begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad B_2 u: \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$B_2 B_1 u = B_1 B_2 u: \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

A stencil is like a column of a matrix. it tells you what the operator does to an elementary unit vector of the form

1d: $\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ ← ith slot

2d:
$$\begin{array}{cc} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} & \leftarrow l \\ \uparrow & \\ j & \end{array}$$

$$AB = BA \Rightarrow (I+A)B = B(I+A)$$

$$AB = BA \Rightarrow BA^{-1} = A^{-1}B$$

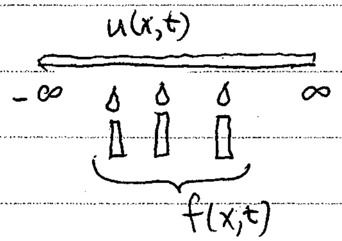
conclusion: $\left(I - \frac{\nu_1}{2} B_1\right)^{-1}$, $\left(I - \frac{\nu_2}{2} B_2\right)^{-1}$

$\left(I + \frac{\nu_1}{2} B_1\right)$, $\left(I + \frac{\nu_2}{2} B_2\right)$

all commute with each other. (only for constant coefficients on a rectangle, though)

Nonzero heat source

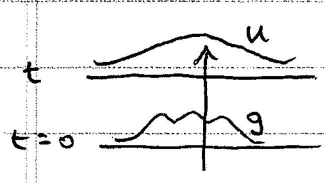
1d:



$$u_t - u_{xx} = f(x,t)$$
$$u(x,0) = g(x)$$

exact solution: Let $U(t)$ be the operator mapping an initial condition to the solution of $u_t = u_{xx}$ at time t:

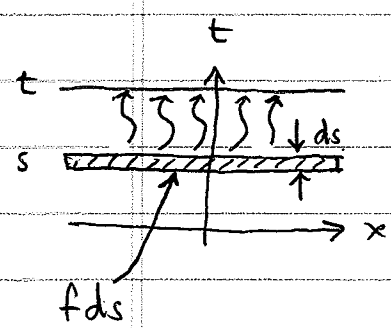
$u = U(t)g$ means $$u(x) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi$$



$U(t)$ maps g to u for the homogeneous problem (with $f=0$)

The solution of the inhomogeneous problem (with $f \neq 0$) is then

$$u(\cdot,t) = U(t)g + \int_0^t U(t-s) f(\cdot,s)\, ds \qquad (*)$$



physical interpretation: (superposition principle) each time slice $f\,ds$ propagates forward like an initial condition g for a time $t-s$

(This is an example of Duhamel's principle, where you build up solutions of an inhomogeneous problem using the representation for the homogeneous initial value problem

if $(*)$ is confusing, it just means

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi + \int_0^t \frac{1}{\sqrt{4\pi(t-s)}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4(t-s)}} f(\xi,s)\, d\xi\, ds$$

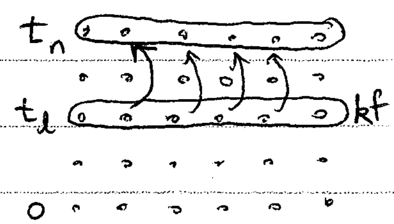
The numerical solution works the same way: $\quad f_j^n = f(jh, nk)$

	homogeneous IVP ($f=0$)	nonzero source
explicit method:	$u^{n+1} = Bu^n$	$u^{n+1} = Bu^n + kf^n$
fully implicit method:	$Au^{n+1} = u^n$	$Au^{n+1} = u^n + kf^{n+1}$
Crank-Nicolson:	$Au^{n+1} = Bu^n$	$Au^{n+1} = Bu^n + \frac{k}{2}[f^n + f^{n+1}]$

use the "discretize space first and choose your favorite ODE method" as your guide for where to evaluate f

e.g. C-N: $(u_j)_t = D^+D^- u_j + f_j \xrightarrow[\text{rule}]{\text{trap.}} u_j^{n+1} = u_j^n + k\left[\frac{D^+D^- u_j^n + f_j^n}{2} + \frac{D^+D^- u_j^{n+1} + f_j^{n+1}}{2}\right]$

The final solution is then a superposition:



explicit: $\quad u^n = B^n u^0 + k\sum_{\ell=0}^{n-1} B^{n-1-\ell} f^\ell$

implicit: $\quad u^n = A^{-n}u^0 + k\sum_{\ell=0}^{n-1} A^{-(n-\ell)} f^{\ell+1}$

C-N: $\quad u^n = (A^{-1}B)^n u^0 + k\sum_{\ell=0}^{n-1} (A^{-1}B)^{n-1-\ell} A^{-1}\left(\frac{f^\ell + f^{\ell+1}}{2}\right)$

kf is propagated forward by the scheme

This may be thought of as a discrete version of Duhamel's principle

The presense of f doesn't affect the error analysis (f is absorbed into τ^n)

~~example: (explicit scheme)~~ $\tau_j^n = \frac{1}{k}\left[\underbrace{u(jh,(n+1)k) - B[u(\cdot h, nk)]_j}_{\text{exact sol'n}} - f_j^n\right]$

numerical sol'n: $u_j^{n+1} = Bu_j^n + kf_j^n$

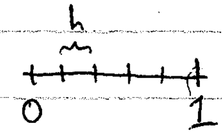
exact sol'n: $u(jh,(n+1)k) = B[u(\cdot h,(n+1)k]_j + kf_j^n + k\tau_j^n$

error: $e_j^{n+1} = Be_j^n - k\tau_j^n$ $\quad \boxed{K = \sup_{0\le nk\le T} \|B(k)^n\|}$

now proceed as before to conclude that $\max_{0\le nk\le T} \|e^n\| \le KT \max_{0\le nk\le T} \|\tau^n\|$

Nonzero boundary conditions

1d example: $\left\{ \begin{array}{l} u_t = u_{xx} \\ u(0,t) = \alpha(t) \\ u(1,t) = \beta(t) \\ u(x,0) = g(x) \end{array} \right\}$ α, β, g given, find u



Let $B : \mathbb{R}^{J+1} \to \mathbb{R}^{J-1}$ be given by

$$B = \begin{array}{cc} \begin{array}{ccccccc} \ell=0 & & & & & & \ell=J \end{array} & \\ \begin{pmatrix} 1 & -2 & 1 & 0 & & & 0 \\ 0 & 1 & -2 & 1 & & 0 & \vdots \\ & & \ddots & \ddots & \ddots & & \vdots \\ & 0 & & & & & 0 \\ & & & & 1 & -2 & 1 \end{pmatrix} & \begin{array}{c} j=1 \\ \\ \\ \\ j=J-1 \end{array} \end{array}$$

Again we let the ODE method in time guide us in where to evaluate α, β:

explicit: $\quad u^{n+1} = u^n + \nu B[\alpha^n; u^n; \beta^n] \qquad \alpha^n = \alpha(nk)$, $\beta^n = \beta(nk)$

fully implicit: $\quad u^{n+1} - \nu B[\alpha^{n+1}; u^{n+1}; \beta^{n+1}] = u^n$

C.N: $\quad u^{n+1} - \frac{\nu}{2} B[\alpha^{n+1}; u^{n+1}; \beta^{n+1}] = u^n + \frac{\nu}{2} B[\alpha^n; u^n; \beta^n]$

It's a little awkward to work with non-square matrices, so let's define

$$B = \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix} \quad \text{and} \quad \tilde{B} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \end{pmatrix}}_{\text{two columns}} \begin{array}{l} \leftarrow j=1 \\ \\ \\ \\ \leftarrow j=J-1 \end{array}$$

(indexed by 0 and J if you like)

If we move all the known stuff to the right hand side, we get:

explicit: $\quad u^{n+1} = (I + \nu B)u^n + \nu \tilde{B}[\alpha^n; \beta^n]$

implicit: $\quad (I - \nu B)u^{n+1} = u^n + \nu \tilde{B}[\alpha^{n+1}; \beta^{n+1}]$

C.N: $\quad (I - \frac{\nu}{2}B)u^{n+1} = (I + \frac{\nu}{2}B)u^n + \nu \tilde{B}\left[\frac{\alpha^{n+1}+\alpha^n}{2}; \frac{\beta^{n+1}+\beta^n}{2}\right]$

so the boundary data appear as source terms (like f_j above) attached to the nodes nearest the boundary (the rows where $\tilde{B}$ has nonzero entries)

$$\begin{cases} \text{for } (x,y) \in \partial\Omega \\ u(x,y,t) = \alpha(x,y,t) \end{cases}$$

$$\boxed{\begin{array}{c} u_t = \Delta u \\ u(x,y,0) = g(x,y) \end{array}}$$

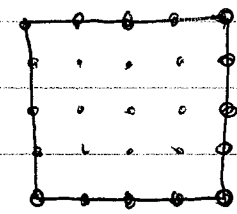
known (given) function $\alpha : \partial\Omega \to \mathbb{R}$

boundary notation

Nonzero b/c's in 2-d

this time instead of $\alpha(t), \beta(t)$ we have $\alpha(x,y,t)$ where x,y range over the boundary

when we discretize, stencils attached to the outer layer of unknowns will ask for boundary data, which we split off into a $\tilde{B}$ operator just as in the 1d case.



$B_1 : \mathbb{R}^9 \to \mathbb{R}^9$, $B_2 : \mathbb{R}^9 \to \mathbb{R}^9$

$\tilde{B}_1 : \mathbb{R}^{16} \to \mathbb{R}^9$, $\tilde{B}_2 : \mathbb{R}^{16} \to \mathbb{R}^9$

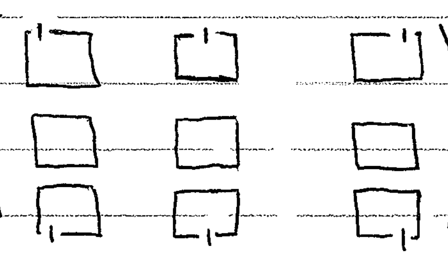
9 interior nodes

16 boundary nodes

for lack of better notation:

$$B_1 = \begin{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & 1 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & -2 & 1 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & -2 \end{pmatrix} \end{pmatrix}, \quad \tilde{B}_1 = \begin{pmatrix} \begin{pmatrix} 0&0&0&0&0 \\ 1&&&&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} & \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} & \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&1 \\ 0&&&&0 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} \\ \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 1&&&&0 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} & \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} & \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 0&&&&1 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} \\ \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 0&&&&0 \\ 1&&&&0 \\ 0&0&0&0&0 \end{pmatrix} & \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&0&0&0&0 \end{pmatrix} & \begin{pmatrix} 0&0&0&0&0 \\ 0&&&&0 \\ 0&&&&0 \\ 0&&&&1 \\ 0&0&0&0&0 \end{pmatrix} \end{pmatrix}$$

$$B_2 = \begin{pmatrix} \begin{pmatrix} -2 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 0 & 0 \\ -2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & -2 \\ 0 & 0 & 1 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -2 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -2 \end{pmatrix} \end{pmatrix}$$

$\tilde{B}_2 =$ 

schemes:

explicit: $\quad u^{n+1} = (I + \nu_1 B_1 + \nu_2 B_2) u^n + (\nu_1 \tilde{B}_1 + \nu_2 \tilde{B}_2) \alpha^n$

implicit: $\quad (I - \nu_1 B_1 - \nu_2 B_2) u^{n+1} = u^n + (\nu_1 \tilde{B}_1 + \nu_2 \tilde{B}_2) \alpha^{n+1}$

C-N: $\quad \left(I - \frac{\nu_1}{2} B_1 - \frac{\nu_2}{2} B_2\right) u^{n+1} = \left(I + \frac{\nu_1}{2} B_1 + \frac{\nu_2}{2} B_2\right) u^n + (\nu_1 \tilde{B}_1 + \nu_2 \tilde{B}_2)\left(\frac{\alpha^n + \alpha^{n+1}}{2}\right)$

not quite right $\to$ ADI: $\quad \left(I - \frac{\nu_1}{2} B_1\right)\left(I - \frac{\nu_2}{2} B_2\right) u^{n+1} = \left(I + \frac{\nu_1}{2} B_1\right)\left(I + \frac{\nu_2}{2} B_2\right) u^n + (\nu_1 \tilde{B}_1 + \nu_2 \tilde{B}_2)\left(\frac{\alpha^n + \alpha^{n+1}}{2}\right) + \frac{1}{4} \nu_1 \nu_2 \tilde{B}_1 \tilde{B}_2 (\alpha^n - \alpha^{n+1})$

For the ADI scheme, this requires a little care ($\tilde{B}_1\tilde{B}_2$ makes no sense)

for the infinite domain, $\left(I-\frac{\nu_1}{2}B_1\right)\left(I-\frac{\nu_2}{2}B_2\right)$

$$= I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}E$$

where the stencil for E is $\begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix}$

<u>The discrete versions of E are:</u>

$$E = \begin{pmatrix} \begin{pmatrix} 4 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} -2 & 4 & -2 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & -2 & 4 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} -2 & 1 & 0 \\ 4 & -2 & 0 \\ -2 & 1 & 0 \end{pmatrix} & \begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix} & \begin{pmatrix} 0 & 1 & -2 \\ 0 & -2 & 4 \\ 0 & 1 & -2 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 \\ -2 & 1 & 0 \\ 4 & -2 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ -2 & 4 & -2 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & -2 & 4 \end{pmatrix} \end{pmatrix},$$

$$\tilde{E} = \begin{pmatrix} \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ -2 & & & & 0 \\ 1 & & & & 0 \\ 0 & & & & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & & & & 0 \\ 0 & & & & 0 \\ 0 & & & & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 1 & -2 & 1 \\ 0 & & & & -2 \\ 0 & & & & 1 \\ 0 & & & & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & & & & 0 \\ -2 & & & & 0 \\ 1 & & & & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & & & & 0 \\ 0 & & & & 0 \\ 0 & & & & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & & & & 1 \\ 0 & & & & -2 \\ 0 & & & & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & & & & 0 \\ 1 & & & & 0 \\ -2 & & & & 0 \\ 1 & -2 & 1 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & & & & 0 \\ 0 & & & & 0 \\ 0 & & & & 0 \\ 0 & 1 & -2 & 1 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & & & & 0 \\ 0 & & & & 1 \\ 0 & & & & -2 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix} \end{pmatrix}$$

and our scheme is

$$\left(I - \frac{\nu_1}{2}B_1 - \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}E\right)u^{n+1} + \left(-\frac{\nu_1}{2}\tilde{B}_1 - \frac{\nu_2}{2}\tilde{B}_2 + \frac{\nu_1\nu_2}{4}\tilde{E}\right)\alpha^{n+1} = \left(I + \frac{\nu_1}{2}B_1 + \frac{\nu_2}{2}B_2 + \frac{\nu_1\nu_2}{4}E\right)u^n + \left(\frac{\nu_1}{2}\tilde{B}_1 + \frac{\nu_2}{2}\tilde{B}_2 + \frac{\nu_1\nu_2}{4}\tilde{E}\right)\alpha^n$$

or (since $E = B_1 B_2$):

$$\left(I - \frac{\nu_1}{2}B_1\right)\left(I - \frac{\nu_2}{2}B_2\right)u^{n+1} = \left(I + \frac{\nu_1}{2}B_1\right)\left(I + \frac{\nu_2}{2}B_2\right)u^n$$

$$\left.\begin{aligned} &+ \left(\nu_1\tilde{B}_1 + \nu_2\tilde{B}_2\right)\left(\frac{\alpha^{n+1}+\alpha^n}{2}\right) \\ &+ \frac{1}{4}\nu_1\nu_2\tilde{E}\left(\alpha^n - \alpha^{n+1}\right) \end{aligned}\right\}$$

again the boundary conditions enter as a source term attached to the nodes adjacent to the boundary

228B Lec 13

Last time: introduction to the wave equation

vibrating string: $\rho u_{tt} = T u_{xx}$, $u(0) = u(L) = 0$ — T = tension, ρ = linear density

free space: $u_{tt} = c^2 \nabla^2 u$ (plane waves)

1d: $\begin{cases} u_{tt} = c^2 u_{xx} \\ u(x,0) = g_0(x) \\ u_t(x,0) = g_1(x) \end{cases}$

d'Alembert's formula

$$u(x,t) = \frac{g_0(x-ct) + g_0(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} g_1(\xi)\, d\xi$$

today: reduction to 1st order system

schemes for the baby (1-way) wave equation $u_t + a u_x = 0$

domain of dependence/influence

CFL condition

stability of a few schemes

reduction to 1st order system:

$$u_{tt} = c^2 u_{xx} \qquad v = \begin{pmatrix} u_x \\ u_t \end{pmatrix}$$

$$v_t = \begin{pmatrix} u_{xt} \\ u_{tt} \end{pmatrix} = \begin{pmatrix} u_{xt} \\ c^2 u_{xx} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ c^2 & 0 \end{pmatrix}}_{A} \begin{pmatrix} u_x \\ u_t \end{pmatrix}_x = A v_x$$

diagonalize $A = U \Lambda U^{-1}$, $\Lambda = \begin{pmatrix} c & \\ & -c \end{pmatrix}$, $U = \begin{pmatrix} 1 & 1 \\ c & -c \end{pmatrix}$, $U^{-1} = \begin{pmatrix} 1/2 & 1/2c \\ 1/2 & -1/2c \end{pmatrix}$

$$v_t = U \Lambda U^{-1} v_x \qquad w = U^{-1} v$$

$$w_t = \Lambda w_x \longleftarrow \begin{cases} (w_1)_t = c\,(w_1)_x \\ (w_2)_t = -c\,(w_2)_x \end{cases}$$

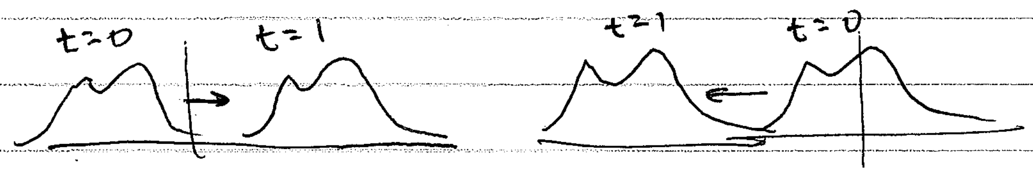
Result: the components of w decouple and satisfy the baby (one-way) wave equation with $a = \pm c$.

---

baby wave equation: $u_t + a u_x = 0$

$u(x,0) = g(x)$ ← initial condition

general solution: $u(x,t) = g(x-at)$



$a > 0$ (right moving) $\qquad\qquad a < 0$ (left moving)

here we plot $u(x,t)$ as fcn of x with t frozen.

<u>schemes</u>: explicit $\left\{\begin{matrix}\text{upwind } (a>0) \\ \text{downwind } (a<0)\end{matrix}\right\}$: $\underbrace{\dfrac{u_j^{n+1} - u_j^n}{k}}_{D_t^+ u_j^n} + a \underbrace{\dfrac{u_j^n - u_{j-1}^n}{h}}_{D_x^- u_j^n} = 0$

explicit $\left\{\begin{matrix}\text{upwind } (a<0) \\ \text{downwind } (a>0)\end{matrix}\right\}$: $D_t^+ u_j^n + a D_x^+ u_j^n = 0$

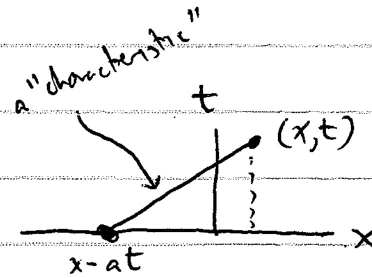
upwind means the space stencil is one-sided in the direction information is coming from (if waves go L to R, stencil "looks left" D_x^-)

explicit centered: $D_t^+ u_j^n + a \underbrace{\dfrac{u_{j+1}^n - u_{j-1}^n}{2h}}_{D_x^0 u_j^n} = 0$

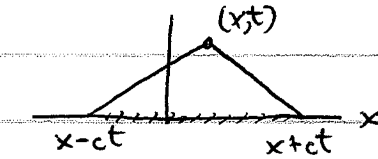
each of these schemes has an implicit counterpart.

Domain of dependence and influence.

The exact solution $u(x,t) = g(x-at)$ depends only on the value of g at one point:

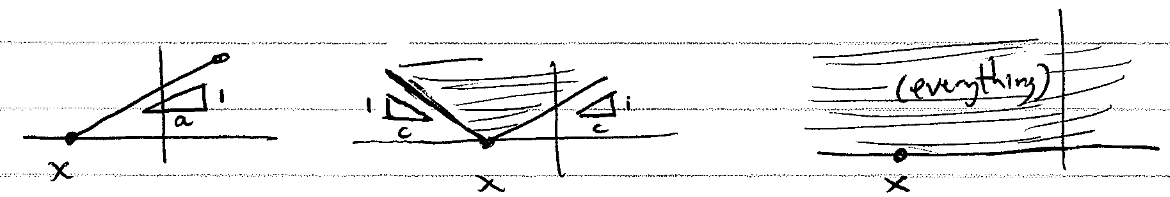


for the full wave equation $u_{tt} = c^2 u_{xx}$, it depends only on the values of g_0, g_1 over a finite range



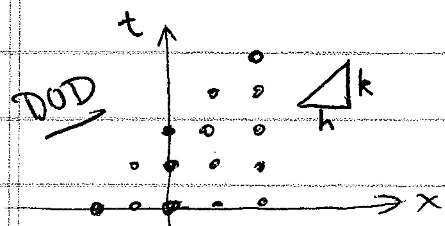
By contrast, the solution $u(x,t)$ of the heat equation depends on $g(x)$ for all $x \in \mathbb{R}$ no matter how small t is.

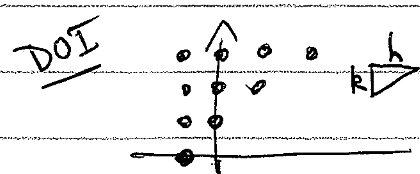
similarly, we can draw the domain of influence of an initial point:



baby wave eqn — full wave eqn — heat eqn

Our schemes have domains of dependence and influence as well.





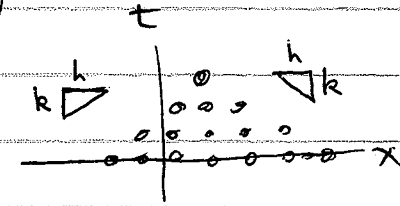
upwind $(a>0)$

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

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

$$\nu = a\frac{k}{h}$$

note: if $\nu = 1$, upwind gives the exact soln ...

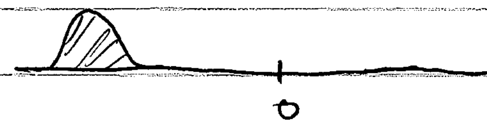


centered

$$u_j^{n+1} = u_j^n - a\frac{k}{h}\left(\frac{u_{j+1}^n - u_{j-1}^n}{2}\right)$$

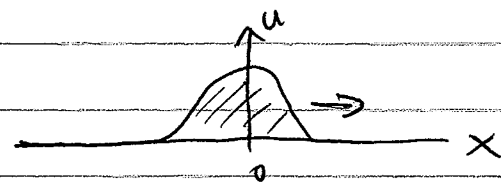
$$= \frac{\nu}{2}u_{j-1}^n + u_j^n - \frac{\nu}{2}u_{j+1}^n$$

The Courant - Friedrichs - Lewy (CFL) condition states that a necessary condition for a scheme to converge is that the DOD of the scheme contain that of the exact solution in the mesh refinement limit.

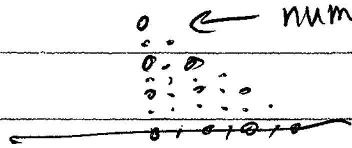
example: $g(x) =$ 

← a bump far off to the left.

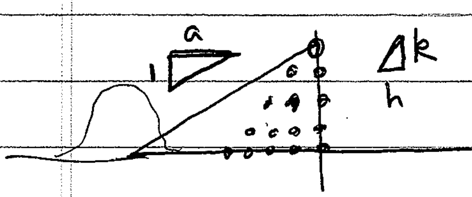
suppose t is in the range where the bump passes the origin:

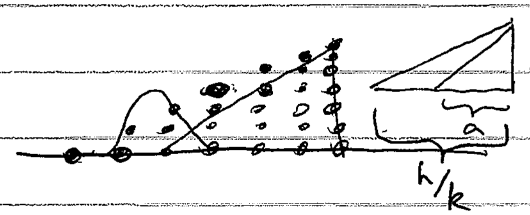
$u(x,t)$ 

The downwind scheme has no hope of getting the answer right:


← numerical solution remains zero no matter how you refine the mesh since $g(x)=0$ for $x \geq 0$.

The upwind scheme requires that $\frac{k}{h}$ is small enough:





$\frac{k}{h} > \frac{1}{a}$ still hopeless, scheme can't know what g is like where it matters.

$\frac{k}{h} < \frac{1}{a}$ finally it's possible to get a good approximation of the solution (but no guarantees, it's only a necessary condition).

We know from the Lax-Richtmyer theory that

consistency + stability $\Rightarrow$ convergence

all our schemes are consistent ~~(since we write them down using D^+, D^-, etc.)~~

so (no CFL) $\Rightarrow$ (no convergence) $\Rightarrow$ (no stability)

Let's compute the norms of our operators B^n and verify this.

upwind $(a>0)$: $\quad u^{n+1} = Bu^n, \quad Bu_j = \nu u_{j-1} + (1-\nu)u_j, \quad \nu = a\frac{k}{h}$

$$B = \begin{pmatrix} \ddots & \ddots & & & \\ & \nu & 1-\nu & & \\ & & \nu & 1-\nu & \\ & & & \nu & 1-\nu \\ & & & & \ddots \quad \ddots \end{pmatrix}$$

$\nu \le 1$: $\|B\|_\infty = \|B\|_1 = |\nu| + |1-\nu|$
$= \nu + 1 - \nu = 1$ (stable)

$\nu > 1$: $\|B\|_\infty = \|B\|_1 = |\nu| + |1-\nu|$
$= \nu + \nu - 1 = 2\nu - 1 > 1$
(unstable)

note: $\nu \le 1$ is the same as $\frac{k}{h} \le \frac{1}{a}$

downwind $(a>0)$ $\quad Bu_j = (1+\nu)u_j - \nu u_{j+1}$

$$B = \begin{pmatrix} \ddots & \ddots & & & \\ & 1+\nu & -\nu & & \\ & 0 & 1+\nu & -\nu & \\ & & 0 & 1+\nu & -\nu \\ & & & \ddots & \ddots \quad \ddots \end{pmatrix}$$

for any ν:
$\|B\|_\infty = \|B\|_1 = |1+\nu| + |-\nu|$
$= 1 + 2\nu > 1$
(unstable)

centered $(a>0)$ $\quad Bu_j = \frac{\nu}{2}u_{j-1} + u_j - \frac{\nu}{2}u_{j+1}$

$$B = \begin{pmatrix} 1 & -\frac{\nu}{2} & & & \\ \nu/2 & 1 & -\nu/2 & & \\ & \nu/2 & 1 & -\nu/2 & \\ & & \nu/2 & 1 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}$$

for any ν, $\|B\|_\infty = \|B\|_1 = 1 + \left|\frac{\nu}{2}\right| + \left|-\frac{\nu}{2}\right|$
$= 1 + \nu > 1$

(unstable)

we can also compute amplification factors to determine the 2-norm:

<u>upwind:</u> $Bu_j = \nu u_{j-1} + (1-\nu) u_j$

$$G(\xi) = \nu e^{-i\xi} + (1-\nu) e^{0}$$

$$= 1 - \nu + \nu e^{-i\xi} \qquad e^{i\theta} = \cos\theta + i \sin\theta$$

$$= (1 - \nu + \nu \cos\xi) - i\nu \sin\xi$$

$$|G(\xi)|^2 = (1-\nu+\nu\cos\xi)^2 + \nu^2 \sin^2\xi$$

$$= (1-\nu)^2 + 2(1-\nu)\nu\cos\xi + \nu^2\cos^2\xi + \nu^2\sin^2\xi$$

$$= 1 - 2\nu + \nu^2 + 2(1-\nu)\nu(1 - 2\sin^2\tfrac{\xi}{2}) + \nu^2$$

$$1 - \cancel{2\nu} + \cancel{2\nu^2} + \cancel{2\nu} - \cancel{2\nu^2} - 4\nu s^2 + 4\nu^2 s^2$$

$$= 1 - 4\nu(1-\nu)\sin^2(\xi/2)$$

so if $\nu \le 1$, $0 \le |G(\xi)|^2 \le 1$ $\Rightarrow$ $\|G\|_\infty = 1$
$|G(0)|^2 = 1$
$\Rightarrow$ (stable)

and if $\nu > 1$, $|G(\pi)|^2 = 1 + 4\nu(\nu-1) > 1 \Rightarrow \|G\|_\infty > 1$
$\Rightarrow$ (unstable)

<u>downwind:</u> $|G(\xi)|^2 = 1 + 4\nu(1+\nu)\sin^2(\xi/2)$

$\|G\|_\infty > 1$ no matter what ν is $\Rightarrow$ (unstable)

<u>centered:</u> $Bu_j = \frac{\nu}{2} u_{j-1} + u_j - \frac{\nu}{2} u_{j+1}$

$$G(\xi) = \frac{\nu}{2} e^{-i\xi} + 1 - \frac{\nu}{2} e^{i\xi} = 1 - i\nu \frac{e^{i\xi} - e^{-i\xi}}{2i}$$

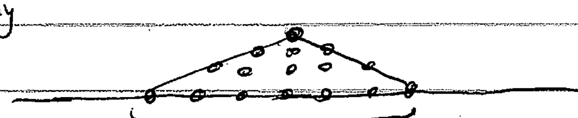
$$= 1 - i\nu \sin\xi$$

$$\|G\|_\infty = \sqrt{1+\nu^2} > 1 \qquad \text{(unstable)}$$

Summary: (1) the CFL condition tells you when the scheme is guaranteed to be bad. (2) Often the stability breakpoint occurs exactly where the CFL condition is satisfied

(3) some schemes (like the forward time centered space) are unstable even though they satisfy CFL. (CFL does not give sufficient conditions)

(4) for the heat equation, you include the DOD only in the limit as k and $h \to 0$ holding $\frac{k}{h^2}$ constant. That's OK because the effect of g decays exponentially



numerical DOD

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-\xi)^2}{4t}} g(\xi)\, d\xi$$

the contribution from ξ outside the numerical DOD goes to zero as the DOD grows.

Note: there's no way to save the downwind scheme by using a different refinement path. It's OK for

$\|B(k)\| \leq 1 + Ck$

since then we have $\|B(k)^n\| \leq (1+Ck)^n \leq e^{Ckn} \leq e^{CT}$

but for the downwind scheme we have

$\|B(k)\|_2 = 1 + 2\nu = 1 + 2a\frac{k}{h}$

so as $h \to 0$, the factor $\frac{2a}{h}$ on k goes to infinity and the C here blows up

it doesn't matter how much we refine k compared to h — you're still only getting information from the right!

on the other hand, we <u>can</u> save the centered scheme, it's just expensive.

for example, let $h = \sqrt{ak}$ be our refinement path.

Then $\|B(k)\|_2 = \sqrt{1+\nu^2} = \sqrt{1+(a\frac{k}{h})^2} = \sqrt{1+ak} \le 1+\frac{1}{2}ak$

so $\|B(k)^n\|_2 \le e^{\frac{1}{2}aT}$

$\therefore$ scheme is stable with this refinement path

<u>But</u>: (1) expensive $k = O(h^2)$

(2) error bound grows exponentially in time.

Next time we'll see how to fix these problems using the Lax-Wendroff and Lax-Friedrichs schemes.

Last time: upwind, downwind, centered schemes for baby wave equation
CFL condition gives a necessary condition for convergence
(tells you when a scheme is guaranteed to be bad)
started analyzing stability of these schemes

today: finish stability analysis
show how to rescue centered scheme with a different refinement path
postpone → Lax Friedrichs, Lax-Wendroff, Crank-Nicolson schemes
heat equation with spherical symmetry

comment on 1-norm & ∞-norm analysis: $u^{n+1} = Bu^n$

showing that $\|B\|_1 \le 1$ or $\|B\|_\infty \le 1$ implies the scheme is stable since $\|B^n\| \le \|B\|^n \le 1$ in that case.
So our upwind scheme

$$Bu_j = \nu u_{j-1} + (1-\nu)u_j \qquad \nu = a\frac{k}{h},\ a>0$$

is stable for $\nu \le 1$. But showing that $\|B\|_1 > 1$ or $\|B\|_\infty > 1$ does not imply the scheme is unstable since $\|B^n\|$ can be less than $\|B\|^n$. Matrix example: $A = \begin{pmatrix} 0 & 10 \\ 0 & 0 \end{pmatrix}$, $A^n = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ for $n \ge 2$

so $\|A^n\| = 0$ while $\|A\|^n = 10^n$ for $n \ge 2$.

But the 2-norm analysis does tell you about $\|B^n\|$ when B is normal (i.e. $B^TB = BB^T$). (Finite difference operators are normal).

we simply compute amplification factors to determine the 2-norm:

<u>upwind:</u> $Bu_j = \nu u_{j-1} + (1-\nu) u_j$

$$G(\xi) = \nu e^{-i\xi} + (1-\nu) e^{0}$$

$$= 1-\nu + \nu e^{-i\xi}$$ $e^{i\theta} = \cos\theta + i\sin\theta$

$$= (1-\nu+\nu\cos\xi) - i\nu\sin\xi$$

$$|G(\xi)|^2 = (1-\nu+\nu\cos\xi)^2 + \nu^2\sin^2\xi$$

$$= (1-\nu)^2 + 2(1-\nu)\nu\cos\xi + \nu^2\cos^2\xi + \nu^2\sin^2\xi$$

$$= 1-2\nu + \nu^2 + 2(1-\nu)\nu(1-2\sin^2\tfrac{\xi}{2}) + \nu^2$$

$$1 - \cancel{2\nu} + \cancel{2\nu^2} + \cancel{2\nu} - \cancel{2\nu^2} - 4\nu s^2 + 4\nu^2 s^2$$

$$= 1 - 4\nu(1-\nu)\sin^2(\xi/2)$$

so if $\nu \le 1$, $0 \le |G(\xi)|^2 \le 1$, $|G(0)|^2 = 1$ $\Rightarrow \|G\|_\infty = 1$ $\Rightarrow$ (stable)

and if $\nu > 1$, $|G(\pi)|^2 = 1 + 4\nu(\nu-1) > 1 \Rightarrow \|G\|_\infty > 1$

since $\|B^n\|_2 = \|G^n\|_\infty = \|G\|_\infty^n$, the scheme is <u>unstable</u> if $\nu = \frac{k}{h} > 1$ is held fixed.

<u>downwind:</u> $|G(\xi)|^2 = 1 + 4\nu(1+\nu)\sin^2(\xi/2)$

$\|G\|_\infty > 1$ no matter what ν is $\Rightarrow$ (unstable)

<u>centered:</u> $Bu_j = \frac{\nu}{2} u_{j-1} + u_j - \frac{\nu}{2} u_{j+1}$

$$G(\xi) = \frac{\nu}{2} e^{-i\xi} + 1 - \frac{\nu}{2} e^{i\xi} = 1 - i\nu\frac{e^{i\xi} - e^{-i\xi}}{2i}$$

$$= 1 - i\nu\sin\xi$$

$\|G\|_\infty = \sqrt{1+\nu^2} > 1$ (unstable if $\nu = \frac{k}{h}$ is held fixed)

Note that the centered scheme is unstable even though it satisfies the CFL condition. (CFL does not give sufficient conditions for convergence)

Saving the centered scheme.

consider the refinement path $h = \sqrt{ak}$, $\nu = a\frac{k}{h} = \sqrt{ak}$

Then $\|B(k)\|_2 = \sqrt{1+\nu^2} = \sqrt{1+ak} \le 1 + \frac{1}{2}ak$

so $\|B(k)^n\|_2 \le \left(e^{\frac{1}{2}ak}\right)^n \le e^{\frac{1}{2}aT}$

$\sqrt{1+\varepsilon} \le 1 + \frac{\varepsilon}{2}$ for all $\varepsilon > 0$

$\left(1+\varepsilon \le 1 + \varepsilon + \frac{\varepsilon^2}{4}\right)$

$\therefore$ scheme is stable with this refinement path.

but ① it's expensive $(k = O(h^2))$

② error bound grows exponentially with time.

Note: this wouldn't have worked in the 1-norm or ∞-norm analysis.

$$\|B\|_1 = \left|\frac{\nu}{2}\right| + |1| + \left|-\frac{\nu}{2}\right| = 1 + \nu = 1 + a\frac{k}{h}$$

$$\|B^n\|_1 \le \|B\|_1^n = \left(1 + a\frac{k}{h}\right)^n \ge 1 + a\frac{nk}{h} \ge 1 + \frac{aT}{2h} \to \infty \text{ as } h \to 0$$

$\frac{T}{2} \le nk \le T$

lost the game here (with 2-norm it's an equality)

$\varepsilon > 0 : (1+\varepsilon)^n = \sum_{\ell=0}^{n} \binom{n}{\ell} 1^{n-\ell}\varepsilon^\ell \ge \underbrace{1 + n\varepsilon}_{\text{first two terms}}$

The downwind scheme can't be saved by using a different refinement path.

$$\|B^n\|_2 = \|B\|_2^n = (1+2\nu)^n \ge 1 + 2a\frac{nk}{h} \ge 1 + \frac{aT}{h} \to \infty$$

as $h \to 0$

It doesn't matter how much we refine k compared to h... you're still only getting information from the right!

$\frac{T}{2} \le nk \le T$

Better schemes for the wave equation

Of the schemes so far, upwind has been the best, but it has 2 drawbacks:

① it's first order in time and space (unless $\nu = 1$ – then it's exact)

② for systems you could have some waves moving L to R and others R to L ... which way's upwind?

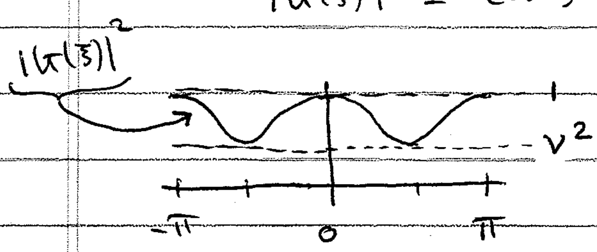
Lax-Friedrichs scheme

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

$$u_j^{n+1} = Bu_j^n = \left(\frac{1}{2} + \frac{\nu}{2}\right)u_{j-1}^n + \left(\frac{1}{2} - \frac{\nu}{2}\right)u_{j+1}^n$$

$$G(\xi) = (1+\nu)\frac{e^{-i\xi}}{2} + (1-\nu)\frac{e^{i\xi}}{2} = \cos\xi - i\nu\sin\xi$$

$$|G(\xi)|^2 = \cos^2\xi + \nu^2\sin^2\xi = 1 - (1-\nu^2)\sin^2\xi$$



$$\|B^n\| = \begin{cases} 1 & \nu \le 1 \quad \text{(stable)} \\ \nu^n & \nu > 1 \quad \text{(unstable)} \end{cases}$$

truncation error of Lax-Friedrichs: $O(k + h^2)$

Lax-Wendroff scheme

this time we derive the scheme by trying to knock off more terms in the Taylor expansion (rather than using a geometric construction)

exact sol'n $\rightarrow$ $$u(x,t+k) = u + ku_t + \frac{k^2}{2}u_{tt} + \cdots$$

$$= u + k(-au_x) + \frac{k^2}{2}(a^2 u_{xx}) + \cdots$$

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

or $$u_j^{n+1} = Bu_j^n = \tfrac{1}{2}\nu(1+\nu)u_{j-1}^n + (1-\nu^2)u_j^n - \tfrac{1}{2}\nu(1-\nu)u_{j+1}^n$$

the amplification factor is

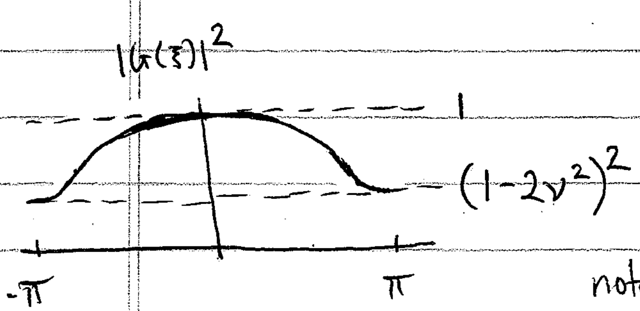
$$G(\xi) = 1 - i\nu\frac{e^{i\xi} - e^{-i\xi}}{2i} + \frac{\nu^2}{2}\left(e^{-i\xi} - 2 + e^{i\xi}\right)$$

$$= 1 - i\nu \sin\xi - \nu^2(1-\cos\xi)$$

$$= 1 - 2\nu^2\sin^2\xi/2 - i\nu\sin\xi$$

$$|G(\xi)|^2 = 1 - 4\nu^2\sin^2\xi/2 + 4\nu^4\sin^4\xi/2 + \underbrace{\nu^2\sin^2\xi}_{4\nu^2\sin^2\xi/2\,\underbrace{\cos^2\xi/2}_{1-\sin^2\xi/2}}$$

$$= 1 - 4\nu^2(1-\nu^2)\sin^4\xi/2$$



so $$\|B^n\|_2 = \begin{cases} 1 & \nu \le 1 \quad \text{(stable)} \\ (1-2\nu^2)^n & \nu > 1 \quad \text{(unstable)} \end{cases}$$

note: if $\nu = 1$ you get the exact solution, just like upwind.

Lax-Wendroff is a rare instance in mathematics where going after more accuracy by including more terms in a Taylor expansion actually improves stability. (Runge-Kutta is another example)

→ method is $O(k^2+h^2)$ <u>and</u> stable for $\nu \le 1$ with no exponential growth in the error bound or special refinement paths required. <u>And</u>, the scheme is centered in space, so it generalizes to <u>systems</u> with right and left moving waves simultaneously. (more later)

Crank-Nicolson $\qquad u_t = -a u_x$

$$\frac{u^{n+1}-u^n}{k} = \frac{1}{2}\left[-a\,\frac{u^{n+1}_{j+1}-u^{n+1}_{j-1}}{2h} - a\,\frac{u^n_{j+1}-u^n_{j-1}}{2h}\right]$$

$$\left(I + \frac{\nu}{2}B_1\right)u^{n+1} = \left(I - \frac{\nu}{2}B_1\right)u^n, \qquad B_1 u_j = \frac{1}{2}\left(u_{j+1}-u_{j-1}\right)$$

$$G_1(\xi) = i\,\frac{e^{i\xi}-e^{-i\xi}}{2i} = i\sin\xi$$

$$u^{n+1} = Bu^n$$

$$G(\xi) = \frac{1 - i\frac{\nu}{2}\sin\xi}{1 + i\frac{\nu}{2}\sin\xi}$$

$|G(\xi)| = 1$ $\qquad$ unconditionally stable for any ν. (implicit methods always satisfy CFL)

truncation error: $O(k^2+h^2)$ ← so probably you'd want to take timesteps with $h \approx ak$ anyways (unconditional stability not as important for wave eqn. as it is for heat eqn.)

Heat equation with spherical symmetry

$[f] = \frac{cal}{cm^3 \cdot s}$

full 3-d heat eqn: $\rho c \frac{\partial u}{\partial t} + \nabla \cdot J = f$

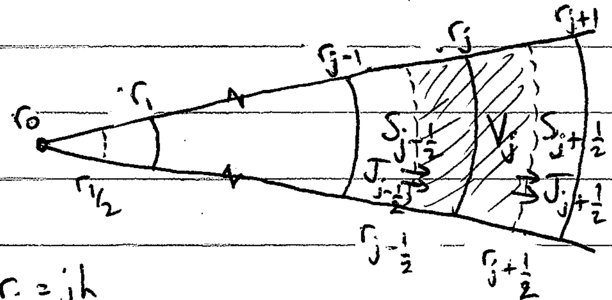
$[J] = \frac{cal}{cm^2 \cdot s}$

Fourier's law: $J = -K \nabla u$

$[K] = \frac{cal}{cm\ s\ K}$

$[c] = \frac{cal}{g \cdot cm}$

$[\rho] = g/cm^3$



$r_j = jh$

integrate over spherical shell

$$\frac{\partial}{\partial t}\left[\rho c \iint_{V_j} u\, dV\right] + \iint_{S_{j-\frac{1}{2}} + S_{j+\frac{1}{2}}} J \cdot n\, dA = \iint_{V_j} f\, dV$$

approximate the integrals:

$\frac{4}{3}\pi\left(r^3_{j+\frac{1}{2}} - r^3_{j-\frac{1}{2}}\right) f_j$

$$\frac{\partial}{\partial t}\left[\rho c \frac{4}{3}\pi\left(r^3_{j+\frac{1}{2}} - r^3_{j-\frac{1}{2}}\right) u_j\right] + \underbrace{4\pi r^2_{j+\frac{1}{2}} J_{j+\frac{1}{2}} - 4\pi r^2_{j-\frac{1}{2}} J_{j-\frac{1}{2}}}_{-4\pi K r^2_{j+\frac{1}{2}} \left.\frac{\partial u}{\partial r}\right|_{r_{j+\frac{1}{2}}} + 4\pi K r^2_{j-\frac{1}{2}} \left.\frac{\partial u}{\partial r}\right|_{r_{j-\frac{1}{2}}}} = (\ )$$

approximate $\left.\frac{\partial u}{\partial r}\right|_{r_{j+\frac{1}{2}}} = \frac{u_{j+1} - u_j}{h}$

$$\frac{\partial u_j}{\partial t} = \frac{3K}{\rho c} \frac{\left(r^2_{j+\frac{1}{2}} \frac{u_{j+1} - u_j}{h} - r^2_{j-\frac{1}{2}} \frac{u_j - u_{j-1}}{h}\right)}{r^3_{j+\frac{1}{2}} - r^3_{j-\frac{1}{2}}} + \frac{f_j}{\rho c}$$

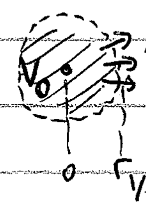
denominator $= (3r_j^2 + h^2/4)h$

Now discretize in time using $\frac{\partial u_j}{\partial t} \approx D_t^+ u_j^n$ and on the RHS use:

explicit: u_j^n, f_j^n or implicit u_j^{n+1}, f_j^{n+1} or C-N: $\frac{1}{2}(u_j^{n+1} + u_j^n)$, $\frac{1}{2}(f_j^{n+1} + f_j^n)$

The origin needs to be dealt with specially since there is no "flux from the left, i.e. from the inside"

integrate over sphere



$$\frac{\partial}{\partial t}\left[\rho C \iiint_{V_0} u \, dV\right] + \iint_{S_{1/2}} J \cdot n \, dA = \iiint_{V_0} f \, dV$$

approx. integral:

$$\frac{\partial}{\partial t}\left[\rho C \frac{4}{3}\pi r_{1/2}^3 u_0\right] - 4\pi K r_{1/2}^2 \left.\frac{\partial u}{\partial r}\right|_{r_{1/2}} = \frac{4}{3}\pi r_{1/2}^3 f_0$$

approx. $\left.\frac{\partial u}{\partial r}\right|_{r_{1/2}} = \frac{u_1 - u_0}{h}$

$$\frac{\partial u_0}{\partial t} = \frac{3K}{\rho C} \frac{r_{1/2}^2 \frac{u_1 - u_0}{h}}{r_{1/2}^3} + \frac{f_0}{\rho C} \qquad r_{1/2} = \frac{h}{2}$$

for definiteness, consider the fully implicit scheme:

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{K}{\rho C} \frac{1}{h^2} B u_j^{n+1} + \frac{f_j^{n+1}}{\rho C}$$

(I absorbed the 3 into B)

$$B u_j = \begin{cases} 6(u_1 - u_0) & j = 0 \\ \dfrac{(j-\frac{1}{2})^2 u_{j-1} - \left[(j+\frac{1}{2})^2 + (j-\frac{1}{2})^2\right] u_j + (j+\frac{1}{2})^2 u_{j+1}}{j^2 + 1/12} & 1 \le j < M \end{cases}$$

B is tridiagonal but is not constant along diagonals since the underlying PDE $\rho C \frac{\partial u}{\partial t} - K r^{-2} \frac{\partial}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) = f$ does not have constant coefficients. Note that for large j, the stencil is close to $(1 \;\; -2 \;\; 1)$ (shell thickness small compared to radius of curvature of shell) shells look like planes)))

Question 1: does the implicit scheme make sense? must show that

$$A = I - \nu B \quad \text{is invertible.} \quad \left(\nu = \frac{K}{\rho C}\frac{k}{h^2}\right)$$

Gershgorin theorem: Let A be an arbitrary $M\times M$ matrix. Then the eigenvalues λ of A are located in the union of the M disks

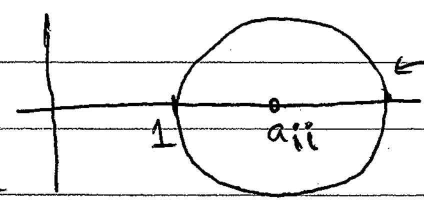
$$|\lambda - a_{ii}| \le \sum_{j\ne i} |a_{ij}|$$

Now, B has the property that it's row sums are zero with off diagonals of the same sign.

$$B = \begin{pmatrix} -6 & 6 & 0 & 0 & 0 & 0 \\ 3/13 & -30/13 & 27/13 & 0 & 0 & 0 \\ 0 & 27/49 & -102/49 & 75/49 & 0 & 0 \\ 0 & 0 & 75/109 & -\frac{222}{109} & \frac{147}{109} & 0 \\ 0 & 0 & 0 & \frac{147}{193} & -\frac{390}{193} & \frac{243}{193} \\ \vdots & \vdots & \vdots & 0 & \frac{243}{301} & -\frac{606}{301} \end{pmatrix}$$

so $a_{ii} = 1 - \nu b_{ii} > 1$ and $\sum_{j\ne i} |a_{ij}| = \sum_{j\ne i} |-\nu b_{ij}| = |\nu b_{ii}|$

complex plane



radius of Gershgorin disk is $a_{ii} - 1$
(disk lies to the right of $1 \in \mathbb{C}$)

Example: $i=0$: $a_{00} = 1 + 6\nu$, $|a_{01}| = |-6\nu| = 6\nu$

$i=1$: $a_{11} = 1 + \frac{30}{13}\nu$, $|a_{10}| + |a_{12}| = \frac{30}{13}\nu$

$i=2$: $a_{22} = 1 + \frac{102}{49}\nu$, $|a_{21}| + |a_{23}| = \frac{102}{49}\nu$

Conclusion: All the eigenvalues of A have real part ≥ 1.

so A is invertible (no zero eigenvalues)

<u>Question 2:</u> B and A are not normal $(B^TB \neq BB^T)$

is there anything like our Fourier analysis to analyze these schemes?

<u>yes.</u> use a weighted norm:

$$\|u\|_{2,h}^2 = \frac{4}{3}\pi\left[r_{1/2}^3 u_0^2 + \sum_{j=1}^{M-1} \left(r_{j+\frac{1}{2}}^3 - r_{j-\frac{1}{2}}^3\right) u_j^2 \right]$$

$$(u,v)_{2,h} = u^T W \bar{v}, \quad W_{jj} = \begin{cases} \frac{\pi h^3}{6} & j=0 \\ 4\pi h^3 (j^2 + \frac{1}{12}) & 1 \le j < M \end{cases}$$

$$(u, Bv)_{2,h} = u^T W B \bar{v}$$

$$(Bu, v)_{2,h} = u^T B^T W \bar{v}$$

claim $\quad WB = (WB)^T = B^T W$

$$(WB)_{i,i+1} = \begin{cases} \pi h^3 & i=0 \\ 4\pi (i+\frac{1}{2})^2 h^3 & i>0 \end{cases}$$

equal

$$(WB)_{i+1,i} = 4\pi \left(i+1-\frac{1}{2}\right)^2 h^3 \qquad i \ge 0$$

So B is self-adjoint in this inner product.

$\therefore$ eigenvalues of B and A are real, eigenvectors are orthonormal.

$$\|A^{-1}\| \le 1$$

implicit scheme is stable for any choice of ν.

☆ 228 Lec 15 ☆

Last time: analysis of the heat equation with spherical symmetry
- non-constant coefficients prevent Fourier analysis from working
- Gershgorin's theorem replaces amplification factor analysis
- mesh weighted norms make the matrix self-adjoint

Today: rescue the centered scheme
better schemes for the wave equation
Lax-Wendroff, Lax-Friedrichs, Crank-Nicolson, Leapfrog

Saving the centered scheme

scheme: $u_j^{n+1} = Bu_j^n = \frac{\nu}{2} u_{j-1} + u_j - \frac{\nu}{2} u_{j+1}$, $\nu = a\frac{k}{h}$

we allow "a" to be positive or negative with this scheme

amplification factor: $G(\xi) = \frac{\nu}{2} e^{-i\xi} + 1 - \frac{\nu}{2} e^{i\xi}$

$$= 1 - i\nu \frac{e^{i\xi} - e^{-i\xi}}{2i} = 1 - i\nu \sin\xi$$

$$|G(\xi)| = \sqrt{1 + \nu^2 \sin^2 \xi}$$

$$\|G\|_\infty = \max_{-\pi \le \xi \le \pi} |G(\xi)| = \sqrt{1+\nu^2}$$

so if we fix $\nu = a\frac{k}{h}$ as we refine k and h, then

$$\|B^n\|_2 = \|G^n\|_\infty = (1+\nu^2)^{n/2} \to \infty \quad \text{as} \quad n \to \infty$$

even keeping $0 < nk \le T$.

∴ unstable.

Note that the centered scheme is unstable even though it satisfies the CFL condition. (CFL does not give sufficient conditions for convergence)

Saving the centered scheme.

consider the refinement path $h = \sqrt{|a|k}$, $\nu = a\frac{k}{h} = \sqrt{|a|k}\,\mathrm{sgn}(a)$

Then $\|B(k)\|_2 = \sqrt{1+\nu^2} = \sqrt{1+|a|k} \le 1 + \frac{1}{2}|a|k$

$\sqrt{1+\varepsilon} \le 1 + \frac{\varepsilon}{2}$ for all $\varepsilon > 0$

$\left(1+\varepsilon \le 1 + \varepsilon + \frac{\varepsilon^2}{4} \right)$

so $\|B(k)^n\|_2 \le \left(e^{\frac{1}{2}|a|k}\right)^n \le e^{\frac{1}{2}|a|T}$

$\therefore$ scheme is stable with this refinement path.

but ① it's expensive $(k = O(h^2))$

② error bound grows exponentially with time.

Note: this wouldn't have worked in the 1-norm or ∞-norm analysis.

$$\|B\|_1 = \left|\frac{\nu}{2}\right| + |1| + \left|-\frac{\nu}{2}\right| = 1 + |\nu| = 1 + |a|\frac{k}{h}$$

$$\|B^n\|_1 \le \|B\|_1^n = \left(1 + |a|\frac{k}{h}\right)^n \ge 1 + |a|\frac{nk}{h} \ge 1 + \frac{|a|T}{2h} \to \infty \text{ as } h \to 0$$

$\frac{T}{2} \le nk \le T$

lost the game here (~~with 2-norm it's~~ an equality) it's a worthless bound since $\|B\|_1^n \to \infty$ as $h \to 0$.

$\varepsilon > 0: \ (1+\varepsilon)^n = \sum_{\ell=0}^{n} \binom{n}{\ell} 1^{n-\ell} \varepsilon^\ell \ge \underbrace{1 + n\varepsilon}_{\text{first two terms}}$

The downwind scheme can't be saved by using a different refinement path.

$$\|B^n\|_2 = \|B\|_2^n = (1+2\nu)^n \ge 1 + 2a\frac{nk}{h} \ge 1 + \frac{aT}{h} \to \infty \text{ as } h \to 0$$

$\frac{T}{2} \le nk \le T$

It doesn't matter how much we refine k compared to h... you're still only getting information from the right!

downwind: $Bu_j = (1+\nu)u_j - \nu u_{j+1}$, $a > 0$

$|\hat{B}(\xi)|^2 = 1 + 4\nu(1+\nu)\sin^2(\xi/2)$

$\|\hat{B}\|_\infty = 1 + 2\nu$

Better schemes for the wave equation

Of the schemes so far, upwind has been the best, but it has 2 drawbacks: ① it's first order in time and space (unless $\nu = 1$ — then it's exact)

② for systems you could have some waves moving L to R and others R to L ... which way's upwind?

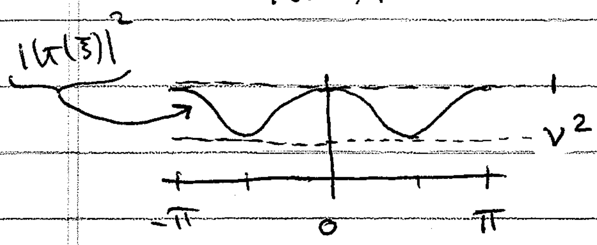
Lax-Friedrichs scheme

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

$$u_j^{n+1} = Bu_j^n = \left(\frac{1}{2} + \frac{\nu}{2}\right)u_{j-1}^n + \left(\frac{1}{2} - \frac{\nu}{2}\right)u_{j+1}^n$$

$$G(\xi) = (1+\nu)\frac{e^{-i\xi}}{2} + (1-\nu)\frac{e^{i\xi}}{2} = \cos\xi - i\nu\sin\xi$$

$$|G(\xi)|^2 = \cos^2\xi + \nu^2\sin^2\xi = 1 - (1-\nu^2)\sin^2\xi$$



$$\|B^n\| = \begin{cases} 1 & \nu \le 1 \quad \text{(stable)} \\ \nu^n & \nu > 1 \quad \text{(unstable)} \end{cases}$$

truncation error of Lax-Friedrichs: $O(k + h^2)$

Lax–Wendroff scheme

this time we derive the scheme by trying to knock off more terms in the Taylor expansion (rather than using a geometric construction)

exact sol'n $\rightarrow$ $$u(x, t+k) = u + k u_t + \frac{k^2}{2} u_{tt} + \cdots$$

$$= u + k(-a u_x) + \frac{k^2}{2}(a^2 u_{xx}) + \cdots$$

scheme:

$$u_j^{n+1} = u_j^n - ak \frac{u_{j+1}^n - u_{j-1}^n}{2h} + \frac{a^2 k^2}{2} \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2}$$

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

the amplification factor is

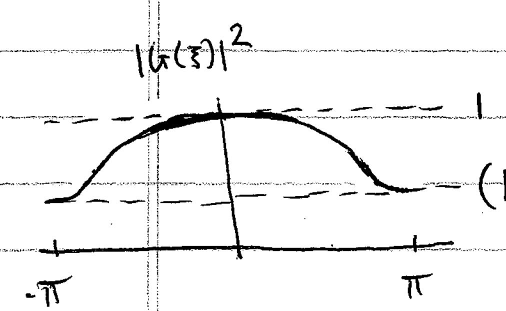
$$G(\xi) = 1 - i\nu \frac{e^{i\xi} - e^{-i\xi}}{2i} + \frac{\nu^2}{2}\left(e^{-i\xi} - 2 + e^{i\xi}\right)$$

$$= 1 - i\nu \sin\xi - \nu^2(1 - \cos\xi)$$

$$= 1 - 2\nu^2 \sin^2 \xi/2 - i\nu \sin\xi$$

$$|G(\xi)|^2 = 1 - 4\nu^2 \sin^2 \xi/2 + 4\nu^4 \sin^4 \xi/2 + \underbrace{\nu^2 \sin^2\xi}_{4\nu^2 \sin^2 \xi/2 \underbrace{\cos^2 \xi/2}_{1 - \sin^2 \xi/2}}$$

$$= 1 - 4\nu^2(1-\nu^2)\sin^4 \xi/2$$



so $$\|B^n\|_2 = \begin{cases} 1 & |\nu| \le 1 \quad \text{(stable)} \\ (1-2\nu^2)^n & |\nu| > 1 \quad \text{(unstable)} \end{cases}$$

note: if $|\nu| = 1$ you get the exact solution, just like upwind.

Lax-Wendroff is a rare instance in mathematics where going after more accuracy by including more terms in a Taylor expansion actually improves stability. (Runge-Kutta is another example)

→ method is $O(k^2+h^2)$ <u>and</u> stable for $|\nu| \le 1$ with no exponential growth in the error bound or special refinement paths required. <u>And</u>, the scheme is centered in space, so it generalizes to <u>systems</u> with right and left moving waves simultaneously. (more later)

Crank-Nicolson $\quad u_t = -a u_x$

$$\frac{u^{n+1} - u^n}{k} = \frac{1}{2}\left[-a \, \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2h} - a \, \frac{u_{j+1}^{n} - u_{j-1}^{n}}{2h} \right]$$

$$\left(I + \frac{\nu}{2} B_1\right) u^{n+1} = \left(I - \frac{\nu}{2} B_1\right) u^n, \qquad B_1 u_j = \frac{1}{2}\left(u_{j+1} - u_{j-1}\right)$$

$$G_1(\xi) = i \, \frac{e^{i\xi} - e^{-i\xi}}{2i} = i \sin \xi$$

$$u^{n+1} = B u^n$$

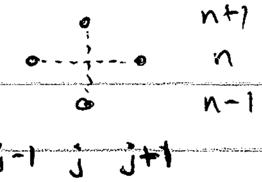
$$G(\xi) = \frac{1 - i \frac{\nu}{2} \sin \xi}{1 + i \frac{\nu}{2} \sin \xi}$$

$$|G(\xi)| = 1$$

unconditionally stable for any ν.
(implicit methods always satisfy CFL)

truncation error: $O(k^2 + h^2)$ ← so probably you'd want to take timesteps with $h \simeq ak$ anyways (unconditional stability not as important for wave eqn. as it is for heat eqn.)

Leapfrog scheme: $\quad u_t = -a u_x$



$$\frac{u_j^{n+1} - u_j^{n-1}}{2k} = -a \frac{u_{j+1}^n - u_{j-1}^n}{2h} \leftarrow D_t^0 u_j^n = -a D_x^0 u_j^n$$

$$u_j^{n+1} = u_j^{n-1} - \nu u_{j+1}^n + \nu u_{j-1}^n$$ (example of a multistep method)

how does the Fourier transform evolve?

$$\sum_j u_j^{n+1} e^{-ij\xi} = \sum_j \left(u_j^{n-1} - \nu u_{j+1}^n + \nu u_{j-1}^n \right) e^{-ij\xi}$$

$$\hat{u}^{n+1}(\xi) = \hat{u}^{n-1}(\xi) - \nu e^{i\xi} \hat{u}^n(\xi) + \nu e^{-i\xi} \hat{u}^n(\xi)$$

$$= \hat{u}^{n-1}(\xi) - (2i\nu \sin\xi)\, \hat{u}^n(\xi)$$

More generally, if $\quad u^{n+1} = B_1 u^n + B_2 u^{n-1}$

then $\quad \hat{u}^{n+1}(\xi) = G_1(\xi)\hat{u}^n(\xi) + G_2(\xi)\hat{u}^{n-1}(\xi)$

in our case, $\quad B_1 u_j = -\nu u_{j+1} + \nu u_{j-1}, \quad B_2 = I$

$$G_1(\xi) = -2i\nu \sin\xi, \quad G_2(\xi) = 1$$

now let's freeze ξ and suppress it in the notation. We need $\hat{u}^0, \hat{u}^1$ to get the recursion going. After that,

$$\hat{u}^{n+1} = G_1 \hat{u}^n + G_2 \hat{u}^{n-1}$$

This recursion may be solved in terms of the roots r_1, r_2 of the polynomial $\quad p(r) = r^2 - G_1 r - G_2$, i.e. $r_{1,2} = \dfrac{G_1 \pm \sqrt{G_1^2 + 4G_2}}{2}$

For leapfrog, $\quad r_{1,2} = \dfrac{-2i\nu \sin\xi \pm \sqrt{-4\nu^2 \sin^2\xi + 4}}{2} = \pm\sqrt{1 - \nu^2 \sin^2\xi} - i\nu \sin\xi$

if $r_1 \neq r_2$, the solution of this recursion is

$$\hat{u}^n = c_1 r_1^n + c_2 r_2^n$$

$\hat{u}^n \leftarrow$ superscript an index

$r_j^n \leftarrow$ superscript a power

to match the initial conditions, we need

$$c_1 + c_2 = \hat{u}^0 \qquad c_1 r_1 + c_2 r_2 = \hat{u}^1 \qquad \text{or} \qquad \begin{pmatrix} 1 & 1 \\ r_1 & r_2 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$$

which gives $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{r_2 - r_1}\begin{pmatrix} r_2 & -1 \\ -r_1 & 1 \end{pmatrix}\begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$

or $$\hat{u}^n = (r_1^n, r_2^n)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{r_2 - r_1}\left(r_1^n r_2 - r_1 r_2^n,\ r_2^n - r_1^n\right)\begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$$

$$= -\hat{u}^0 \frac{r_1 r_2 (r_1^{n-1} - r_2^{n-1})}{r_1 - r_2} + \hat{u}^1 \frac{r_1^n - r_2^n}{r_1 - r_2}$$

in the limit as $r_1 \to r_2$, we can use $r_1^n - r_2^n = (r_1 - r_2)\sum_{k=0}^{n-1} r_1^k r_2^{n-1-k}$ to obtain:

$$\hat{u}^n = -\hat{u}^0 r_1 r_2 \sum_{k=0}^{n-2} r_1^k r_2^{n-2-k} + \hat{u}^1 \sum_{k=0}^{n-1} r_1^k r_2^{n-1-k}$$

$$\underset{\substack{\uparrow \\ \text{in limit} \\ \text{as } r_1 \to r_2}}{=} -(n-1) r_1^n \hat{u}_0 + n r_1^{n-1} \hat{u}_1$$

We also could have derived this result directly:

general solution when $r_1 = r_2$:

$$\hat{u}^n = c_1 r_1^n + c_2 n r_1^{n-1} = \begin{cases} c_1 & n=0 \\ c_1 r_1 + c_2 & n=1 \\ c_1 r_1^n + c_2 n r_1^{n-1} & n \geq 2 \end{cases}$$

the initial conditions yield:

$$\begin{pmatrix} 1 & 0 \\ r_1 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix} \rightarrow \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -r_1 & 1 \end{pmatrix} \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$$

or $$\hat{u}^n = \hat{u}^0 r_1^n + \left(-r_1 \hat{u}^0 + \hat{u}^1\right) n r_1^{n-1}$$

$$= -(n-1) r_1^n \hat{u}^0 + n r_1^{n-1} \hat{u}^1$$

Summary: for the leapfrog scheme, the Fourier coefficients $\hat{u}^n(\xi)$ evolve acording to a two step recurrence (i.e. difference equation) $$\hat{u}^{n+1} = G_1 \hat{u}^n + G_2 \hat{u}^{n+1}$$

This equation can be solved in terms of the roots r_1, r_2 of the polynomal $p(r) = r^2 - G_1 r - G_2$.

The solution $\hat{u}^n$ remains bounded for all n and all initial conditions $\hat{u}^0, \hat{u}^1$ iff p satisfies the root condition

$$|r_1| \leq 1, \quad |r_2| \leq 1, \quad \text{if } r_1 = r_2 \text{ then } |r_1| < 1$$

Last time:

1. The centered scheme is unstable even though it satisfies CFL

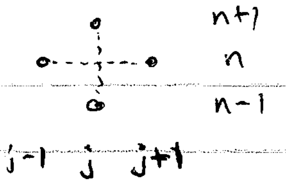
2. it can be made stable by choosing a different refinement path, but the resulting method is expensive and inaccurate (solutions still grow exponentially in time)

3. analyzed Lax-Friedrichs, Lax-Wendroff, Crank-Nicolson

4. introduced leapfrog scheme

Today: analyze Leapfrog scheme.

Leapfrog scheme: $\qquad u_t = -a u_x$



$$\frac{u_j^{n+1}-u_j^{n-1}}{2k} = -a\,\frac{u_{j+1}^n - u_{j-1}^n}{2h} \longleftarrow \quad D_t^0 u_j^n = -a\, D_x^0 u_j^n$$

$$u_j^{n+1} = u_j^{n-1} - \nu u_{j+1}^n + \nu u_{j-1}^n \qquad \left(\text{example of a multistep method}\right)$$

how does the Fourier transform evolve?

$$\sum_j u_j^{n+1} e^{-ij\xi} = \sum_j \left(u_j^{n-1} - \nu u_{j+1}^n + \nu u_{j-1}^n\right) e^{-ij\xi}$$

$$\hat{u}^{n+1}(\xi) = \hat{u}^{n-1}(\xi) - \nu e^{i\xi}\hat{u}^n(\xi) + \nu e^{-i\xi}\hat{u}^n(\xi)$$

$$= \hat{u}^{n-1}(\xi) - (2i\nu \sin\xi)\,\hat{u}^n(\xi)$$

More generally, if $\quad u^{n+1} = B_1 u^n + B_2 u^{n-1}$

then $\quad \hat{u}^{n+1}(\xi) = G_1(\xi)\hat{u}^n(\xi) + G_2(\xi)\hat{u}^{n-1}(\xi)$

in our case, $\quad B_1 u_j = -\nu u_{j+1} + \nu u_{j-1}, \qquad B_2 = I$

$$G_1(\xi) = -2i\nu\sin\xi, \qquad G_2(\xi) = 1$$

now let's freeze ξ and suppress it in the notation. We need $\hat{u}^0, \hat{u}^1$ to get the recursion going. After that,

$$\hat{u}^{n+1} = G_1\hat{u}^n + G_2\hat{u}^{n-1}$$

This recursion may be solved in terms of the roots r_1, r_2 of the polynomial $\quad p(r) = r^2 - G_1 r - G_2$, i.e. $r_{1,2} = \dfrac{G_1 \pm \sqrt{G_1^2 + 4G_2}}{2}$

For leapfrog, $\quad r_{1,2} = \dfrac{-2i\nu\sin\xi \pm \sqrt{-4\nu^2\sin^2\xi + 4}}{2} = \pm\sqrt{1-\nu^2\sin^2\xi} - i\nu\sin\xi$

if $r_1 \neq r_2$, the solution of this recursion is

$$\hat{u}^n = c_1 r_1^n + c_2 r_2^n$$

$\hat{u}^n \leftarrow$ superscript an index

$r_j^n \leftarrow$ superscript a power

to match the initial conditions, we need

$$c_1 + c_2 = \hat{u}^0 \qquad c_1 r_1 + c_2 r_2 = \hat{u}^1 \quad \text{or} \quad \begin{pmatrix} 1 & 1 \\ r_1 & r_2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$$

which gives $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{r_2 - r_1} \begin{pmatrix} r_2 & -1 \\ -r_1 & 1 \end{pmatrix} \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$

or $$\hat{u}^n = (r_1^n, r_2^n) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{r_2 - r_1} \left(r_1^n r_2 - r_1 r_2^n ,\ r_2^n - r_1^n \right) \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$$

$$= -\hat{u}^0 \frac{r_1 r_2 (r_1^{n-1} - r_2^{n-1})}{r_1 - r_2} + \hat{u}^1 \frac{r_1^n - r_2^n}{r_1 - r_2}$$

in the limit as $r_1 \to r_2$, we can use $r_1^n - r_2^n = (r_1 - r_2) \sum_{k=0}^{n-1} r_1^k r_2^{n-1-k}$

to obtain:

$$\hat{u}^n = -\hat{u}^0 r_1 r_2 \sum_{k=0}^{n-2} r_1^k r_2^{n-2-k} + \hat{u}^1 \sum_{k=0}^{n-1} r_1^k r_2^{n-1-k}$$

$$\underset{\substack{\uparrow \\ \text{in limit} \\ \text{as } r_1 \to r_2}}{=} -(n-1) r_1^n \hat{u}_0 + n r_1^{n-1} \hat{u}_1$$

We also could have derived this result directly:

general solution when $r_1 = r_2$:

$$\hat{u}^n = c_1 r_1^n + c_2 n r_1^{n-1} = \begin{cases} c_1 & n=0 \\ c_1 r_1 + c_2 & n=1 \\ c_1 r_1^n + c_2 n r_1^{n-1} & n \geq 2 \end{cases}$$

the initial conditions yield:

$$\begin{pmatrix} 1 & 0 \\ r_1 & 1 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix} \rightarrow \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -r_1 & 1 \end{pmatrix}\begin{pmatrix} \hat{u}^0 \\ \hat{u}^1 \end{pmatrix}$$

or
$$\hat{u}^n = \hat{u}^0 r_1^n + \left(-r_1 \hat{u}^0 + \hat{u}^1\right) n r_1^{n-1}$$
$$= -(n-1) r_1^n \hat{u}^0 + n r_1^{n-1} \hat{u}^1 \quad \leftarrow \text{multiple roots lead to polynomial growth.}$$

<u>Summary</u>: for the leapfrog scheme, the Fourier coefficients $\hat{u}^n(\xi)$ evolve according to a two step recurrence (i.e. difference equation)
$$\hat{u}^{n+1} = G_1 \hat{u}^n + G_2 \hat{u}^{n+1}$$

This equation can be solved in terms of the roots r_1, r_2 of the polynomial $p(r) = r^2 - G_1 r - G_2$.

The solution $\hat{u}^n$ remains bounded for all n and all initial conditions $\hat{u}^0, \hat{u}^1$ iff p satisfies the root condition

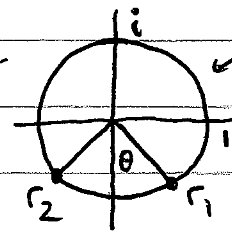
$$|r_1| \leq 1, \quad |r_2| \leq 1, \quad \text{if } r_1 = r_2 \text{ then } |r_1| < 1$$

The question we really care about is: $\left(\begin{array}{c} \text{how big is } \|\hat{u}^n\|_{L^2} \\ \text{for } 0 \le nk \le T \text{ ?} \end{array} \right)$

so we have to look more closely at the recursion for different choices of ξ.

For the leapfrog scheme $\left\{ \begin{array}{l} r_1 = \sqrt{1-\nu^2\sin^2\xi} - i\nu\sin\xi \\ r_2 = -\sqrt{1-\nu^2\sin^2\xi} - i\nu\sin\xi \end{array} \right\}$ $\left(\begin{array}{c} |\nu|>1 \\ \Rightarrow \\ r_1 \text{ or } r_2 \\ \text{outside} \\ \text{unit circle} \\ \text{for} \\ \xi = \frac{\pi}{2} \end{array} \right)$

if $|\nu| \le 1$



$$r_1 = -ie^{i\theta} \qquad \sin\theta = \sqrt{1-\nu^2\sin^2\xi}$$
$$r_2 = -ie^{-i\theta}$$

we saw that $\displaystyle \hat{u}^n = -\hat{u}^0 \frac{r_1 r_2 (r_1^{n-1} - r_2^{n-1})}{r_1 - r_2} + \hat{u}^1 \frac{r_1^n - r_2^n}{r_1 - r_2}$

$$r_1^n - r_2^n = (-i)^n \left(e^{in\theta} - e^{-in\theta} \right)$$
$$= (-i)^n (2i) \frac{e^{in\theta} - e^{-in\theta}}{2i} = 2(-i)^{n-1} \sin n\theta$$

$$r_1 - r_2 = (-i)(2i) \frac{e^{i\theta} - e^{-i\theta}}{2i} = 2\sin\theta$$

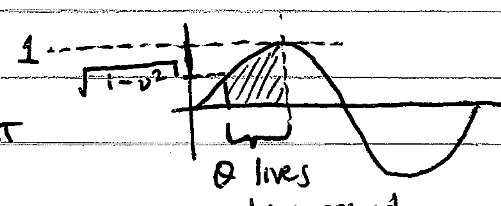
so

$$\hat{u}^n = -\hat{u}^0 \frac{\overbrace{(-1)}^{r_1 r_2}(2)(-i)^{n-2}\sin((n-1)\theta)}{2\sin\theta} + \hat{u}^1 \frac{2(-i)^{n-1}\sin n\theta}{2\sin\theta}$$

$$= -(-i)^n \frac{\sin((n-1)\theta)}{\sin\theta} \hat{u}^0 + (-i)^{n-1} \frac{\sin n\theta}{\sin\theta} \hat{u}^1$$

note that $\sin\theta = \sqrt{1-\nu^2\sin^2\xi}$ lies in the range

$$\underbrace{\sqrt{1-\nu^2}}_{\xi = \frac{\pi}{2}} \le \sin\theta \le \underset{\uparrow \, \xi = 0, \pm\pi}{1}$$



$$\sin\theta \geq \sqrt{1-\nu^2}$$

if $|\nu| < 1$, then $\left|\frac{\sin n\theta}{\sin\theta}\right| \leq \frac{1}{|\sin\theta|} \leq \frac{1}{\sqrt{1-\nu^2}}$

So

$$|\hat{u}^n(\xi)| \leq \frac{1}{\sqrt{1-\nu^2}}\left(|\hat{u}^0(\xi)| + |\hat{u}^1(\xi)|\right)$$

$$|\hat{u}(\xi)|^2 \leq \frac{2}{1-\nu^2}\left(|\hat{u}^0(\xi)|^2 + |\hat{u}^1(\xi)|^2\right)$$

here we used $(a+b)^2 = a^2 + 2ab + b^2 \leq 2(a^2+b^2)$

$0 \leq (a-b)^2 = a^2 - 2ab + b^2 \Rightarrow 2ab \leq a^2 + b^2$

$$\therefore \quad \|\hat{u}^n\|^2_{L^2(-\pi,\pi)} = \int_{-\pi}^{\pi} |\hat{u}(\xi)|^2 d\xi \leq \frac{2}{1-\nu^2}\left[\int_{-\pi}^{\pi} |\hat{u}^0(\xi)|^2 d\xi + \int_{-\pi}^{\pi} |\hat{u}^1(\xi)|^2 d\xi\right]$$

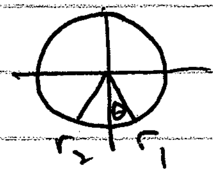
$$\therefore \quad \|\hat{u}^n\|_{L^2} \leq \sqrt{\frac{2}{1-\nu^2}}\sqrt{\|\hat{u}^0\|^2_{L^2} + \|\hat{u}^1\|^2_{L^2}} \leq \sqrt{\frac{2}{1-\nu^2}}\left(\|\hat{u}^0\| + \|\hat{u}^1\|\right)$$

$$\therefore \quad \|u^n\|_{2,h} \leq \sqrt{\frac{2}{1-\nu^2}}\left(\|u^0\|_{2,h} + \|u^1\|_{2,h}\right)$$

uniform bound on growth of solution which does not blow up as $k,h \to 0$

$\therefore$ scheme is stable for $|\nu| < 1$, but the error bound gets worse and worse as ν approachs 1.

so what happens if $|\nu| = 1$? The roots r_1, r_2 still live on the unit circle, but when $\xi = \pi/2$ we get a double root, which spells trouble...



$$\sin\theta = \sqrt{1-\nu^2\sin^2\xi} = 0 \quad \text{when } \xi = \frac{\pi}{2},\ |\nu| = 1$$

This time we'll estimate $\left|\frac{\sin n\theta}{\sin\theta}\right|$ using:

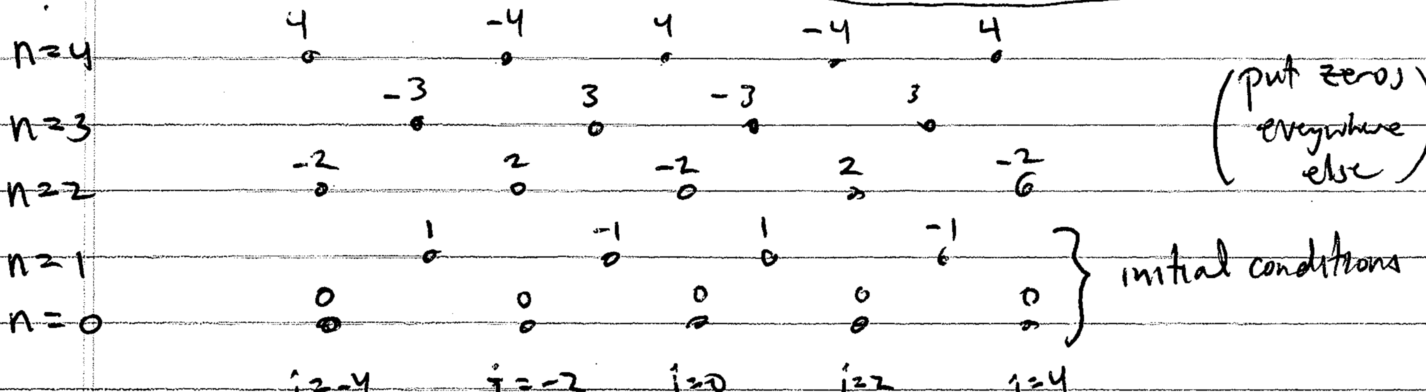
$$\begin{aligned}
|\sin n\theta| &= \left|\sin((n-1)\theta)\cos\theta + \sin\theta\cos((n-1)\theta)\right| \\
&\leq |\sin((n-1)\theta)| + |\sin\theta| \\
&\leq |\sin(n-2)\theta| + 2|\sin\theta| \\
&\ \ \vdots \\
&\leq n|\sin\theta|
\end{aligned}$$

so $$\left|\frac{\sin n\theta}{\sin\theta}\right| \leq \min\left(\frac{1}{\sqrt{1-\nu^2\sin^2\xi}},\ n\right)$$

if we know nothing about the initial data, we can't do much better than

$$\|u^n\|_{2,h} \leq \sqrt{2}\,n\left(\|u^0\|_{2,h} + \|u^1\|_{2,h}\right)$$

and in fact, this linear growth with the number of timesteps does actually happen stencil with $\nu = 1$: $\boxed{u_j^{n+1} = u_j^{n-1} + u_{j-1}^n + u_{j+1}^n}$



on a periodic lattice, we would get $\|u^n\|_{2,h} = n\left(\|u^0\|_{2,h} + \|u^1\|_{2,h}\right)$

$\therefore$ scheme is unstable, but the instability is fairly mild. Since this problem is linear, you could borrow a factor of k from the truncation error to turn n into $nk \le T$, so the convergence tests might actually indicate that the method is $O(k + h^2)$ rather than unstable.

Also, only modes close to $\frac{\pi}{2}$ will get amplified indefinitely, so if the initial condition happens to satisfy $\hat{u}^0(\xi) = 0$, $\hat{u}^1(\xi) = 0$ for ξ close to $\pi/2$, you won't see the instability (note that initial conditions of the form $g(x) = \sin 2\pi x$ or $\sin 4\pi x$ or $\sin 2N\pi x$ for some fixed N have this property for a fine enough mesh).

Finally, if $g(x)$ is a periodic, analytic function, its fourier coefficients will decay exponentially (eventually), and so for small enough h the magnitude of $\hat{u}^0(\xi)$ for ξ near $\pi/2$ will decay as you refine the mesh, possibly saving the instability of the leapfrog scheme with $|\nu| = 1$.

☆ 228B Lec 17 ☆

Last time: Analysis of leapfrog method (and other 2-step schemes)

Today: (1) spectrally accurate differentiation & integration of periodic functions
(2) schemes for hyperbolic systems
(3) boundary conditions for hyperbolic systems.

Discrete Fourier transform (in matlab: fft, ifft. beware of "off by 1" errors)

$$w_k = \sum_{j=0}^{N-1} e^{-2\pi i jk/N} u_j \qquad w_{k+N} = w_k, \quad k \in \mathbb{Z}$$

$$u_j = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i jk/N} w_k = \sum_{k=-N/2}^{\frac{N}{2}-1} e^{2\pi i jk/N} w_k$$

most sensible range of indices

Now suppose $u_j = u(x_j)$, $x_j = jh = \frac{jL}{N}$

j: 0 1 2 … N

x: 0 h $2h$ … L

Then $u(x_j) = \frac{1}{N}\sum_k e^{ij\frac{2\pi k}{N}} w_k = \frac{1}{N}\sum_k e^{i\left(\frac{x_j}{h}\right)\xi_k} w_k$, $\xi_k = \frac{2\pi k}{N}$

interpolate between sampled points using same formula ($x_j \to x$)

spectrally accurate differentiation and integration formulas:

$$u'(x) = \frac{1}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} e^{i\left(\frac{x}{h}\right)\xi_k}\left(i\frac{\xi_k}{h}w_k\right), \quad \int_0^x u(s)\,ds = \frac{1}{N}\sum_{k\neq 0} e^{i\frac{x}{h}\xi_k}\left(\frac{w_k}{i\,\xi_k/h}\right)$$

↑ best to zero out Nyquist frequency $w_{N/2} = 0$

↑ need $w_0 = 0$ for this to make sense

schemes for hyperbolic systems $\left(\nu = \frac{k}{h},\ -a \to A,\ \text{real distinct eigenvalues}\right)$

goal: solve $\vec{u}_t = A\vec{u}_x$ where $\vec{u}$ is a vector

Lax-Friedrichs $\quad \vec{u}_j^{n+1} = \frac{1}{2}(I - \nu A)\vec{u}_{j-1}^n + \frac{1}{2}(I + \nu A)\vec{u}_{j+1}^n$

Lax-Wendroff $\quad \vec{u}_j^{n+1} = \vec{u}_j^n + \nu A\left(\frac{\vec{u}_{j+1}^n - \vec{u}_{j-1}^n}{2}\right) + \frac{\nu^2}{2}A^2\left(\vec{u}_{j-1}^n - 2\vec{u}_j^n + \vec{u}_{j+1}^n\right)$

Leapfrog $\quad \vec{u}_j^{n+1} = \vec{u}_j^{n-1} - \nu A\vec{u}_{j-1}^n + \nu A\vec{u}_{j+1}^n$

The Fourier analysis of these schemes is similar to the scalar case, but the amplification factor $G(\xi)$ becomes an amplification matrix (still called $G(\xi)$)

Lax-Friedrichs: $\quad \hat{\vec{u}}^{n+1}(\xi) = G(\xi)\hat{\vec{u}}^n(\xi),\quad G(\xi) = \frac{1}{2}(I - \nu A)e^{-i\xi} + \frac{1}{2}(I + \nu A)e^{i\xi}$

$$= (\cos\xi)I + i\nu(\sin\xi)A$$

Lax-Wendroff: $\quad G(\xi) = I + i\nu(\sin\xi)A + \frac{\nu^2}{2}\left(e^{-i\xi} - 2 + e^{i\xi}\right)A^2$

$$= I - 2\nu^2\sin^2(\xi/2)A^2 + i\nu(\sin\xi)A$$

Leapfrog: $\quad \hat{\vec{u}}^{n+1}(\xi) = G_1(\xi)\hat{\vec{u}}^n(\xi) + G_2(\xi)\hat{\vec{u}}^{n-1}(\xi)$

$$G_1(\xi) = 2i\nu(\sin\xi)A, \quad G_2(\xi) = I$$

These amplification matrices are diagonalized along with A:

$$A = U\Lambda U^{-1} \Rightarrow G(\xi) = U\begin{pmatrix} G(\xi,\lambda_1) & & \\ & \ddots & \\ & & G(\xi,\lambda_N)\end{pmatrix}U^{-1}$$

scaler amplification factors for $u_t = \lambda_j u_x$ (scheme applied to)

so for fixed ξ,

$$|\hat{\vec{u}}^n(\xi)| \le \|G(\xi)^n\| \cdot |\hat{\vec{u}}^0(\xi)|$$

$$\|G(\xi)^n\| \le \|U\| \cdot \|U^{-1}\| \cdot \max_{1\le \ell \le N} |G(\xi,\lambda_\ell)|^n$$

and our 2-norm estimate looks like

$$\int_{-\pi}^{\pi} |\hat{\vec{u}}^n(\xi)|^2 \, d\xi \le \left(\max_{-\pi\le\xi\le\pi} \|G(\xi)^n\|^2\right)\int_{-\pi}^{\pi} |\hat{\vec{u}}^0(\xi)|^2 \, d\xi$$

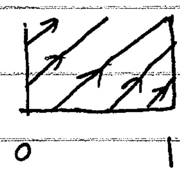
or $$\|\vec{u}^n\|_{2,h} \le \max_{1\le \ell\le N}\left(\max_{-\pi\le\xi\le\pi} |G(\xi,\lambda_\ell)|^n\right)\|U\|\cdot\|U^{-1}\|\cdot\|\vec{u}^0\|_{2,h}$$

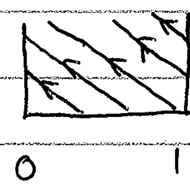
so stability boils down to the scalar scheme applied to each eigenvalue separately. The same is true for the recursion involved in the leapfrog scheme.

Boundary conditions

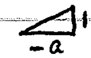
periodic b/c's are easy to implement, but Dirichlet & Neumann conditions are tricky for wave equations.

<u>scalar equation:</u> $u_t = a u_x$





solution constant along lines $x + at = \text{const}$

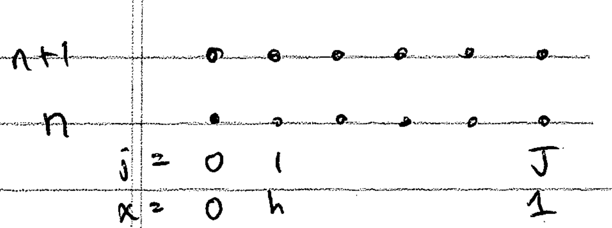


$a<0$ $\quad$ $a>0$

$a<0$: need a b.c. on the left wall
illegal to impose one on the right

$a>0$: need one on right wall, illegal on left.

example: solve $u_t = -a u_x$, $a>0$, $u(x,0) = g(x)$
using Lax-Wendroff $\quad u(0,t) = f(t)$



<u>scheme:</u> $u_0^{n+1} = f(t_{n+1})$

$$u_j^{n+1} = u_j^n - \nu a\left(\frac{u_{j+1}^n - u_{j-1}^n}{2}\right) + \frac{\nu^2 a^2}{2}\overbrace{\cancel{\left(u_{j-1}^n + u_j^n + u_{j+1}^n\right)}}^{u_{j-1}^n - 2u_j^n + u_{j+1}^n}$$

$1 \le j \le J-1$

$u_J^{n+1} = ?$

most common choice: just use upwind on right bdry: $u_J^{n+1} = (1 - a\nu)u_J^n + a\nu u_{J-1}^n$

<u>in matrix form:</u>

$$u^{n+1} = B u^n + \tilde{B} f^n \qquad B = \begin{pmatrix} \alpha & \gamma & & & \\ \beta & \alpha & \gamma & & \\ & \ddots & \ddots & \ddots & \\ & & \beta & \alpha & \gamma \\ & & & \kappa & \theta \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} \beta \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

$\alpha = 1 - a^2\nu^2$
$\beta = \frac{1}{2} a\nu(1 + a\nu)$
$\gamma = \frac{1}{2} a\nu(-1 + a\nu)$
$\kappa = a\nu$
$\theta = 1 - a\nu$

all I see that is easy to prove is $\|B\|_\infty = 1$.

it's not ~~clear~~ obvious whether introducing a first order error at the right endpoint will wreck the 2nd order convergence of Lax-Wendroff.

Another option would be $u_J^{n+1} = c_2 u_{J-2} + c_1 u_{J-1} + c_0 u_J$

choosing the coefficients c_0, c_1, c_2 to match Taylor coefficients:

$$\underbrace{u + k u_t}_{-au_x} + \underbrace{\frac{k^2}{2} u_{tt}}_{a^2 u_{xx}} = c_2\left(u - (2h) u_x + \frac{(2h)^2}{2} u_{xx}\right) + c_1\left(u - h u_x + \frac{h^2}{2} u_{xx}\right) + c_0 u$$

$$(1 - c_0 - c_1 - c_2) u + (-ka + hc_1 + 2hc_2) u_x + \left(\frac{k^2}{2} a^2 - \frac{h^2}{2} c_1 - 2h^2 c_2\right) u_{xx} = 0$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ a\nu \\ a^2\nu^2 \end{pmatrix} \Rightarrow \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & -3/2 & 1/2 \\ 0 & 2 & -1 \\ 0 & -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ a\nu \\ a^2\nu^2 \end{pmatrix}$$

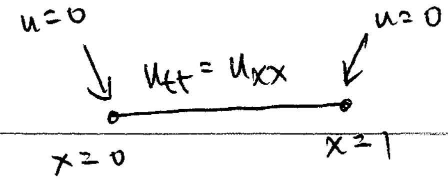
or $$u_J^{n+1} = -\frac{1}{2} a\nu(1-a\nu) u_{J-2} + a\nu(2-a\nu) u_{J-1} + \left(1 - \frac{a\nu}{2}(3-a\nu)\right) u_J$$

But I suspect this scheme is unstable, i.e. B has an eigenvalue $|\lambda_j| > 1$:

$$B = \begin{pmatrix} \alpha & \gamma & & & \\ \beta & \alpha & \gamma & & \\ & \ddots & \ddots & \ddots & \\ & & \beta & \alpha & \gamma \\ & & c_2 & c_1 & c_0 \end{pmatrix}$$

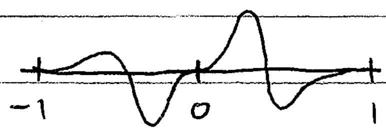
So boundary conditions for wave equations are difficult to deal with. Fortunately, for most physically important problems you can use symmetry to derive the correct B.C.'s to use.

vibrating string



$$\left.\begin{aligned} u(x,0) &= g_0(x) \\ u_t(x,0) &= g_1(x) \end{aligned}\right\} \text{given initial conditions}$$

idea: turn the Dirichlet conditions into periodic conditions:



← new problem: $\begin{cases} u_{tt} = u_{xx} \\ u(1,t) = u(-1,t) \\ u_x(1,t) = u_x(-1,t) \\ u(x,0) = g_0(x) & -1 \le x \le 1 \\ u_t(x,0) = g_1(x) & -1 \le x \le 1 \end{cases}$

for $x < 0$, we define

$$g_0(x) = -g_0(-x)$$
$$g_1(x) = -g_1(-x)$$

whatever the solution $u(x,t)$ of this problem is, the function

$$v(x,t) = -u(-x,t) \qquad -1 \le x \le 1, \ t \ge 0$$

is also a solution: $v_{tt} = -u_{tt}(-x,t)$

$$v_x = u_x(-x,t)$$
$$v_{xx} = -u_{xx}(-x,t) = -u_{tt}(-x,t) = v_{tt}$$

since u & v satisfy the same (periodic) b.c.'s and the same initial conditions, they are equal (uniqueness of solutions).

So $u(x,t) = -u(-x,t) \qquad -1 \le x \le 1, \ t \ge 0$

in particular: $u(0,t) = -u(0,t) \Rightarrow u(0,t) = 0$

$$u(1,t) = -u(-1,t) \underset{\substack{\uparrow \\ \text{periodicity}}}{=} -u(1,t) \Rightarrow u(1,t) = 0$$

So the new problem gives the soln to the original problem.

Next we want to figure out b.c.'s to impose on the original problem to mimic the periodic problem without actually computing any values at $x_j < 0$

1st order system: $\vec{w} = \begin{pmatrix} u_t \\ u_x \end{pmatrix}$, $\vec{w}_t = A\vec{w}_x$, $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

for the periodic system, the Lax-Wendroff update at $j=0$ would be

$$\vec{w}_0^{n+1} = \vec{w}_0^n + \nu A\left(\frac{\vec{w}_1^n - \vec{w}_{-1}^n}{2}\right) + \frac{\nu^2}{2}\underbrace{A^2}_{I}\left(w_{-1}^n - 2\vec{w}_0^n + \vec{w}_1^n\right)$$

$$= (1-\nu^2)\vec{w}_0^n + \nu A\left(\frac{\vec{w}_1^n - \vec{w}_{-1}^n}{2}\right) + \nu^2\left(\frac{\vec{w}_1^n + \vec{w}_{-1}^n}{2}\right)$$

But $\vec{w}_{-1}^n = \begin{pmatrix} u_t(-h,t_n) \\ u_x(-h,t_n) \end{pmatrix} = \begin{pmatrix} -u_t(h,t_n) \\ u_x(h,t_n) \end{pmatrix}$, $w_1^n = \begin{pmatrix} u_t(h,t_n) \\ u_x(h,t_n) \end{pmatrix}$

$u(x,t) = -u(-x,t)$

so

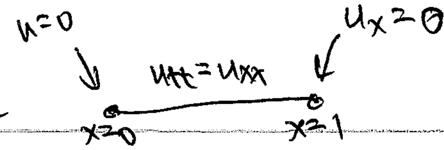
$$\vec{w}_0^{n+1} = (1-\nu^2)\vec{w}_0^n + \nu A\begin{pmatrix} u_t(h,t_n) \\ 0 \end{pmatrix} + \nu^2\begin{pmatrix} 0 \\ u_x(h,t_n) \end{pmatrix}$$

$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

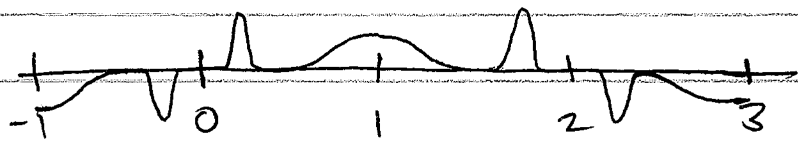
$$\boxed{\vec{w}_0^{n+1} = (1-\nu^2)\vec{w}_0^n + \begin{pmatrix} 0 & 0 \\ \nu & \nu^2 \end{pmatrix}\vec{w}_1^n}$$

At the right endpoint, a similar analysis gives

$$\boxed{\vec{w}_J^{n+1} = (1-\nu^2)\vec{w}_J^n + \begin{pmatrix} 0 & 0 \\ -\nu & \nu^2 \end{pmatrix}\vec{w}_{J-1}^n}$$

For the case 

we would actually extend by odd symmetry about the origin and even symmetry about $x=1$ to get a periodic domain 4 times as big



$$u(3,t)=u(-1,t)$$

$$u(x,t) = -u(-x,t) \qquad -1 \le x \le 0$$

$$u(x,t) = u(2-x,t) \qquad 1 \le x \le 3 \quad \leftarrow \text{or } u(1+x,t)=u(1-x,t) \quad 0 \le x \le 2$$

you only actually want to compute values of u (or w) between $0 \le x \le 1$

at 0: proceed as before: $\vec{w}_0^{\,n+1} = (1-\nu^2)\vec{w}_0^{\,n} + \begin{pmatrix} 0 & 0 \\ \nu & \nu^2 \end{pmatrix} \vec{w}_1^{\,n}$

at 1: $\vec{w}_J^{\,n+1} = (1-\nu^2)\vec{w}_J^{\,n} + \nu A\left(\dfrac{\vec{w}_{J+1}^{\,n} - \vec{w}_{J-1}^{\,n}}{2}\right) + \nu^2\left(\dfrac{\vec{w}_{J+1}^{\,n} + \vec{w}_{J-1}^{\,n}}{2}\right)$

$$w_{J+1}^n = \begin{pmatrix} u_t(1+h,t_n) \\ u_x(1+h,t_n) \end{pmatrix} \underset{\substack{\uparrow \\ u(1+x,t)=u(1-x,t)}}{=} \begin{pmatrix} u_t(1-h,t_n) \\ -u_x(1-h,t_n) \end{pmatrix}, \qquad w_{J-1} = \begin{pmatrix} u_t(1-h,t_n) \\ u_x(1-h,t_n) \end{pmatrix}$$

so

$$\vec{w}_J^{\,n+1} = (1-\nu^2)\vec{w}_J^{\,n} + \nu\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ -u_x(1-h,t_n) \end{pmatrix} + \nu^2\begin{pmatrix} u_t(1-h,t_n) \\ 0 \end{pmatrix}$$

$$\boxed{w_J^{n+1} = (1-\nu^2)\vec{w}_J^{\,n} + \begin{pmatrix} \nu^2 & -\nu \\ 0 & 0 \end{pmatrix} \vec{w}_{J-1}^{\,n}}$$

228B Lec 18

Dissipation and Dispersion

$u_t = -a u_x$, initial condition $u(x,0) = g(x) = e^{i\frac{x}{h}\xi}$

exact solution: $u(x,t) = g(x-at) = e^{i\frac{x-at}{h}\xi} = e^{-i\frac{at}{h}\xi} e^{i\frac{x}{h}\xi}$

Sampled on grid: $u(x_j, t_n) = \left(e^{-ia\frac{nk}{h}\xi}\right) e^{ij\xi} = \left(e^{-ia\frac{k}{h}\xi}\right)^n u(x_j, 0)$

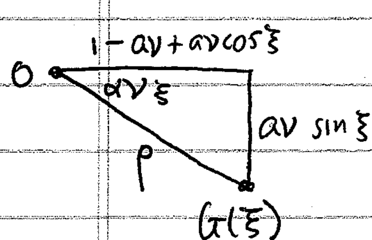
numerical solution: $u_j^n = G(\xi)^n u_j^0 \leftarrow \left\{ \begin{array}{l} \text{special initial data} \\ u_j^0 = e^{ij\xi} \end{array} \right.$

so want $G(\xi) \approx e^{-ia\nu\xi}$, $\nu = \frac{k}{h}$

upwind $(a>0)$: $u_j^{n+1} = a\nu\, u_{j-1}^n + (1-a\nu) u_j^n$

$$G(\xi) = a\nu e^{-i\xi} + (1-a\nu) \qquad e^{-i\xi} = \cos\xi - i\sin\xi$$

$$= \left(1 - a\nu + a\nu\cos\xi\right) - i a\nu \sin\xi$$



$$= \rho e^{-i\alpha\nu\xi}$$

$$\rho^2 = |G(\xi)|^2 = (1-a\nu+a\nu\cos\xi)^2 + (a\nu)^2\sin^2\xi$$

scheme is dissipative of order 2 $\longrightarrow$ $\rho = \sqrt{1 - 4a\nu(1-a\nu)\sin^2(\xi/2)}$

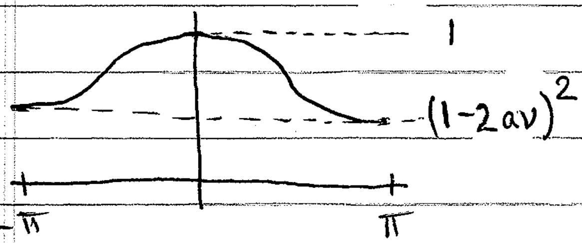
(exact has $\rho = 1$, no damping)

$$\tan(\alpha\nu\xi) = \frac{a\nu\sin\xi}{1 - a\nu(1-\cos\xi)}$$

numerical wave speed $\longrightarrow$ $\dfrac{\alpha}{a} = \dfrac{1}{a\nu\xi} \arctan\left(\dfrac{a\nu\sin\xi}{1-a\nu(1-\cos\xi)}\right)$

exact wave speed $\longrightarrow$

$\rho^2 = 1 - 4a\nu(1-a\nu)\sin^2(\xi/2)$



def:
we say a scheme is dissipative of order $2r$ if $\exists\, C>0$ s.t.

$$|G(\xi)|^2 \le 1 - C\sin^2(\xi/2)$$

for $-\pi \le \xi \le \pi$

so upwind is dissipative of order 2 if $0 < a\nu < 1$

Dispersion refers to the idea that waves with different (spatial) frequencies travel at different speeds.

for upwind, we have:

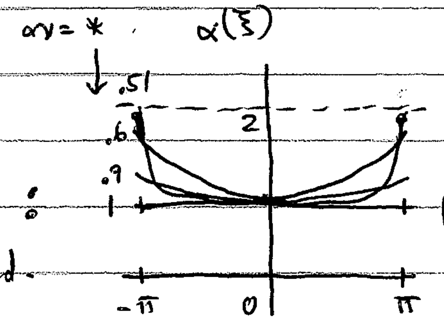
$a\nu > 1$: unstable

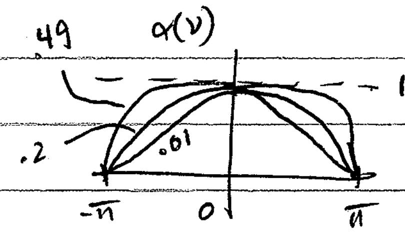
$a\nu = 1$: exact solution $\left(\frac{\alpha}{a} = 1\right)$

$\frac{1}{2} < a\nu < 1$: $\alpha(\xi) > a$ numerical wave speed is faster than true wave speed.

interesting case → $a\nu = \frac{1}{2}$: wave speeds exact $(\alpha(\xi) = 1)$ but dissipation is large

$0 < a\nu < \frac{1}{2}$: $\alpha(\xi) < a$ numerical wave speed slower than exact





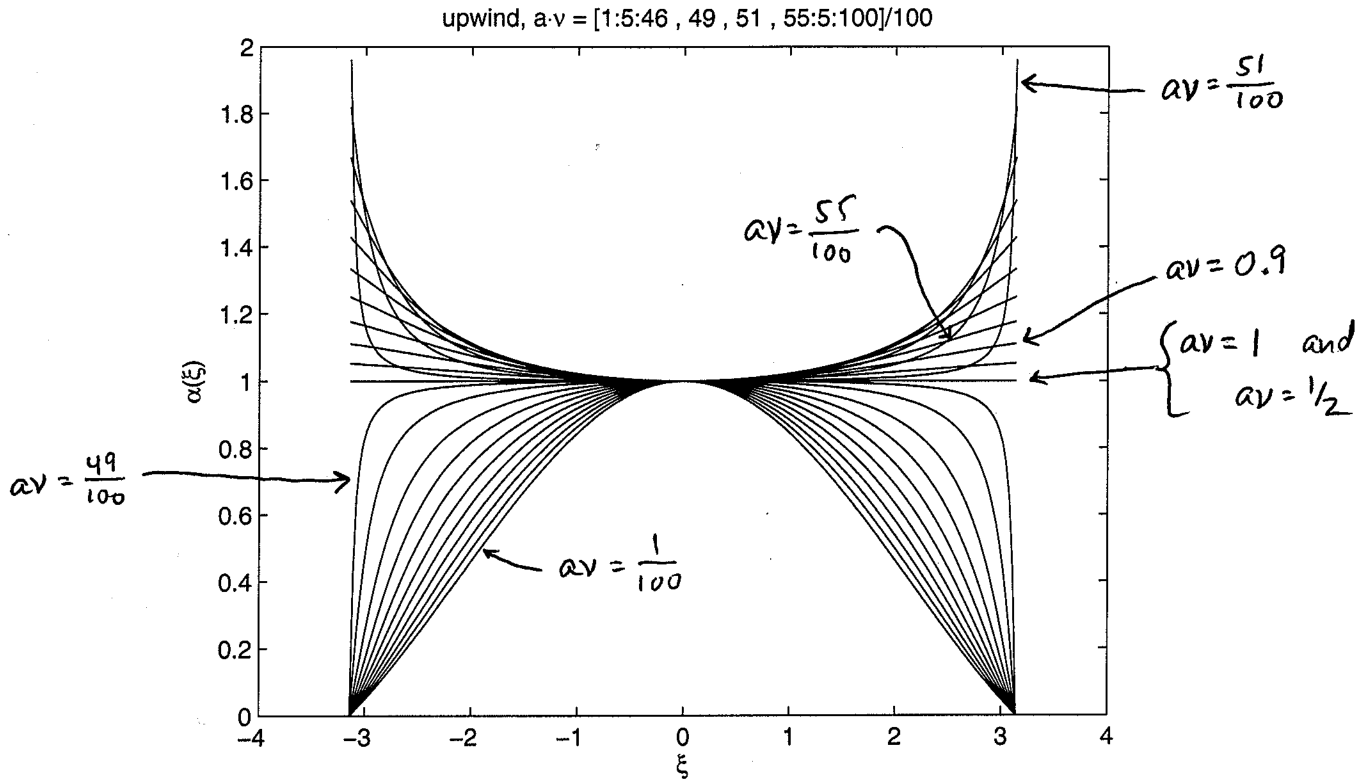
See computer plots for better pictures ↗

when ξ is small, we can Taylor expand $\frac{\alpha(\xi)}{a} = \frac{1}{a\nu\xi}\arctan\left(\frac{a\nu\sin\xi}{1-a\nu(1-\cos\xi)}\right)$

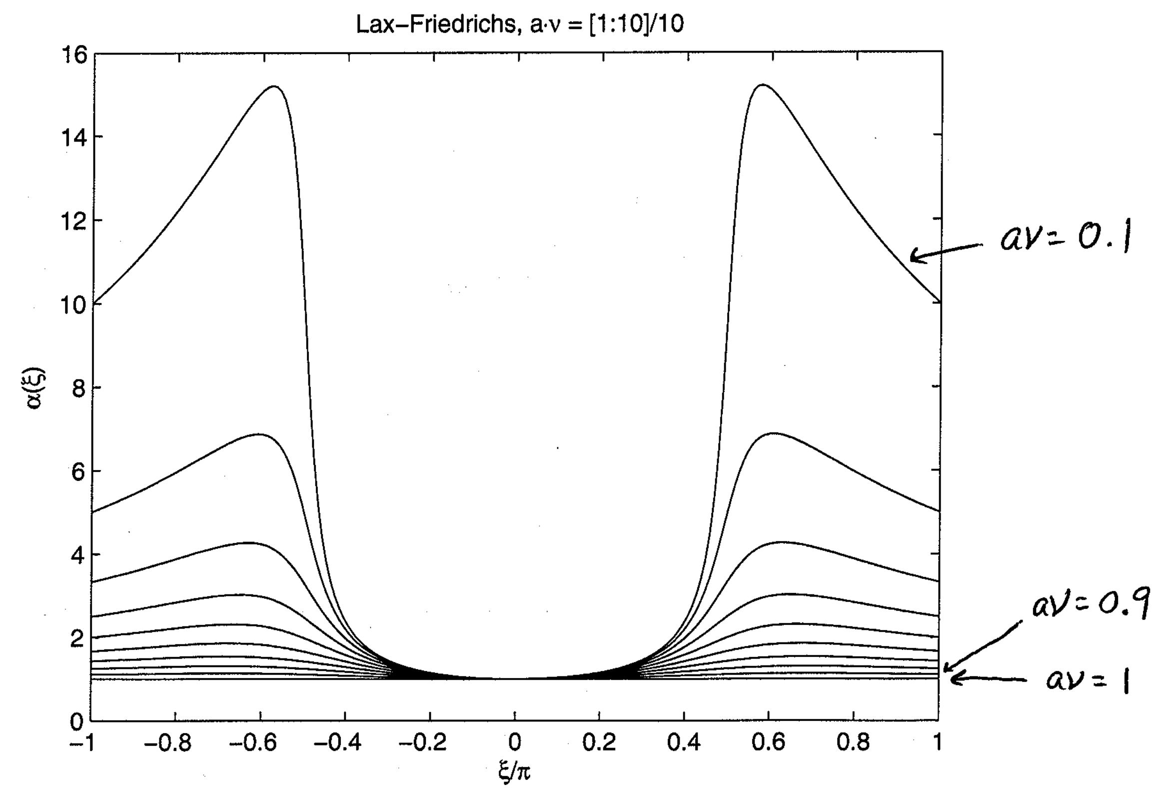
to learn:

$$\frac{\alpha(\xi)}{a} = 1 + \frac{(1-a\nu)(2a\nu-1)}{6}\xi^2 + O(\xi^4)$$

RHS happens to be exactly 1 when $a\nu = 1$, $a\nu = \frac{1}{2}$



upwind, a·ν = [1:5:46 , 49 , 51 , 55:5:100]/100
α(ξ)
ξ
aν = 51/100
aν = 55/100
aν = 0.9
aν = 1 and aν = 1/2
aν = 49/100
aν = 1/100



Lax−Friedrichs, a·ν = [1:10]/10
α(ξ)
ξ/π
aν = 0.1
aν = 0.9
aν = 1

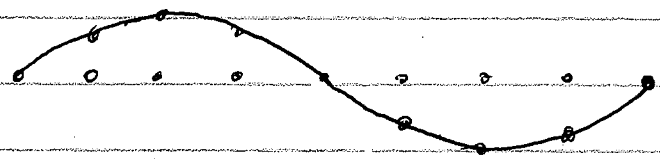
$\kappa x_j = j\xi$

$\xi = \kappa h$

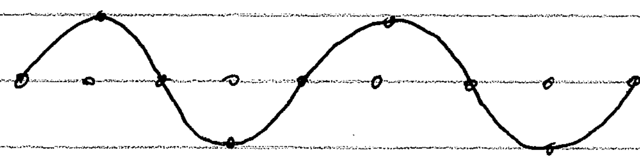
What does a graph plotting ρ or α vs. ξ tell us?

$e^{i(\kappa x - \omega t)}$: κ wave number, ω angular frequency

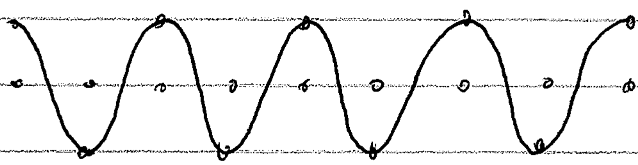
Think of ξ as a wave number relative to the mesh spacing



$\xi = \frac{2\pi}{8} = \frac{\pi}{4}$



$\xi = \frac{2\pi}{4} = \frac{\pi}{2}$



$\xi = \frac{2\pi}{2} = \pi$

On a fixed periodic domain $[0,L]$, the permitted wave numbers are

$$\kappa = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \pm\frac{6\pi}{L}, \dots \pm\frac{2\pi m}{L} \dots$$

once we discretize into J segments of width $h = \frac{L}{J}$, the corresponding "allowable" values of ξ are

$$\xi = \kappa h = 0, \pm\frac{2\pi}{J}, \pm\frac{4\pi}{J}, \dots \pm\frac{2\pi m}{J}$$

The corresponding functions $e^{ij\xi}$ on the grid are linearly independent only for $-\frac{J}{2} \le m \le \frac{J}{2} - 1$. For m outside this range you get aliasing effects (a high frequency mode looks like a lower frequency one)

$\rho = 1 - c\xi^{2r} + O(\xi^{2r+2})$ the bigger the r the better

$\alpha/a = 1 + c\xi^{2r} + O(\xi^{2r+2})$

ξ close to zero: lots of gridpoints to resolve the wave

$\rightarrow$ better get ρ and α right for ξ near zero

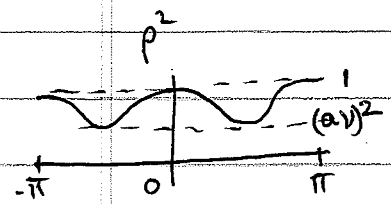
ξ close to $\pm\pi$: waves like this oscillate wildly on the grid. don't expect accuracy here. Good to have $\rho < 1$ to damp them out.

$$u_t = -a u_x$$

Lax-Friedrichs

$$u_j^{n+1} = \tfrac{1}{2}(1 + a\nu)\, u_{j-1}^n + \tfrac{1}{2}(1 - a\nu)\, u_{j+1}^n$$

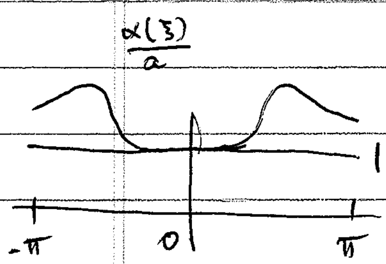
$$G(\xi) = \cos\xi - i\, a\nu \sin\xi = \rho e^{-i\alpha\nu\xi}$$



$$\rho^2 = \cos^2\xi + (a\nu)^2 \sin^2\xi = 1 - (1 - (a\nu)^2)\sin^2\xi$$

$\longrightarrow$ not dissipative $(\rho(\pi) = 1)$

$$\tan(\alpha\nu\xi) = \frac{a\nu \sin\xi}{\cos\xi} = a\nu \tan\xi$$



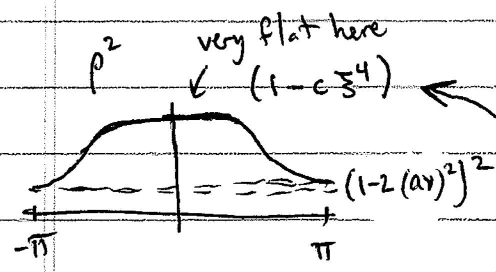
$$\frac{\alpha}{a} = \frac{1}{a\nu\xi} \arctan\left(a\nu \tan\xi\right) \quad \leftarrow \text{always faster than true wavespeed}$$

$$= 1 + \frac{1 - (a\nu)^2}{3}\xi^2 + O(\xi^4)$$

Lax-Wendroff $(u_t = -a u_x)$

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

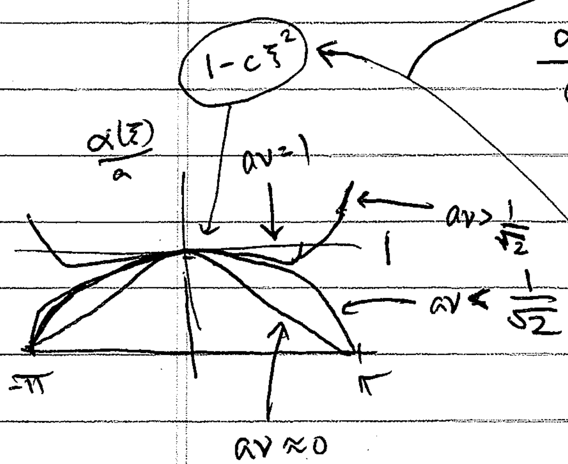
$$G(\xi) = 1 - 2(a\nu)^2 \sin^2 \xi/2 - i(a\nu)\sin\xi = \rho e^{-i\alpha\nu\xi}$$



$$\rho^2 = 1 - 4(a\nu)^2(1 - (a\nu)^2)\sin^4(\xi/2)$$

~~(dissipative of order 4)~~

$$\frac{\alpha}{a} = \frac{1}{a\nu\xi} \arctan\left(\frac{a\nu \sin\xi}{1 - 2(a\nu)^2 \sin^2(\xi/2)}\right)$$



$$= 1 - \frac{1 - (a\nu)^2}{6}\xi^2 + O(\xi^4)$$

for small ξ: numerical wave speed too slow

if $a\nu > \frac{1}{\sqrt{2}}$ and ξ close to $\pm\pi$: waves go too fast

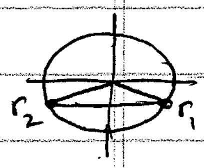
$\boxed{\text{main error in L-W is due to dispersion.}}$

Leapfrog: $u_t = -au_x$ $\qquad u_j^{n+1} = -a\nu(u_{j+1}^n - u_{j-1}^n) + u_j^{n-1}$

$$\hat{u}^{n+1} = G_1(\xi)\hat{u}^n(\xi) + G_2(\xi)\hat{u}^{n-1}(\xi)$$

$$G_1(\xi) = -2ia\nu \sin\xi, \qquad G_2(\xi) \equiv 1$$

general solution: $\hat{u}^n(\xi) = C_1(\xi)\, r_1(\xi)^n + C_2(\xi)\, r_2(\xi)^n$



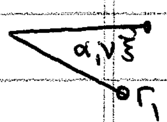
$$r_1 = \sqrt{1-(a\nu)^2\sin^2\xi} - i(a\nu)\sin\xi \qquad \leftarrow r_{1,2} = \frac{G_1 \pm \sqrt{G_1^2 + 4G_2}}{2}$$

$$r_2 = -\sqrt{1-(a\nu)^2\sin^2\xi} - i(a\nu)\sin\xi$$

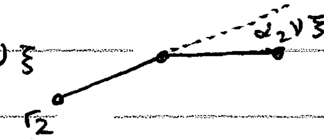
we can interpret the cases $C_1(\xi)=1,\ C_2(\xi)=0$ and $C_1(\xi)=0, C_2(\xi)=1$ as travelling wave solutions

$$u_j^n = r_1(\xi)^n e^{ij\xi}, \qquad u_j^n = r_2(\xi)^n e^{ij\xi}$$

we again want to know how these compare to the exact sol'n $\left(e^{-ia\nu\xi}\right)^n e^{ij\xi}$



$$r_1(\xi) = \rho_1 e^{-i\alpha_1\nu\xi}, \qquad r_2(\xi) = -\rho_2 e^{i\alpha_2\nu\xi}$$



$\rho_1 = \rho_2 = 1$ $\leftarrow$ Leapfrog is strictly non-dissipative (no mode is damped)

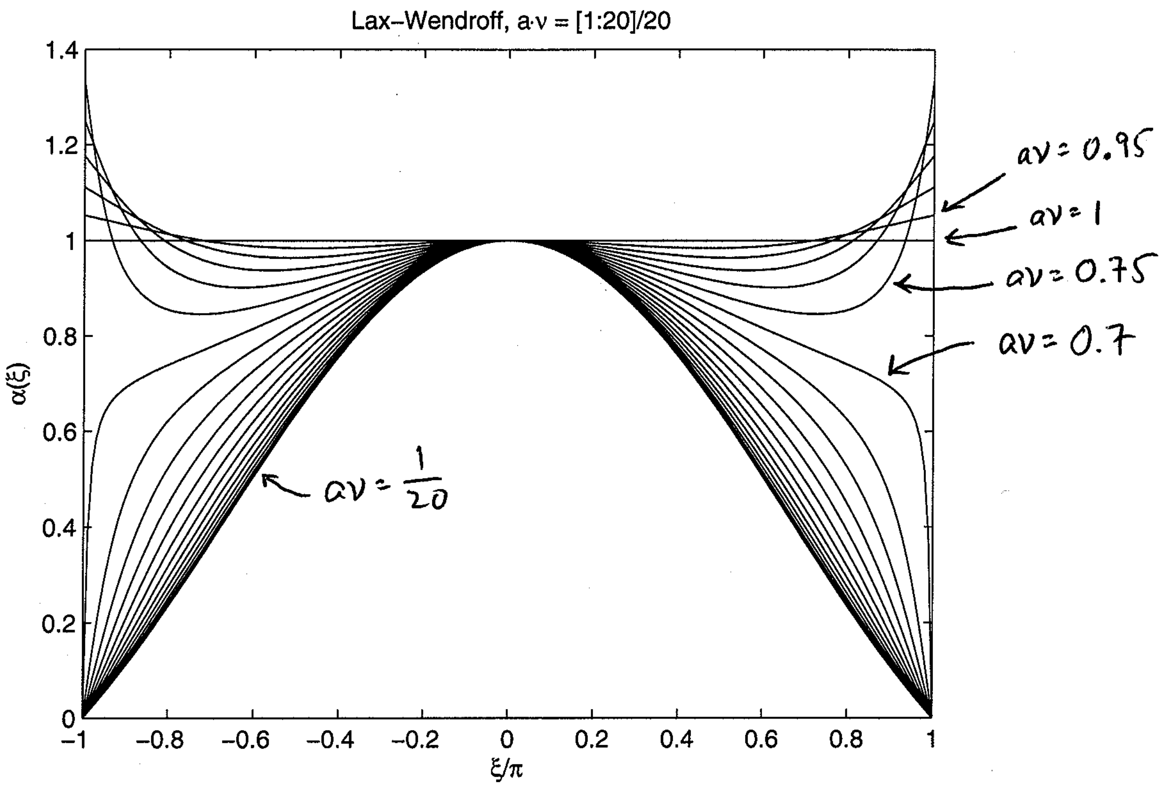
$$\frac{\alpha_1}{a} = \frac{\alpha_2}{a} = \frac{1}{a\nu\xi}\arctan\left(\frac{a\nu\sin\xi}{\sqrt{1-(a\nu)^2\sin^2\xi}}\right) = 1 - \frac{1-(a\nu)^2}{6}\xi^2 + O(\xi^4)$$

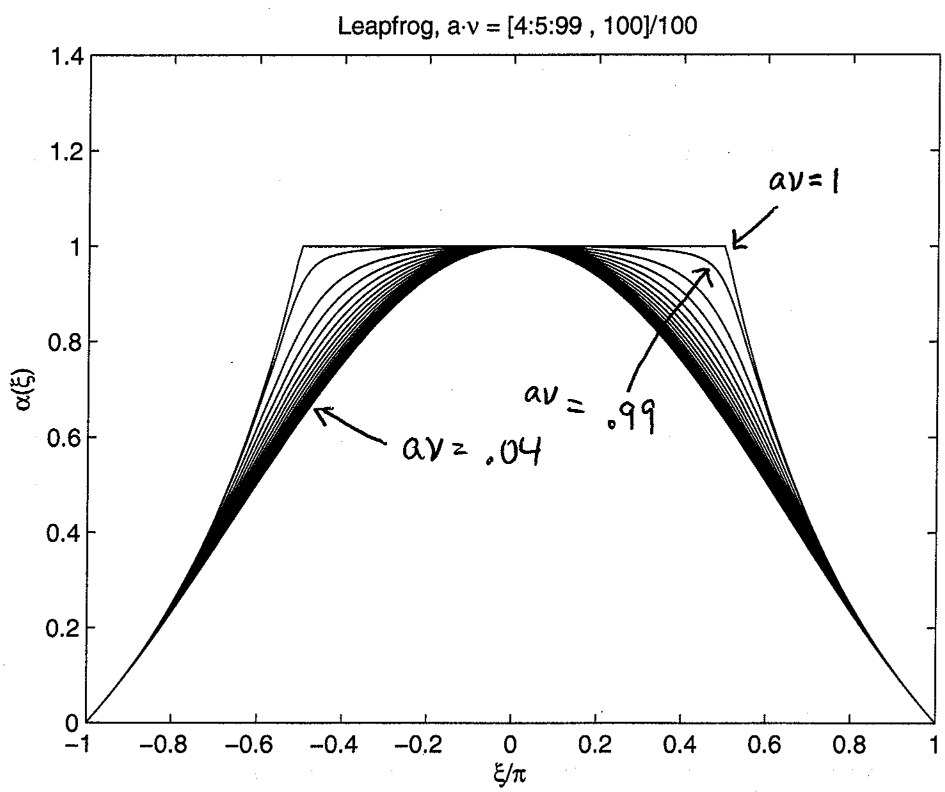
α_1 is the same (through 2nd order) as Lax-Wendroff, but α_2 is oscillating and travelling in the wrong direction. (parasitic mode)

$$r_2(\xi) = -\rho_2 e^{+i\alpha_2\nu\xi}$$

going wrong way

oscillating


Lax–Wendroff, a·ν = [1:20]/20
α(ξ)
ξ/π
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
aν = 0.95
aν = 1
aν = 0.75
aν = 0.7
aν = 1/20


Leapfrog, a·ν = [4:5:99 , 100]/100
α(ξ)
ξ/π
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
aν = 1
aν = .99
aν = .04

you can add numerical dissipation to any scheme.

for leapfrog, two natural candidates are:

$$\text{①} \quad u_j^{n+1} = u_j^{n-1} - a\nu\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\varepsilon}{4}\overbrace{\left(u_{j-1}^{n-1} - 2u_j^{n-1} + u_{j+1}^{n-1}\right)}^{k^2 D^+D^- u_j^{n-1}}$$

$$\text{②} \quad u_j^{n+1} = u_j^{n-1} - a\nu\left(u_{j+1}^n - u_{j-1}^n\right) - \frac{\varepsilon}{16} k^4 \left(D^+D^-\right)^2 u_j^{n-1}$$

the recursion $\hat{u}^{n+1} = G_1 \hat{u}^n + G_2 \hat{u}^{n-1}$

holds with ① $G_1 = -2i a\nu \sin\xi \qquad G_2 = 1 - \varepsilon \sin^2 \xi/2$

or ② $G_1 = -2i a\nu \sin\xi \qquad G_2 = 1 - \varepsilon \sin^4 \xi/2$

the roots become

$$\text{①} \quad r_\pm = \frac{G_1 \pm \sqrt{G_1^2 + 4G_2}}{2} = \pm\sqrt{1 - (a\nu)^2 \sin^2\xi - \varepsilon \sin^2 \xi/2} - i(a\nu)\sin\xi$$

or

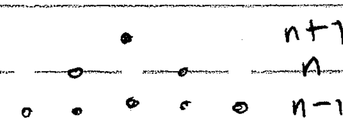
$$\text{②} \quad r_\pm = \pm\sqrt{\underbrace{1 - (a\nu)^2 \sin^2\xi - \varepsilon \sin^4 \xi/2}_{}} - i(a\nu)\sin\xi$$

stability requires this to stay positive. Can't choose ε too large...

drawback to ①: $\rho_\pm = 1 - \varepsilon \sin^2 \xi/2$ is not $O(\xi^3)$

~~so method is only first order now~~

drawback to ② stencil is wider



we can probably find something better with this additional degree of freedom.

Last time: dissipation & dispersion $\quad G(\xi) = \rho e^{-i\alpha \frac{\xi}{h} k}$ $\quad \frac{\xi}{h}k = \nu\xi$

amplification factor (complex) — $G(\xi)$

amplification factor (magnitude) — ρ

wave speed — α

wave number — $\frac{\xi}{h}$

time increment — k

dissipation: $\rho(\xi)$ $\leftarrow$ decay rate of different Fourier modes

dispersion: $\alpha(\xi)$ $\leftarrow$ different modes travel at different speeds

Today: ① dispersion of the leapfrog scheme

② aliasing in the grid based Fourier transform

③ group velocity and wave packets.

recap: if we start with a sequence u_j^0 and run the scheme $u^{n+1} = Bu^n$,

the Fourier transform $\hat{u}^n(\xi)$ evolves via $\hat{u}^{n+1}(\xi) = G(\xi)\hat{u}^n(\xi)$

and we can interpret the inversion formula

$$u_j^n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ij\xi}\hat{u}^n(\xi)\,d\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[G(\xi)^n e^{ij\xi}\right]\hat{u}^0(\xi)\,d\xi$$

as a superposition of travelling Fourier modes on the mesh.

Note that an initial condition of the form $u_j^0 = e^{ij\xi}$ with ξ frozen

advances under the scheme $u^{n+1} = Bu^n$ via $u_j^n = G(\xi)^n e^{ij\xi}$

it advances under the PDE $u_t = -au_x$ via $u_j^n = e^{i\frac{x_j - at_n}{h}\xi} = \left(e^{-ia\frac{k}{h}\xi}\right)^n e^{ij\xi}$

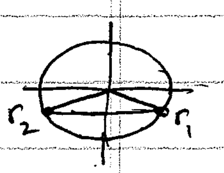
so we want to know how close $G(\xi) = \rho e^{-i\alpha\nu\xi}$ is to $e^{-ia\nu\xi}$.

Leapfrog: $u_t = -au_x$ $\qquad u_j^{n+1} = -a\nu\left(u_{j+1}^n - u_{j-1}^n\right) + u_j^{n-1}$

$$\hat{u}^{n+1} = G_1(\xi)\,\hat{u}^n(\xi) + G_2(\xi)\,\hat{u}^{n-1}(\xi)$$

$$\cancel{G_1(\xi) = -2ia\nu\sin\xi}\,, \quad \cancel{G_2(\xi) = 1}$$

general solution: $\hat{u}^n(\xi) = C_1(\xi)\,r_1(\xi)^n + C_2(\xi)\,r_2(\xi)^n$



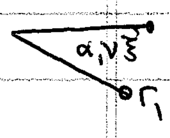
$$r_1 = \sqrt{1-(a\nu)^2\sin^2\xi} - i(a\nu)\sin\xi \qquad \leftarrow r_{1,2} = \frac{G_1 \pm \sqrt{G_1^2 + 4G_2}}{2}$$

$$r_2 = -\sqrt{1-(a\nu)^2\sin^2\xi} - i(a\nu)\sin\xi$$

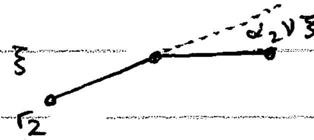
we can interpret the cases $C_1(\xi)=1,\ C_2(\xi)=0$ and $C_1(\xi)=0, C_2(\xi)=1$ as travelling wave solutions

$$u_j^n = r_1(\xi)^n e^{ij\xi}\,, \qquad u_j^n = r_2(\xi)^n e^{ij\xi}$$

we again want to know how these compare to the exact sol'n $\left(e^{-ia\nu\xi}\right)^n e^{ij\xi}$



$$r_1(\xi) = \rho_1 e^{-i\alpha_1\nu\xi}\,, \qquad r_2(\xi) = -\rho_2 e^{i\alpha_2\nu\xi}$$



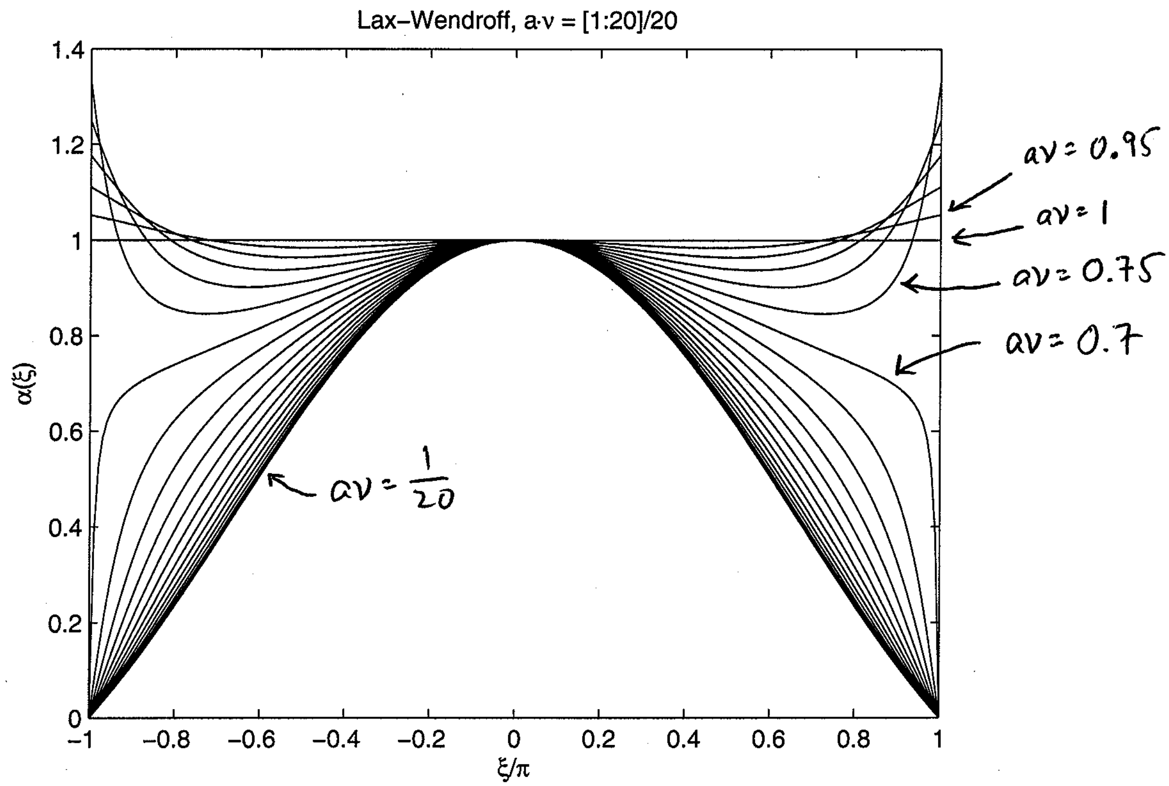
$\rho_1 = \rho_2 = 1$ $\leftarrow$ Leapfrog is strictly non-dissipative (no mode is damped)

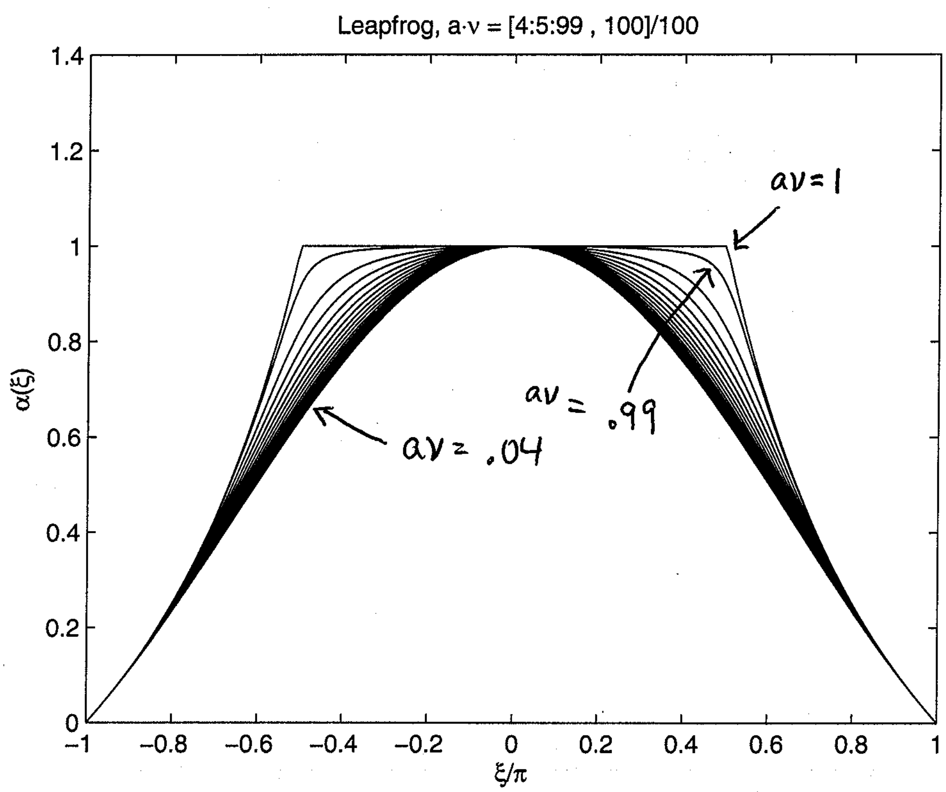
$$\frac{\alpha_1}{a} = \frac{\alpha_2}{a} = \frac{1}{a\nu\xi}\arctan\left(\frac{a\nu\sin\xi}{\sqrt{1-(a\nu)^2\sin^2\xi}}\right) = 1 - \frac{1-(a\nu)^2}{6}\xi^2 + O(\xi^4)$$

α_1 is the same (through 2nd order) as Lax-Wendroff, but α_2 is oscillating and travelling in the wrong direction. (parasitic mode)

(The goal of the first step is to activate as little of the parasitic mode as possible.)

$$r_2(\xi) = \underbrace{-}_{\text{oscillating}}\rho_2 e^{\overbrace{+i\alpha_2\nu\xi}^{\text{going wrong way}}}$$


Lax–Wendroff, a·ν = [1:20]/20
α(ξ)
ξ/π
aν = 0.95
aν = 1
aν = 0.75
aν = 0.7
aν = 1/20


Leapfrog, a·ν = [4:5:99 , 100]/100
α(ξ)
ξ/π
aν = 1
aν = .99
aν = .04

so v is a sequence and u is a function

<u>aliasing</u>: suppose $u(x)$ is a smooth function decaying rapidly to zero as $x \to \pm\infty$, and let $v_j = u(x_j) = u(jh)$.

let's modify our grid based fourier transform a little:

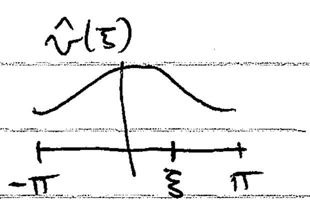
$$\hat{v}(\xi) = h \sum_j e^{-ij\xi} v_j \qquad (h \text{ is newly added})$$

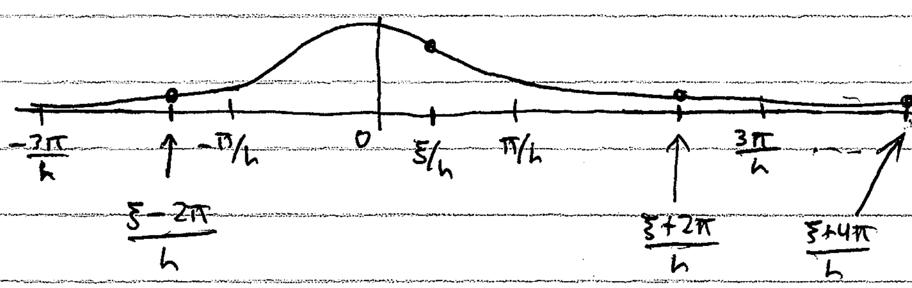
Poisson summation formula:

$$h \sum_j e^{-ij\xi} u(jh) = \sum_n \hat{u}\left(\frac{\xi + 2\pi n}{h}\right)$$

$$\hat{u}(k) = \int_{-\infty}^{\infty} e^{-ikx} u(x)\, dx \qquad \leftarrow k \text{ is a wave number here, not a timestep}$$

$$\text{so} \quad \hat{v}(\xi) = \hat{u}\left(\frac{\xi}{h}\right) + \sum_{n \neq 0} \hat{u}\left(\frac{\xi + 2\pi n}{h}\right)$$





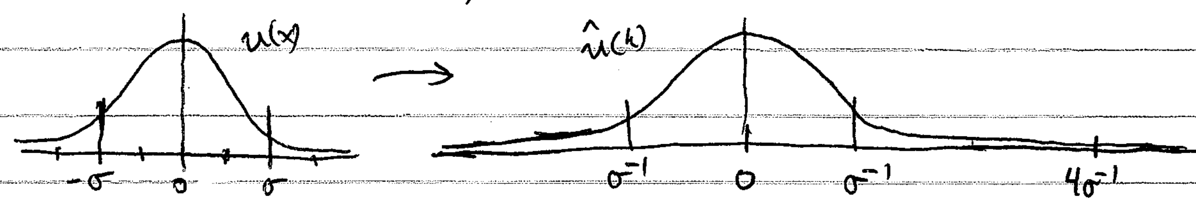
The values of $\hat{u}$ outside the interval $\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$ get mapped back into this interval (they are aliased) when the integral in the fourier transform is replaced by a discrete sum.

$$\hat{u}(k) = \int_{-\infty}^{\infty} e^{-ikx} u(x)\, dx, \qquad \hat{v}(\xi) = \sum_{j=-\infty}^{\infty} e^{-ij\xi} u(x_j)\, h$$

$$j\xi = \frac{x_j}{h}\xi = x_j \frac{\xi}{h} = x_j k$$

so if h is small enough that $\hat{u}(k)$ is negligible for $|k| \geq \pi/h$, then $\hat{v}(\xi) \approx \hat{u}(k/h)$ to high accuracy.

<u>example:</u> $u(x) = e^{-x^2/2\sigma^2}$, $\hat{u}(k) = \sqrt{2\pi}\,\sigma\, e^{-\frac{k^2\sigma^2}{2}}$



say $h = \frac{\sigma}{2}$ $\qquad$ $\hat{v}(\xi) = \hat{u}\left(\frac{\xi}{h}\right) + \sum_{n \neq 0} \hat{u}\left(\frac{\xi + 2\pi n}{h}\right)$

when $k = \pi/h$, $\frac{k^2\sigma^2}{2} = \frac{\pi^2\sigma^2}{2h^2} = 2\pi^2 = -19.7$

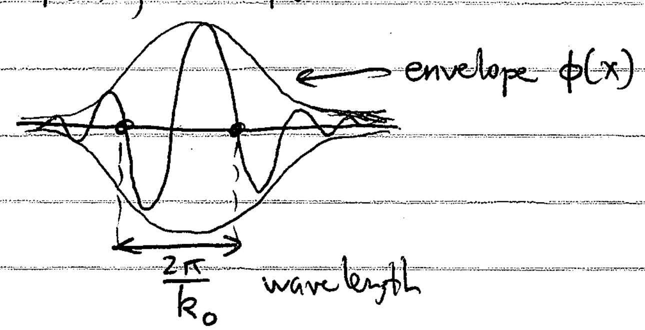
$\exp\left(-\frac{k^2\sigma^2}{2}\right) = 2.7 \times 10^{-9}$

so this sum is tiny and $\hat{v} \approx \hat{u}$

---

example: $u(x) = e^{ik_0 x}\phi(x)$, $\phi(x) = e^{-x^2/2\sigma^2}$

wave packet:

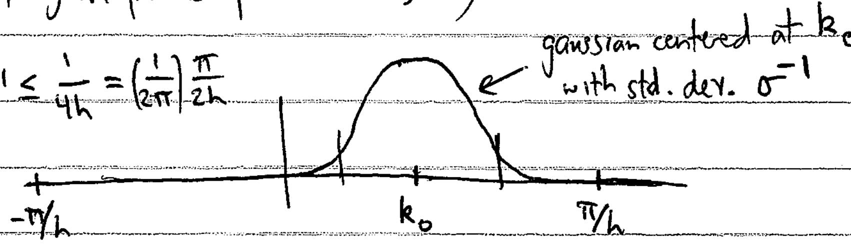


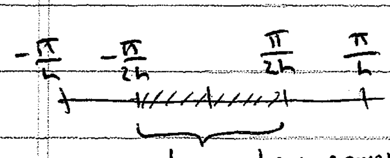
$\hat{u}(k) = \hat{\phi}(k - k_0)$

choose $h \leq \frac{\pi}{2k_0}$ (only 4 grid points per wavelength)

assume $\sigma \geq 4h$ so $\sigma^{-1} \leq \frac{1}{4h} = \left(\frac{1}{2\pi}\right)\frac{\pi}{2h}$

Then $\hat{u}(k)$ looks like





the distance from k_0 to $\pm\frac{\pi}{h}$ is $\geq \frac{\pi}{2h} \geq 2\pi\sigma^{-1}$

$\therefore$ boundary where aliasing kicks in is more than 6 standard deviations away. $\therefore \hat{v}(\xi) \approx \hat{u}(\xi/h)$

summary: you don't have to sample a wave packet very closely for the discrete fourier transform $\hat{v}(\xi)$ to accurately approximate the continuous F.T. $\hat{u}(\xi/h)$.

group velocity: so what will our schemes do to a wave packet?

Equation: $u_t = -a u_x$

initial condition: $u(x,0) = e^{ik_0 x}\phi(x)$, $\phi(x) = e^{-x^2/2\sigma^2}$

exact solution: $u(x,t) = e^{ik_0(x-at)}\phi(x-at)$

numerical solution: $u_j^n \approx \rho(\xi_0)^n e^{ik_0(x_j - \underbrace{\alpha(\xi_0)}_{\text{phase velocity}} t_n)} \phi(x_j - \underbrace{\gamma(\xi_0)}_{\text{group velocity}} t_n)$

$\xi_0 = k_0 h$

what is γ and why is it not equal to α?

$$u_j^n = \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi}\left(\rho e^{-i\alpha\nu\xi}\right)^n e^{ij\xi}\,\hat{u}^0(\xi)\,d\xi$$

$\hat{u}$ is the grid based F.T. here (stopped using $\hat{v}$ for this)

$$= \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi}\rho^n e^{i\left(\frac{\xi}{h}x_j - \alpha\frac{\xi}{h}t_n\right)}\,\hat{u}^0(\xi)\,d\xi$$

wave number $k = \xi/h$

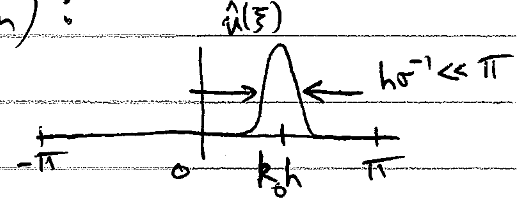
angular frequency $\omega = \alpha\,\xi/h$

(we'll write Δt for the timestep today)

phase velocity: $\frac{\omega}{k} = \alpha$

group velocity: $\frac{d\omega}{dk} = \frac{d\xi}{dk}\frac{d\omega}{d\xi} = h\cdot h^{-1}\frac{d}{d\xi}(\alpha\xi) = \frac{d}{d\xi}(\alpha\xi) = \gamma$

So now let's imagine that $\hat{u}^0(\xi)$ is narrowly peaked near $k_0 h$ (i.e. σ is large compared to h):



Then $\hat{u}^0(\xi) \approx \hat{\phi}\left(\frac{\xi}{h} - k_0\right)$

and

$$u_j^n = \frac{h^{-1}}{2\pi} \int_{-\pi}^{\pi} \rho(\xi)^n e^{i\left(\frac{\xi}{h} x_j - \alpha(\xi)\frac{\xi}{h} t_n\right)} \hat{u}^0(\xi)\, d\xi$$

in region where $\hat{u}^0(\xi)$ is nonzero:
$\xi_0 = k_0 h$, $\rho(\xi) \approx \rho(\xi_0)$, $\alpha(\xi)\xi \approx \alpha_0 \xi_0 + \gamma_0(\xi - \xi_0)$

$$\approx \frac{h^{-1}}{2\pi} \int_{-\pi}^{\pi} \underbrace{\rho(\xi_0)^n e^{i\frac{\xi_0}{h}(x_j - \alpha_0 t_n)}}_{\text{indep. of } \xi} e^{i\left(\frac{\xi - \xi_0}{h}\right)(x_j - \gamma_0 t_n)} \hat{\phi}\left(\frac{\xi}{h} - k_0\right) d\xi$$

$k_0 = \xi_0/h$

$k = \frac{\xi}{h} - k_0$

$dk = \frac{d\xi}{h}$

$$= \rho(\xi_0)^n e^{i k_0 (x_j - \alpha_0 t_n)} \frac{1}{2\pi} \int_{-\frac{\pi}{h} - k_0}^{\frac{\pi}{h} - k_0} e^{ik(x_j - \gamma_0 t_n)} \hat{\phi}(k)\, dk$$

but $\hat{\phi}(k) \approx 0$ outside this region so replace with $\int_{-\infty}^{\infty} \cdots$

$$= \rho(\xi_0)^n e^{i k_0 (x_j - \alpha_0 t_n)} \phi(x_j - \gamma_0 t_n)$$

as claimed.

Lec 20 , Math 228 B

group velocity and wave packets

$$\hat{\phi}(k) = \sqrt{2\pi}\,\sigma\, e^{-\frac{k^2\sigma^2}{2}}$$

$$u_t = -a u_x \ , \quad u(x,0) = e^{ik_0 x}\phi(x) \ , \qquad \phi(x) = e^{-x^2/2\sigma^2}$$

exact solution: $\quad u(x,t) = e^{ik_0(x-at)}\phi(x-at)$

numerical sol'n: $\quad v_j^n = \rho(\xi_0)^n e^{ik_0(x_j - \alpha(\xi_0)t_n)}\phi(x_j - \gamma(\xi_0)t_n)$

<u>scheme</u>: $v_j^0 = u(x_j,0)$ $\leftarrow$ sample initial condition on grid

$v^{n+1} = Bv^n$ $\leftarrow$ apply finite difference scheme

$$\hat{v}^n(\xi) = h\sum_j e^{-ij\xi} v_j^n \ , \qquad \hat{u}(k,t) = \int_{-\infty}^{\infty} e^{-ikx}u(x,t)\,dx$$

$$\hat{v}^0(\xi) = h\sum_j e^{-ij\xi}u(jh,0) = \hat{u}\left(\tfrac{\xi}{h},0\right) + \underbrace{\sum_{n\neq 0}\hat{u}\left(\frac{\xi+2\pi n}{h}\right)}_{\text{extremely small}}$$

$$\therefore \ \hat{v}^0(\xi) \approx \hat{\phi}\left(\tfrac{\xi}{h} - k_0\right)$$

e.g. if $hk_0 \le \frac{\pi}{2}$ then

$h < \frac{\sigma}{4} \Rightarrow \left(\sum_{n\neq 0}\cdots\right) \le \sigma e^{-18}$

$h < \frac{\sigma}{6} \Rightarrow \left(\sum\cdots\right) \le \sigma e^{-43}$

our Fourier analysis tells us

$$v_j^n = \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi}\left(\rho e^{-i\alpha\nu\xi}\right)^n e^{ij\xi}\,\hat{v}^0(\xi)\,d\xi \qquad G(\xi) = \rho e^{-i\alpha\nu\xi}$$

$$= \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi}\rho^n e^{i\left(\frac{\xi}{h}x_j - \alpha\frac{\xi}{h}t_n\right)}\,\hat{u}^0(\xi)\,d\xi$$

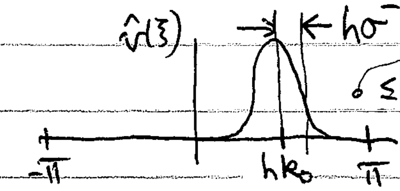
$k = \xi/h \qquad \omega = \alpha\xi/h$

phase velocity $\ \alpha = \frac{\omega}{k} \qquad$ group velocity: $\ \gamma = \frac{d\omega}{dk} = \frac{d}{d\xi}\left(\xi\alpha(\xi)\right)$

so now let's imagine that $\hat{v}^0(\xi)$ is narrowly peaked near $\xi_0 = hk_0$ (i.e. $h\sigma^{-1} \ll \frac{\pi}{2}$)

$\leftarrow h\sigma^{-1} \ll \frac{\pi}{2} \le \delta$

$\delta \le \min(\pi - hk_0, hk_0 + \pi)$



we'll approximate $\left[\begin{array}{l} \rho(\xi) \approx \rho(\xi_0) = \rho_0 \\ \alpha(\xi)\xi \approx \alpha_0\xi_0 + \gamma_0(\xi - \xi_0) \end{array}\right]$ in the region where $\hat{u}^0(\xi)$ is significant

so

$$v_j^n = \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi} \rho(\xi)^n e^{i\left(\frac{\xi}{h}x_j - \alpha(\xi)\frac{\xi}{h}t_n\right)} \hat{v}^0(\xi)\, d\xi$$

$$\approx \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi} \underbrace{\rho(\xi_0)^n e^{i\frac{\xi_0}{h}(x_j - \alpha_0 t_n)}}_{\text{indep. of } \xi} e^{i\left(\frac{\xi - \xi_0}{h}\right)(x_j - \gamma_0 t_n)} \hat{\phi}\left(\frac{\xi}{h} - k_0\right) d\xi$$

$$\boxed{\begin{array}{c} k = \frac{\xi}{h} - k_0 \\ dk = h^{-1} d\xi \end{array}}$$

$$= \rho(\xi_0)^n e^{ik_0(x_j - \alpha_0 t_n)} \frac{1}{2\pi}\int_{-\frac{\pi}{h} - k_0}^{\frac{\pi}{h} - k_0} e^{ik(x_j - \gamma_0 t_n)} \hat{\phi}(k)\, dk$$

but $\hat{\phi}(k) \approx 0$ outside this region so replace with $\int_{-\infty}^{\infty} \cdots$

$$v_j^n = \underbrace{\rho(\xi_0)^n}_{\substack{\text{numerical} \\ \text{dissipation}}} e^{ik_0(x_j - \alpha_0 t_n)} \phi(x_j - \gamma_0 t_n)$$

carrier signal travels at phase velocity

wave envelope travels at group velocity.

in practice this is exactly what you see happen.

Stability analysis for PDE's with solutions that grow (can't expect $\|B\| \le 1$ since solution grows)

example: $u_t = -u_x + u$ exact soln: $u(x,t) = e^t g(x-t)$

scheme: let's look for something like Lax-Wendroff

$$u(x,t+k) = u(x,t) + k\, u_t(x,t) + \frac{k^2}{2} u_{tt}(x,t) + \cdots$$

$$u_t = -u_x + u$$

$$u_{tt} = -u_{xt} + u_t = u_{xx} - u_x - u_x + u = u_{xx} - 2u_x + u$$

$$u_j^{n+1} = u_j^n + k\left[-\frac{u_{j+1}^n - u_{j-1}^n}{2h} + u_j^n\right] + \frac{k^2}{2}\left[\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + u_j^n - 2\frac{u_{j+1}^n - u_{j-1}^n}{2h}\right]$$

$$= u_j^n - \nu\frac{u_{j+1}^n - u_{j-1}^n}{2} + \frac{\nu^2}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) \leftarrow \text{L.W. for } u_t = -u_x$$

$$+ k\left[\left(1+\frac{k}{2}\right)u_j^n - \nu\frac{u_{j+1}^n - u_{j-1}^n}{2}\right] \leftarrow \text{terms associated with } +u \text{ part of equation}$$

$$G(\xi) = \left(1 - 2\nu^2 \sin^2 \xi/2 - i\nu \sin\xi\right) + k\left[\left(1+\frac{k}{2}\right) - i\nu\sin\xi\right]$$

$$|G(\xi)| \le \sqrt{1 - 4\nu^2(1-\nu^2)\sin^4 \xi/2} + k\sqrt{\left(1+\frac{k}{2}\right)^2 + \nu^2\sin^2\xi}$$

$$\|B\|_2 = \|G\|_\infty \le 1 + 2k \quad \leftarrow \text{assuming } \nu \le 1 \text{ and } k \le 1$$

$$\|B^n\| \le \|B\|^n \le (1+2k)^n \le (e^{2k})^n = e^{2kn} \le e^{2T}$$

$\therefore$ scheme is stable. (numerical solution $u^n = B^n u^0$ grows exponentially in time, but that's OK. The true solution does, too. Bad schemes grow exponentially in n without a k to balance it.)

Fourier collocation / pseudo-spectral methods

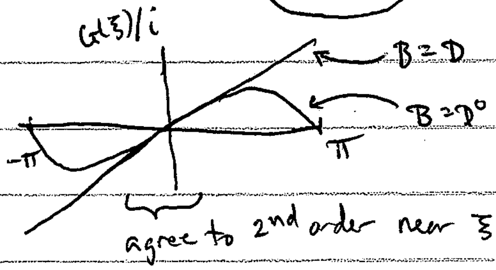
the amplification factors of D_x^0 and $D_x^+ D_x^-$ are

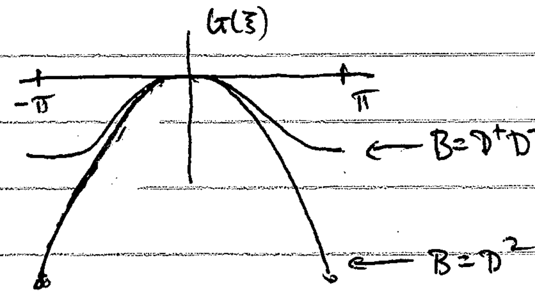
$$\frac{e^{i\xi} - e^{-i\xi}}{2h} = ih^{-1}\sin\xi \quad \text{and} \quad \frac{e^{i\xi} - 2 + e^{-i\xi}}{h^2} = -4h^{-2}\sin^2 \xi/2$$

if we compute $\frac{\partial}{\partial x}$ and $\frac{\partial^2}{\partial x^2}$ of $e^{i\frac{x}{h}\xi}$ we get

$j = x_j/h$

$$ih^{-1}\xi \quad \text{and} \quad -h^{-2}\xi^2$$





here $D : \ell^2 \to \ell^2$ is given by $Du_j = \frac{h^{-1}}{2\pi}\int_{-\pi}^{\pi} ih^{-1}\xi\, e^{ij\xi}\hat{u}(\xi)\,d\xi$

where $\hat{u}(\xi) = h\sum_{j=-\infty}^{\infty} e^{-ij\xi} u_j$

now let's try the method of lines (discretize space first) using D:

$$u_t = u_{xx} \quad \longrightarrow \quad v_t = D^2 v$$

exact $\rightarrow u(x,t)$

numerical $\rightarrow v_j(t) \approx u(x_j, t)$

finally, let's discretize time using e.g. a Runge-Kutta method

$$w_1 = D^2(v^n)$$
$$w_2 = D^2\left(v^n + \frac{k}{2}w_1\right)$$
$$v^{n+1} = v^n + kw_2$$

$\leftarrow$ explicit midpoint rule (2nd order RK method)

(for $y' = f(t,y)$:

$k_1 = f(t_n, y_n)$

$k_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1)$

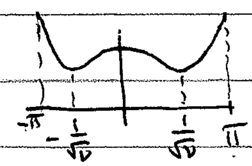
$y_{n+1} = y_n + hk_2$)

if we combine the stages, we get $v^{n+1} = v^n + kD^2v^n + \frac{k^2}{2}D^4v^n = Bv^n$

the amplification factor is: $G(\xi) = 1 - \frac{k}{h^2}\xi^2 + \frac{k^2}{2h^4}\xi^4$

$$= 1 - \nu\xi^2 + \frac{\nu^2}{2}\xi^4$$

maximum value: $G'(\xi) = -2\nu\xi + 2\nu^2\xi^3 = 0 \Rightarrow \left[\begin{array}{l}\xi = 0 \quad \text{or} \\ \xi^2 = 1/\nu\end{array}\right.$



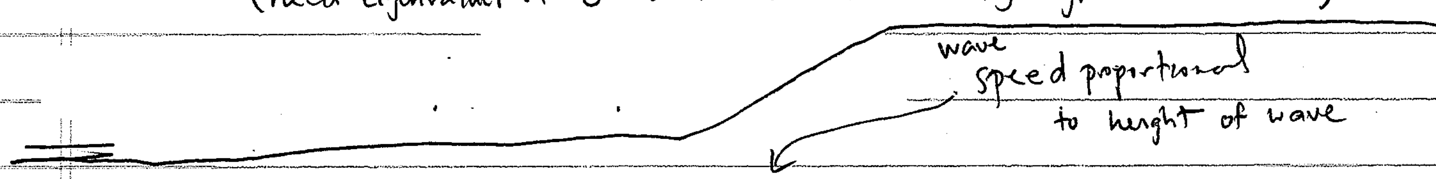
$G(\frac{1}{\sqrt{\nu}}) = 1 - 1 + \frac{1}{2} = \frac{1}{2} \le 1$ ✓

$G(0) = 1$ ✓, $G(\pi) = 1 - \nu\pi^2 + \frac{\nu^2}{2}\pi^4 \le 1$

$$\Rightarrow \nu^2\pi^4 \le 2\nu\pi^2 \Rightarrow \nu \le \frac{2}{\pi^2}$$

$$\|G\|_\infty = \begin{cases} 1 & \nu \le \frac{2}{\pi^2} = .2026 \quad \leftarrow \text{stable} \\ 1 - \nu\pi^2 + \frac{\nu^2}{2}\pi^4 & \nu > \frac{2}{\pi^2} \quad \leftarrow \text{unstable} \end{cases}$$

you can do the same sort of analysis for higher order RK methods.
(need eigenvalues of B to lie inside the stability region of the scheme)



wave speed proportional to height of wave

viscous Burger's eqn: $u_t + \underbrace{uu_x}_{\text{nonlinearity}} = \varepsilon u_{xx}$

initial condition: $u(x,t) = \sin x$, $-\pi \le x \le \pi$

periodic b.c.'s

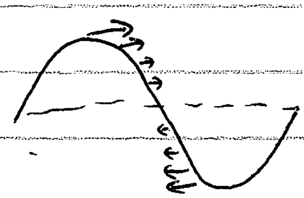
method of lines: $v_t + D\left[\frac{1}{2}v^2\right] = \varepsilon D^2 v$

evaluate v^2 on the grid

spectral derivative operator (using FFT) now timestep this using your favorite ODE method, e.g.

$$\begin{cases} w_1 = -D\left[\frac{1}{2}(v^n)^2\right] + \varepsilon D^2 v^n \\ w_2 = -D\left[\frac{1}{2}\left(v^n + \frac{k}{2}w_1\right)^2\right] + \varepsilon D^2\left(v^n + \frac{k}{2}w_1\right) \\ v^{n+1} = v^n + k w_2 \end{cases}$$

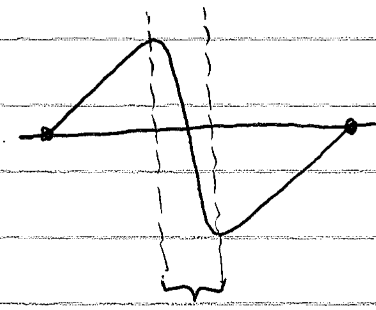
expect



speed proportional to height

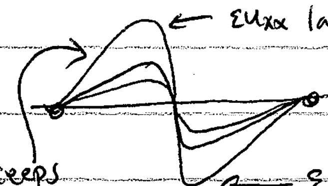
(εu_{xx} doesn't kick in yet)

shock forms:



shock region $O(\varepsilon)$

then solution decays (due to εu_{xx})

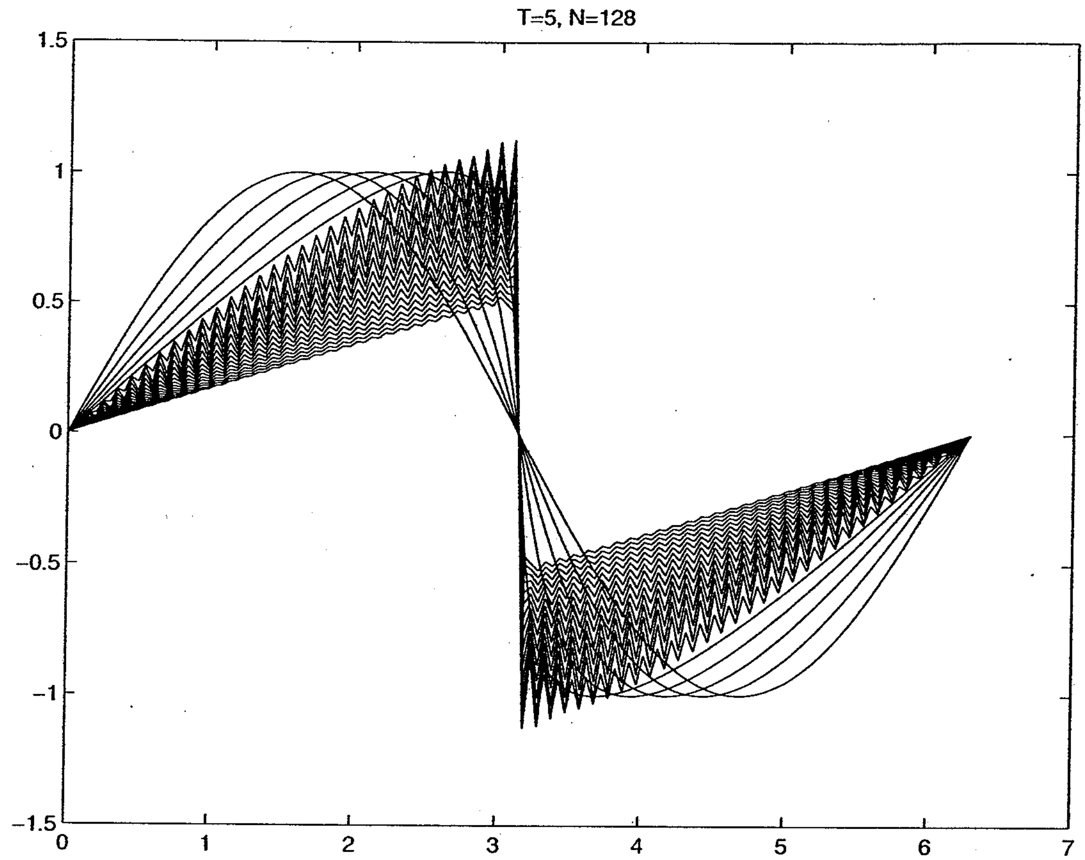


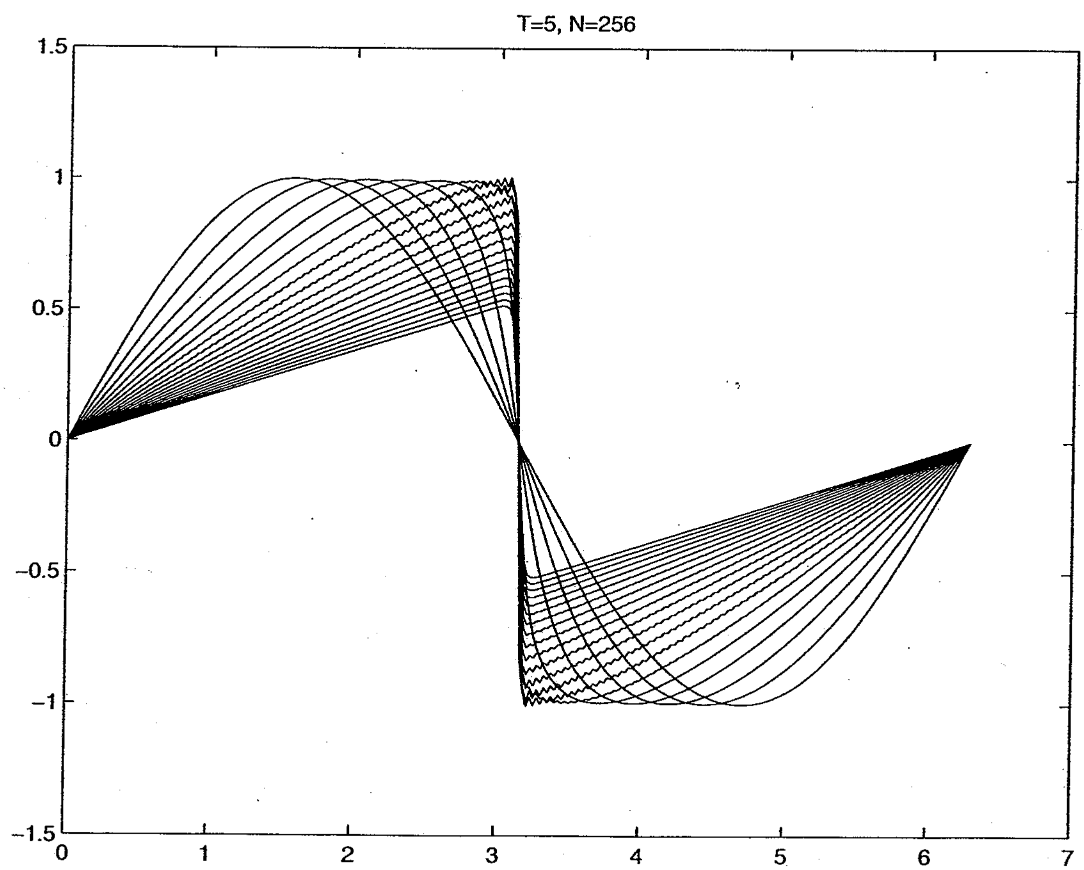
$\leftarrow$ εu_{xx} large negative (decays)

(keeps moving left in proportion to height)

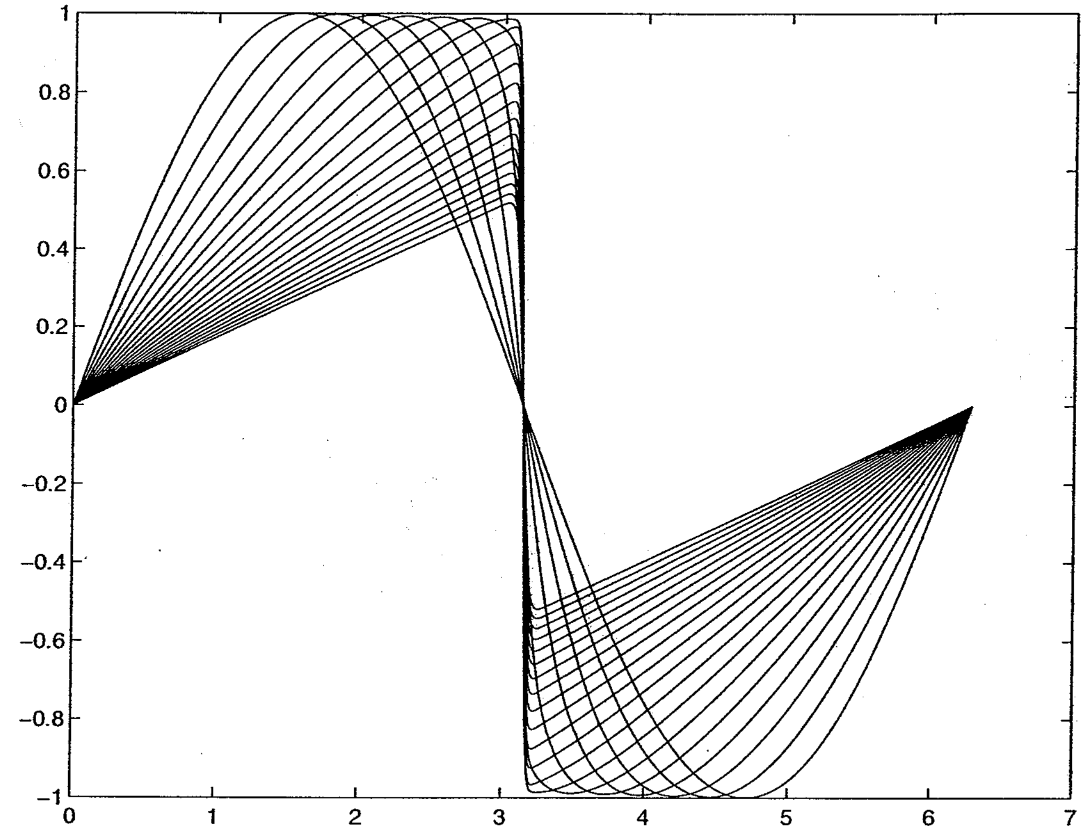
$\leftarrow$ εu_{xx} large positive (decays)

need h small compared to ε to resolve the calculation

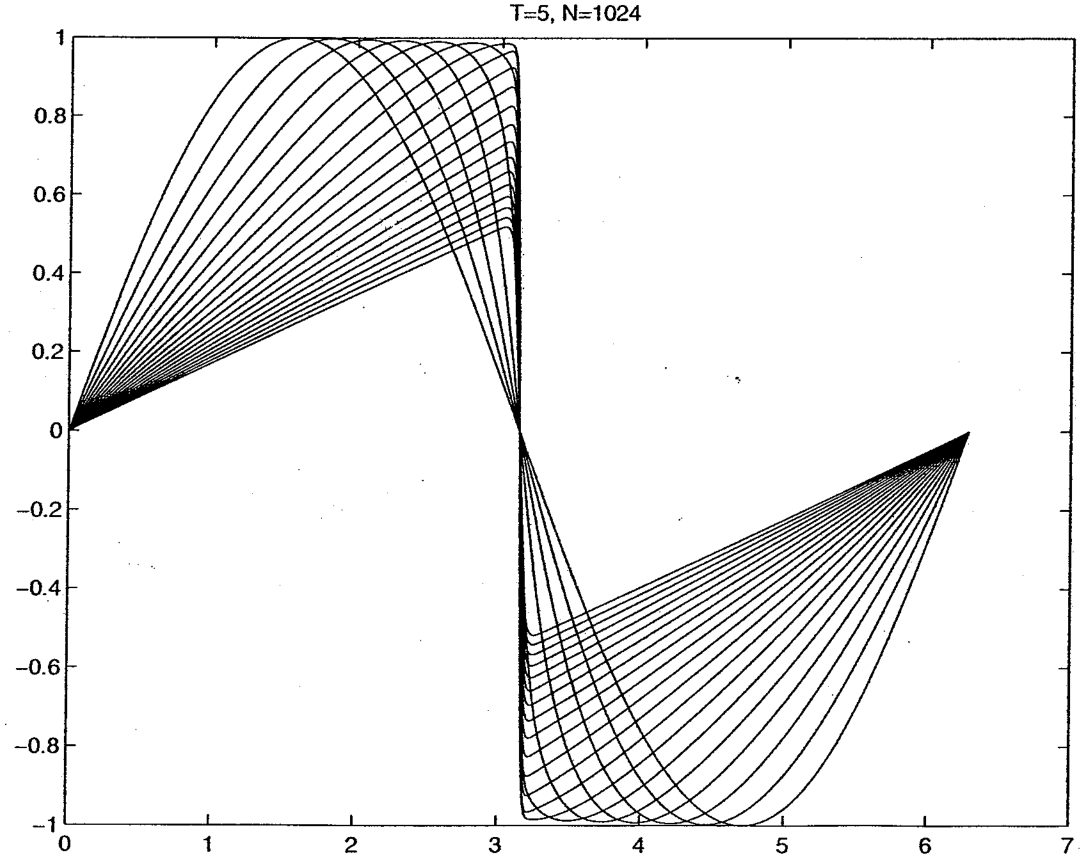

T=5, N=128
1.5
1
0.5
0
−0.5
−1
−1.5
0
1
2
3
4
5
6
7

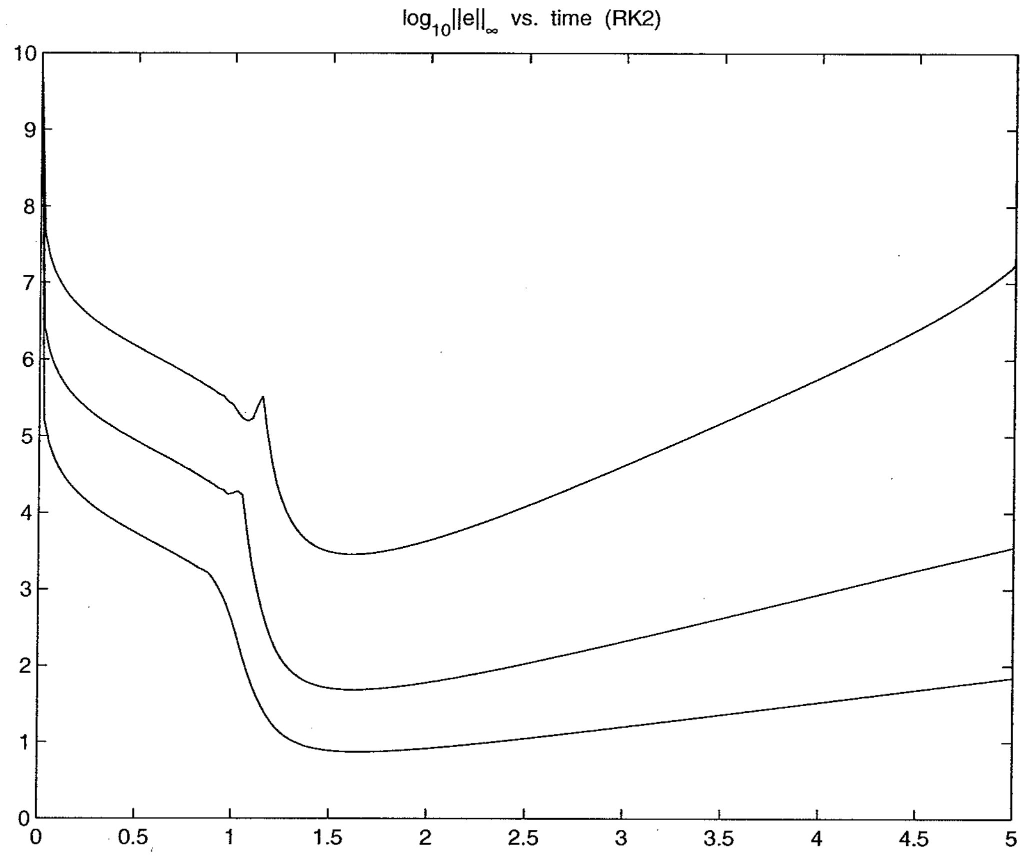

T=5, N=256
1.5
1
0.5
0
−0.5
−1
−1.5
0
1
2
3
4
5
6
7

T=5, N=512

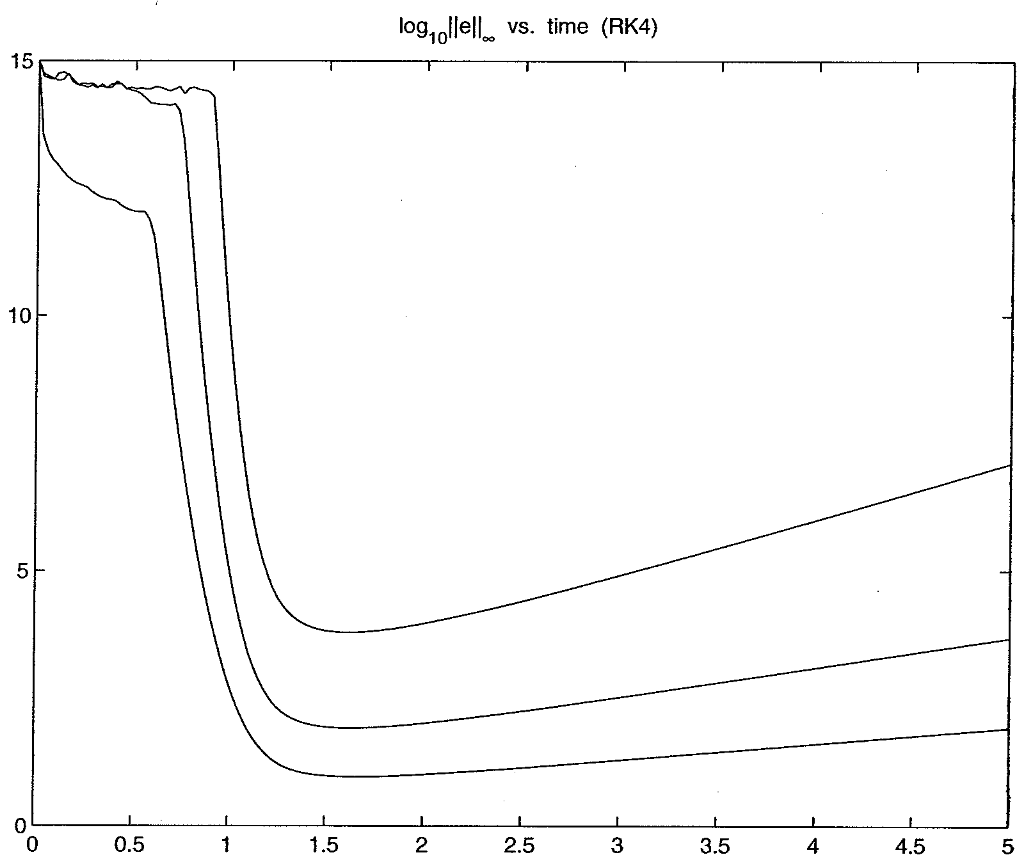


T=5, N=1024



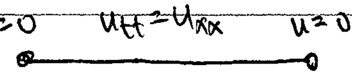


log10||e||∞ vs. time (RK2)
10
9
8
7
6
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5



log10||e||∞ vs. time (RK4)
15
10
5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Spectral integration in matlab (for differentiation, mult. by $ih^{-1}\xi_k$ rather than divide)

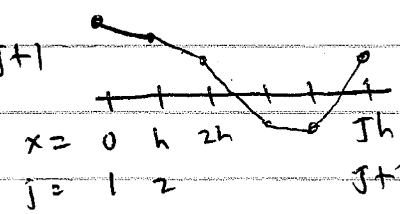
example 

Dirichlet conditions

suppose you have computed $\begin{pmatrix} u_t \\ u_x \end{pmatrix}$ on the grid and now you want to recover $u(x_j)$ from $u_x(x_j)$, $x_j = (j-1)h$

known: $\quad ux(j) = u_x((j-1)h) \qquad 1 \le j \le J+1$

wanted: $\quad u(j) = u(x_j)$



Let $N = 2J$ and extend by even symmetry

$$ux(J+1+j) = ux(J+1-j) \qquad 1 \le j \le J-1$$

Now define $w = \text{fft}(u)$, i.e.

$$w_k = \sum_{j=1}^{N} e^{-2\pi i (j-1)(k-1)/N} u_j = \sum_{j=1}^{N} e^{-ij\xi_k} u_j \qquad 1 \le k \le N$$

$$\xi_k = \frac{2\pi}{N} \cdot \begin{cases} k-1 & 1 \le k \le N/2 \\ k-1-N & \frac{N}{2}+1 \le k \le N \end{cases}$$

N=8 example

k	1	2	3	4	5	6	7	8
$\frac{N}{2\pi}\xi_k$	0	1	2	3	±4	-3	-2	-1

Nyquist frequency always causes trouble... just zero out that mode.

using ξ_k avoids actually re-shuffling the components of w

Now we can integrate the inversion formula $ux = \text{ifft}(w)$ term by term:

$$ux(x_j) = \frac{1}{N}\sum_{k=1}^{N} e^{i\frac{x_j}{h}\xi_k} w_k \Rightarrow u(x_j) = \frac{1}{N}\sum_{k=1}^{N} e^{ij\xi_k}\left(\frac{w_k}{ih^{-1}\xi_k}\right) + C$$

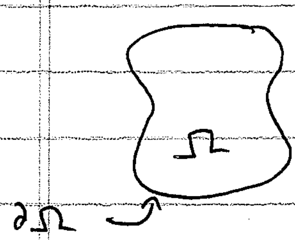
algorithmically, you just have to set

$$\tilde{w}_k = \frac{h}{i\xi_k} w_k \qquad 1 \le k \le N$$

and then define $u = \text{ifft}(\tilde{w})$. Easy...

C chosen so $u(0) = u(1) = 0$
in fact, $C=0$ since the even symmetry of ux gives u odd symmetry when C is omitted

the Finite element method



Poisson equation

$$-\Delta u = f \quad \text{on } \Omega$$
$$u = 0 \quad \text{on } \partial\Omega \quad \leftarrow \text{Dirichlet B.C.'s}$$

$$\left(\Delta = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = \text{Laplacian} \right)$$

mathematical questions: do solutions exist?
are they unique?
how nice are they? (regularity)

computational questions: how can I find an approximate solution?
how close is the " " to the true sol'n?
how fast can I compute the solution?
how much memory does the computer need, etc...

There are several approaches to studying existence, uniqueness and regularity of elliptic equations. They all have numerical counterparts:

(1) fundamental solutions, Green's functions, potential theory $\rightsquigarrow$ boundary integral methods

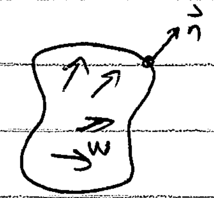
(2) maximum principle, subharmonic functions $\rightsquigarrow$ useful for proving error bounds in finite difference methods

(3) Hilbert space methods $\rightsquigarrow$ finite elements.

idea of 3^{rd} approach: multiply by a test function v and integrate by parts

$$\int -v\Delta u \, dx = \int v f \, dx \qquad \leftarrow dx \text{ means } dx\,dy \text{ or } dA$$

divergence theorem: $\displaystyle\int_\Omega \nabla\cdot\vec{w}\, dx = \int_{\partial\Omega} \vec{w}\cdot\vec{n}\, ds$



vector identity: $\nabla\cdot(v\nabla u) = \nabla v\cdot\nabla u + v\Delta u$

so
$$\int -v\Delta u\, dx = \int_\Omega \nabla v\cdot\nabla u\, dx - \int_\Omega \nabla\cdot(v\nabla u)\, dx$$
$$= \int_\Omega \nabla v\cdot\nabla u\, dx - \int_{\partial\Omega} v \underbrace{(\nabla u\cdot n)}_{\frac{\partial u}{\partial n}\ \text{normal derivative}}\, ds$$

Green's identity

so if $v=0$ on $\partial\Omega$, the boundary term is zero and we get

$$\int \nabla v\cdot\nabla u\, dx = \int f v\, dx$$

We now introduce the Sobolev spaces

$H^1(\Omega)$ = "space of L^2 functions with one <u>weak</u> derivative in L^2" (defined later)

$H^1_0(\Omega)$ = "H^1 functions that vanish on the boundary"

weak formulation of the Dirichlet problem:

find $u\in H^1_0(\Omega)$ such that for all test functions $v\in H^1_0(\Omega)$

$$\int \nabla u\cdot\nabla v\, dx = \int f v\, dx$$

The finite element framework parallels the theoretical one.

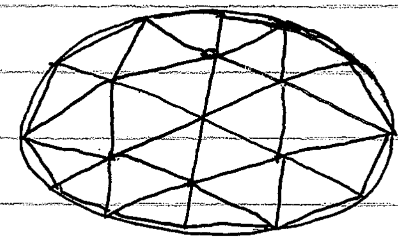
A <u>conforming</u> FE space is a finite dimensional subspace

$$S_h \subseteq H^1_0(\Omega)$$

$\uparrow$

h is a parameter (e.g. the diameter of the largest triangle in the mesh)

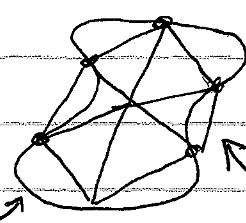
e.g.



$$\Omega_h = \bigcup_k T_k \subseteq \Omega$$

S_h = continuous functions $u \in C(\Omega)$ that are piecewise linear on each triangle of a mesh and zero at the boundary nodes (and on $\Omega \setminus \Omega_h$)

example of a non-conforming FE space:



Ω is not convex

a non-trivial piecewise linear function on this triangle is nonzero on $\partial\Omega$, so doesn't belong to $H^1_0(\Omega)$

Discrete problem:

find $u_h \in S_h$ s.t. $\int \nabla u_h \cdot \nabla v \, dx = \int f v \, dx \quad \forall v \in S_h$

Note: the solution space is smaller, but it is now easier to be a solution since the space of test functions is also smaller

The theoretical tools that are used to study the weak formulation of the continuous problem carry over directly to the discrete problem (and provide error estimates) what are these theoretical tools?

[today: high level overview
later: details

(1) The spaces $H^1(\Omega)$ and $H^1_0(\Omega)$ are <u>Hilbert spaces</u> with inner product (complete inner product spaces)

$$(u,v)_1 = \int_\Omega uv\,dx + \int_\Omega \nabla u \cdot \nabla v\,dx$$

as always, the norm of a Hilbert space is given by $\|u\|_1 = \sqrt{(u,u)_1}$

<u>warning</u>: the derivatives here are weak derivatives (defined later).
if you consider only differentiable functions, the spaces are not complete.

Our finite dimensional subspace $S_h \subseteq H^1_0(\Omega)$ inherits the inner product from the ambient space. (S_h is also a Hilbert space in its own right)

(2) The equation we're trying to solve has the structure:

find $u \in H$ s.t. $a(u,v) = \langle l, v\rangle \quad \forall v \in H$

Here $a(\cdot,\cdot)$ is a <u>bilinear form</u> and $\langle l, \cdot \rangle$ is a <u>linear functional</u>

$a: H \times H \to \mathbb{R}$	$l: H \to \mathbb{R}$
$a(u, \alpha v + \beta w) = \alpha\, a(u,v) + \beta (u,w)$ $a(\alpha u + \beta v, w) = \alpha\, a(u,w) + \beta (v,w)$ for any $u,v,w \in H$ and $\alpha,\beta \in \mathbb{R}$	$\langle l, \alpha u + \beta v\rangle = \alpha \langle l, u\rangle + \beta \langle l, v\rangle$ for any $u,v \in H$ and $\alpha,\beta \in \mathbb{R}$

In our case, $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$

$$\langle l, v \rangle = \int_\Omega f v \, dx$$

There's a powerful theorem called the <u>Lax-Milgram</u> theorem that says that if $a(\cdot,\cdot)$ is a <u>coercive</u>, <u>continuous</u> bilinear form and $\langle l, \cdot \rangle$ is a <u>bounded</u>, linear functional then there is a unique solution u satisfying

$$a(u,v) = \langle l, v \rangle \quad \forall v \in H.$$

true on any Hilbert space H.

coercivity means $\exists\, \alpha > 0$ s.t. $\alpha \|u\|^2 \le a(u,u) \quad \forall u \in H$

continuity (or boundedness) means $\exists\, C < \infty$ s.t. $|a(u,v)| \le C \|u\| \cdot \|v\| \quad \forall u,v \in H$

Same idea for linear functionals. l is bounded (or continuous) if

$$\exists\, C < \infty \quad \text{s.t.} \quad |\langle l, v \rangle| \le C \|v\| \quad \forall v \in H$$

The smallest choice of C is denoted $\|l\|$ (the norm of l)

In our case, $\|u\|_1^2 = \int_\Omega u^2 \, dx + \int_\Omega |\nabla u|^2 \, dx$

so it's easy to show $|a(u,v)| \le \|u\|_1 \cdot \|v\|_1$ ← $C=1$ works for a

and $|\langle l, v \rangle| \le \underbrace{\|f\|_0}_{\sqrt{\int_\Omega f^2 \, dx}} \cdot \|v\|_1$ ← upper bound for $\|l\|$

coercivity is harder to prove, i.e. that $\exists\, \alpha > 0$ s.t.

$$\alpha \|u\|_1^2 \le a(u,u) \qquad \forall u \in H_0^1(\Omega)$$

(proof is based on the Poincaré - Friedrichs inequality ... discussed next week)

summary: the Lax-Milgram theorem gives us existence and uniqueness of the continuous and discrete systems:

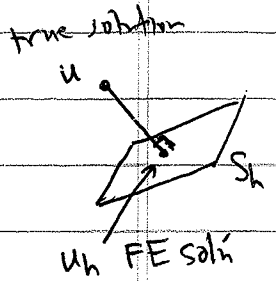
$$a(u,v) = \langle \ell, v \rangle \qquad \forall v \in H_0^1(\Omega)$$

$$(\circledast) \qquad a(u_h,v) = \langle \ell, v \rangle \qquad \forall v \in S_h$$

(3) we now want to estimate the error, $\|u_h - u\|_1$

from $\circledast$, we have

$$a(u-u_h, v) = 0 \qquad \forall v \in S_h$$



Galerkin orthogonality (closest solution in the a-norm $\|u\|_a = \sqrt{a(u,u)}$)

using coercivity, we have

$$
\begin{aligned}
\alpha \|u-u_h\|_1^2 &\le a(u-u_h, u-u_h) \\
&= a(u-u_h, u \underbrace{- v_h + v_h}_{0 \text{ for any } v_h \in S_h} - u_h) \\
&= a(u-u_h, u-v_h) + \underbrace{a(u-u_h, v_h-u_h)}_{0 \text{ since } v_h - u_h \in S_h} \\
&\le C \|u-u_h\| \cdot \|u-v_h\|
\end{aligned}
$$

$\therefore$ $\|u-u_h\|_1 \le \frac{C}{\alpha} \|u-v_h\|$ for every $v_h \in S_h$ ← Cea's Lemma this reduces the error analysis to determining how well the true solú can be approximated by <u>any</u> function in the FE space.

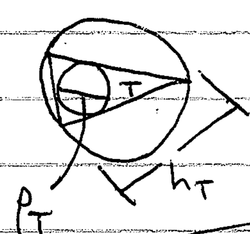
(4) Now we look for other functions in S_h that we can guarantee are close to u, namely $v_h = I_h u$

interpolation operator. Evaluate the exact solution at the nodes and interpolate on the elements

By Cea: $\|u-u_h\|_1 \le \frac{C}{\alpha}\|u - I_h u\|_1$

---

we will see that there is a constant C_2 depending on the mesh quality K



$$\frac{h_T}{\rho_T} \le K \quad \forall T \text{ in the triangulation}$$

such that $\|u - I_h u\|_{1,h} \le C_2 h\, |u|_{2,h} \quad \forall u \in H^2(\Omega_h)$

Sobolev space of square integrable functions with two weak derivatives

here $|u|_{2,h}^2 = \int_\Omega u_{xx}^2 + u_{xy}^2 + u_{yy}^2 \, dx$

---

finally, there's a theorem (the elliptic regularity theorem) that says that if Ω is convex, there is a constant C_3 s.t. the solution of the Dirichlet problem $\left[\begin{array}{ll} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{array}\right]$ satisfies

$\|u\|_2^2 = \|u\|_0^2 + |u|_1^2 + |u|_2^2$

$\|u\|_2 \le C_3 \|f\|_0$. final error estimate: $\boxed{\|u-u_h\|_1 \le \frac{C C_2 C_3}{\alpha} h \|f\|_0}$

$(\Omega = \Omega_h = \text{convex polygon})$

Sobolev spaces

Let $1 \le p < \infty$, define $L^p(\Omega) = \left\{ f: \Omega \to \mathbb{R} : \int |f(x)|^p dx < \infty \right\}$

functions which are equal a.e. are identified
(we don't distinguish between f & g if $\int_\Omega |f-g| dx = 0$)

most common cases: $p = 1, 2, \infty$

$$L^\infty = \left\{ f : \Omega \to \mathbb{R} \quad : \quad \sup_{x \in \Omega} |f(x)| < \infty \right\}$$

The norms on the L^p spaces are:

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx \right)^{1/p} \qquad 1 \le p < \infty$$

$$\|f\|_\infty = \sup_x |f(x)|$$

The space $L^2(\Omega)$ is special. Not only is it a Banach space, it's also a Hilbert space with the inner product

$$(f, g) = \int_\Omega f(x) g(x) \, dx \qquad \left(\text{or } \int f(x) \overline{g(x)} \, dx \text{ if complex valued fcns are considered} \right)$$

$$\|f\| = \sqrt{(f, f)}$$

we will write $\|f\|_0$ to mean $\|f\|_{L^2(\Omega)}$

and $(f, g)_0$ to mean $\int_\Omega f g \, dx$

A Hilbert space is a complete inner product space.

completeness is the reason for studying weak derivatives.

for example, we can endow

$C'(I)$, $I=(-1,1)$ with the inner product

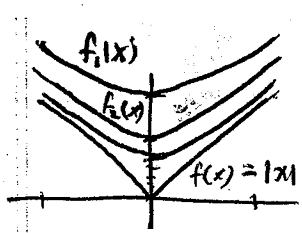
$$(f,g) = \int_{-1}^{1} f(x)g(x)\,dx + \int_{-1}^{1} f'(x)g'(x)\,dx$$

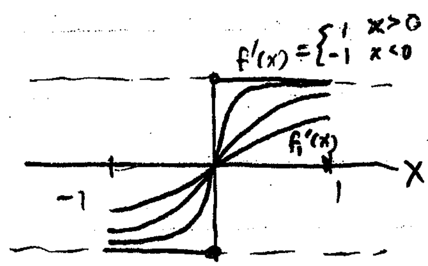
but the sequence $f_n(x) = \sqrt{x^2+1/n^2}$ is a Cauchy sequence which doesn't converge to any function $f \in C'(I)$.

(it converges to $f(x) = |x|$ which is not differentiable at $x=0$)

$f_n(x) = \sqrt{x^2 + 1/n^2}$

$f_n'(x) = \dfrac{x}{\sqrt{x^2+1/n^2}}$





$\{f_n\}$ is Cauchy:

$$\|f_n - f_m\| \le \|f_n - f\| + \|f_m - f\| \to 0 \quad \text{as } m,n\to\infty$$

since $\|f_n - f\|^2 = \int_{-1}^{1} (f_n(x) - f(x))^2\,dx + \int_{-1}^{1} (f_n'(x) - f'(x))^2\,dx \to 0$

as $n\to\infty$ by dominated convergence theorem.

weak derivatives

def: $C_c^\infty(\Omega) = \{ \phi : \Omega \to \mathbb{R} \; : \; \begin{array}{l} \phi \text{ is smooth } (\infty\text{-differentiable}) \\ \phi \text{ has compact support} \end{array} \}$

$\operatorname{spt}\phi = \overline{\{x : \phi(x) \neq 0\}}$ ← closure in Ω

$= \{ x \in \Omega : \exists \text{ sequence } x_n \to x \text{ s.t. } x_n \in \Omega \text{ and } \phi(x_n) \neq 0 \}$

$\operatorname{spt}\phi$ is compact if it is closed and bounded.

as a subset of $\mathbb{R}^n$

motivation: suppose $u \in C^1(\Omega)$ and $\phi \in C_c^\infty(\Omega)$

$$\operatorname{div}\begin{pmatrix} 0 \\ u\phi \\ 0 \\ 0 \end{pmatrix} \leftarrow i\text{th slot} \quad = (\partial_i u)\phi + u\,\partial_i\phi \qquad \partial_i = \frac{\partial}{\partial x_i}$$

so
$$\int_\Omega (\partial_i u)\phi + u\,\partial_i\phi \, dx = \int_\Omega \operatorname{div}\begin{pmatrix} 0 \\ u\phi \\ 0 \end{pmatrix} dx \underset{\text{div. thm}}{=} \int_{\partial\Omega} \begin{pmatrix} 0 \\ u\phi \\ 0 \\ 0 \end{pmatrix}\cdot n \, dA \underset{\phi = 0 \text{ on } \partial\Omega}{=} 0$$

$$\therefore \int_\Omega u\,\partial_i\phi \, dx = -\int_\Omega (\partial_i u)\,\phi \, dx \qquad (i = 1, \dots, n)$$

integration by parts formula

conclusion: Let $\Omega \subseteq \mathbb{R}^n$ be open and connected.

if $u \in C^1(\Omega)$ and $\phi \in C_c^\infty(\Omega)$ then

$$\int_\Omega u\, \partial_i \phi \, dx = -\int_\Omega (\partial_i u)\, \phi \, dx \qquad (i=1,\dots,n)$$

more generally, if $u \in C^k(\Omega)$ and $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index, then

$$\int_\Omega u\, \partial^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega (\partial^\alpha u)\, \phi \, dx$$

Here each α_i is an integer ≥ 0 and

$$\partial^\alpha u = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} u$$

further notation: $|\alpha| = \alpha_1 + \dots + \alpha_n$

for $x \in \mathbb{R}^n$ we write $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$

$$\alpha! = \alpha_1! \, \alpha_2! \cdots \alpha_n!$$

$\alpha \leq \beta$ means $\alpha_1 \leq \beta_1, \dots, \alpha_n \leq \beta_n$

how many ways can you put k balls in n bins?

the number of multi-indices of order k is

$$\#\{\alpha : |\alpha| = k\} = \binom{n-1+k}{n-1} = \frac{(n-1+k)!}{(n-1)!\, k!}$$

example: choosing 2 from 6 leaves 4 objects partitioned into 3 groups:

$n=3, k=4$

$\circ \; \odot \; \circ \; \circ \; \odot \; \circ \iff \alpha = (1,2,1), \; \partial^\alpha = \partial_1 \partial_2^2 \partial_3$

$\odot \; \circ \; \odot \; \circ \; \circ \; \circ \iff \alpha = (0,1,3), \; \partial^\alpha = \partial_2 \partial_3^3$

def: Suppose $u, v \in L^2(\Omega)$ and α is a multi-index. We say v is the αth partial derivative of u ($v = \partial^\alpha u$) provided that

$$\int_\Omega u\, \partial^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega v \phi \, dx$$

for all test functions $\phi \in C_c^\infty(\Omega)$. Note: this is the definition of $\partial^\alpha u$.

we just saw that continuous classical derivatives are weak derivatives.

Thm: weak derivatives are unique.

pf: Sp $v = \partial^\alpha u$ and $\tilde{v} = \partial^\alpha u$.

Then $$(-1)^{|\alpha|} \int_\Omega v \phi \, dx = \int u\, \partial^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega \tilde{v} \phi \, dx$$

i.e. $$\int_\Omega (v - \tilde{v}) \phi \, dx = 0 \quad \forall \phi \in C_c^\infty(\Omega)$$

$\therefore\ v = \tilde{v}$ a.e. (see e.g. Evans, PDE; Lieb & Loss, Analysis)

Rk: The definition only really requires
$u, v \in L^1_{loc}(\Omega) = \{u : \Omega \to \mathbb{R} : \int_K |u(x)| \, dx < \infty \ \forall \text{ cpt } K \subset \Omega\}$
but we will only ever need the L^2 theory of weak derivatives.

Sobolev spaces: $H^m(\Omega) = \left\{ u : \Omega \to \mathbb{R} \;\middle|\; \begin{array}{c} \forall \alpha \text{ with } |\alpha| \le m, \\ \partial^\alpha u \text{ exists weakly} \\ \text{and belongs to } L^2(\Omega) \end{array} \right\}$

Scaler product

$$(u,v)_m = \sum_{|\alpha| \le m} (\partial^\alpha u, \partial^\alpha v)_0 \leftarrow L^2 \text{ inner product}$$

$(H^0 = L^2)$

$(f,g)_0 = \int_\Omega fg \, dx$

norm:

$$\|u\|_m = \sqrt{(u,u)_m} = \sqrt{\sum_{|\alpha| \le m} \|\partial^\alpha u\|_0^2}$$

semi-norm:

$$|u|_m = \sqrt{\sum_{|\alpha| = m} \|\partial^\alpha u\|_0^2}$$

a seminorm on a vector space X is a mapping $x \mapsto \|x\|$ from X to $[0,\infty)$ s.t.

$\|x+y\| \le \|x\| + \|y\|$

$\|\lambda x\| = |\lambda| \, \|x\|$

A norm is a semi norm s.t. $\|x\| = 0$ iff $x = 0$

---

important cases:

$$(u,v)_0 = \int_\Omega uv \, dx$$

$$(u,v)_1 = \int_\Omega uv \, dx + \int_\Omega \nabla u \cdot \nabla v \, dx$$

note that $\sum_{|\alpha|=1} (\partial^\alpha u, \partial^\alpha v)_0 = (\partial_1 u)(\partial_1 v) + \dots + (\partial_n u)(\partial_n v)$
$= \nabla u \cdot \nabla v$

$$(u,v)_2 = (u,v)_1 + \int_\Omega \operatorname{lowtri}(D^2 u) : \operatorname{lowtri}(D^2 v) \, dx$$

Hessian matrix $(D^2 u)_{ij} = \partial_i \partial_j u$, $1 \le i,j \le n$

$$\operatorname{lowtri}(A) = \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix}$$

$$A : B = \sum_{i,j} A_{ij} B_{ij}$$

228B Lec 23

Last time: weak derivatives, definition of Sobolev spaces

Today: finish discussing Sobolev spaces,
prove Poincaré-Friedrichs inequality (coercivity of $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$)

recap:

$\alpha = (\alpha_1, \dots, \alpha_n)$ ← non-negative integers

weak derivatives: Suppose $u, v \in L^2(\Omega)$ and α is a multi-index.
we say v is the αth partial derivative of u ($v = \partial^\alpha u$)
provided that

$$\int_\Omega u \, \partial^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega v \phi \, dx$$

for all test functions $\phi \in C_c^\infty(\Omega)$. Note: this is
the definition of $\partial^\alpha u$.

Sobolev spaces: $H^m(\Omega) = \left\{ u : \Omega \to \mathbb{R} \;\middle|\; \begin{array}{c} \forall \alpha \text{ with } |\alpha| \le m \\ \partial^\alpha u \text{ exists weakly} \\ \text{and belongs to } L^2(\Omega) \end{array} \right\}$

scalar product: $(u,v)_m = \sum_{|\alpha| \le m} (\partial^\alpha u, \partial^\alpha v)_0$ ← L^2 $\quad (H^0 = L^2)$

$(f,g)_0 = \int_\Omega f g \, dx$

norm: $\|u\|_m = \sqrt{(u,u)_m} = \sqrt{\sum_{|\alpha| \le m} \|\partial^\alpha u\|_0^2}$

seminorm: $|u|_m = \sqrt{\sum_{|\alpha| = m} \|\partial^\alpha u\|_0^2}$

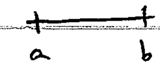
seminorm properties:
$\|x\| \ge 0$
$\|x+y\| \le \|x\| + \|y\|$
$\|\lambda x\| = |\lambda| \, \|x\|$
norm: above and
$\|x\| = 0$ iff $x = 0$

important cases

$(u,v)_0 = \int_\Omega uv \, dx, \quad (u,v)_1 = \int_\Omega uv \, dx + \int_\Omega \nabla u \cdot \nabla v \, dx$

note that $\sum_{|\alpha|=1} (\partial^\alpha u, \partial^\alpha v)_0 = (\partial_1 u, \partial_1 v)_0 + \dots + (\partial_n u, \partial_n v)_0 = \int_\Omega \nabla u \cdot \nabla v \, dx$

Remark: In 1d, there's already a generalization of derivative beyond $C^k(a,b)$ $\quad \Omega = (a,b)$ open interval



Fundamental theorem of Calculus for Lebesgue integrals:

if $a<b$ and $F:[a,b]\to\mathbb{R}$, TFAE:

(1) F is absolutely continuous on $[a,b]$

(2) $F(x) - F(a) = \int_a^x f(t)\,dt$ for some $f \in L^1(a,b)$

(3) F is differentiable almost everywhere on $[a,b]$,
$F' \in L^1(a,b)$, and $F(x) - F(a) = \int_a^x F'(t)\,dt$

Here (1) means that $\forall \varepsilon > 0 \;\exists \delta > 0$ s.t.
if $(a_1,b_1), \dots, (a_N,b_N)$ is any disjoint collection of intervals in $[a,b]$ with $\sum_1^N (b_i - a_i) < \delta$
then $\sum_1^N |F(b_i) - F(a_i)| < \varepsilon$

proof of theorem: see Folland's book on Real Analysis

Lemma: if F, G are abs. cont. on $[a,b]$ then so is FG and
$$\int_a^b (FG' + F'G)\,dx = FG\Big|_a^b$$

Theorem: $u \in H^1(a,b)$ iff (u is equal a.e. to an absolutely continuous function on $[a,b]$, and $u' \in L^2(a,b)$)

proof: SKIP!

$\Leftarrow$) Let $\phi \in C_c^\infty(a,b)$. Since u and ϕ are abs. cont.,

the lemma gives

$$\int_a^b u\phi'\,dx = \underbrace{u\phi\Big|_a^b}_{0} - \int_a^b u'\phi\,dx$$

so u' (defined a.e. as $\lim_{h\to 0} \frac{u(x+h)-u(x)}{h}$) is a weak derivative of u. Since $u, u' \in L^2$, $u \in H^1$.

$\Rightarrow$) Given $u, v \in L^2(a,b)$ s.t. $\int_a^b u\phi'\,dx = -\int_a^b v\phi\,dx$ $\forall \phi \in C_c^\infty(a,b)$

define $\tilde{u}(x) = \int_a^x v(t)\,dt$.

Then $\tilde{u}$ is abs. cont. and $\tilde{u}' = v$ a.e.

and hence weakly by $\Leftarrow$) above.

$\therefore \int (u-\tilde{u})\phi'\,dx = 0 \quad \forall \phi \in C_c^\infty(a,b)$.

Choose any $\phi_0 \in C_c^\infty(a,b)$ with $\int_a^b \phi_0(x)\,dx = 1$

For any $\phi \in C_c^\infty(a,b)$ we have

$$\phi(x) = \underbrace{\phi(x) - \alpha\phi_0(x)}_{\text{has mean zero}} + \alpha\phi_0(x) \qquad \alpha = \int_a^b \phi(x)\,dx$$

hence equals $\psi'(x)$ for $\psi(x) = \int_a^x \phi(t) - \alpha\phi_0(t)\,dt \in C_c^\infty(a,b)$

$\therefore \int (u-\tilde{u})\phi\,dx = \int (u-\tilde{u})\,\alpha\phi_0(x)\,dx = \alpha c = c\int_a^b \phi(x)\,dx$ $\quad c = \int_a^b (u-\tilde{u})\phi_0\,dx$

$\therefore \int (u-\tilde{u}-c)\phi\,dx = 0 \quad \forall \phi$. $\therefore u = \tilde{u} + c$ a.e.

$\therefore u$ is abs. cont.

In 2-d there are unbdd fcns in H^1.

Claim $u(x,y) = \log\log\frac{2}{r}$ belongs to $H^1(\Omega)$, $\Omega = \{x^2+y^2<1\}$
unit disk

pf: $r=\sqrt{x^2+y^2}$, $\partial_x r = \frac{x}{r}$ $\partial_y r = \frac{y}{r}$

$$\nabla u = \frac{-1}{r^2\log\frac{2}{r}}\begin{pmatrix} x \\ y \end{pmatrix} \qquad |\nabla u|^2 = \frac{1}{r^2\log^2\frac{2}{r}}$$

(1) $$\int_\Omega |\nabla u|^2\, dx = \int_0^{2\pi}\int_0^1 \frac{1}{r^2\log^2\frac{2}{r}}\, r\,dr\,d\theta = 2\pi\int_0^1 \frac{1}{r\log^2\frac{2}{r}}\,dr$$

$$\log\frac{2}{r} = \log 2 - \log r \rightarrow \frac{\partial}{\partial r}\left(\log\frac{2}{r}\right)^{-1} = -\left(\log\frac{2}{r}\right)^{-2}\left(-\frac{1}{r}\right)$$

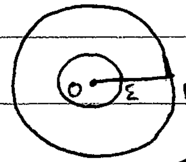
$$= 2\pi\left(\log\frac{2}{r}\right)^{-1}\Big|_0^1 = \frac{2\pi}{\log 2} - 0 < \infty$$

(2) $$\int_\Omega u^2\, dx = 2\pi\int_0^1 \underbrace{r\log^2\log\frac{2}{r}}_{\text{bounded}}\,dr < \infty$$

bounded ($\to 0$ as $r\to 0$ by l'Hôpital's rule:

$$\lim_{r\to 0}\frac{\log\log\frac{2}{r}}{r^{-1/2}} = \lim_{r\to 0}\frac{\frac{-1}{r\log\frac{2}{r}}}{-\frac{1}{2}r^{-3/2}} = \lim_{r\to 0}\frac{2r^{1/2}}{\log 2/r} = 0$$

(3) Let $\Omega_\varepsilon = \Omega \setminus B(0,\varepsilon)$



Then for any $\phi \in C_c^\infty(\Omega)$,

Green's identity is valid on Ω_ε since u & ϕ are smooth there

$$\int_{\Omega_\varepsilon} u\, \partial_i \phi \, dx = \underbrace{\int_{\partial\Omega_\varepsilon} u \phi\, n_i \, ds}_{\lesssim\, 2\pi\varepsilon\left(\log\log\frac{2}{\varepsilon}\right)\|\phi\|_{L^\infty} \,\to\, 0 \text{ as } \varepsilon \to 0} - \int_{\Omega_\varepsilon} \partial_i u\, \phi \, dx$$

$\therefore \int_\Omega u\, \partial_i \phi \, dx = - \int_\Omega \partial_i u\, \phi \, dx$

$\therefore$ the classical derivative blows up slowly enough at the origin that it is a weak derivative.

$\therefore$ $u \in H^1(\Omega)$ as claimed.

Exercise: Show that for $n \geq 3$, $u(x) = r^{-\alpha}$ is an H^1 function on the unit ball for $\alpha < \frac{n-2}{2}$.

<u>def</u>: A subset X of a normed space Y is <u>dense</u> if

$$\forall y \in Y \text{ and } \varepsilon > 0 \quad \exists\, x \in X \text{ s.t. } \|x-y\| < \varepsilon .$$

The norm here is the Y norm (since $X \subseteq Y$, X inherits this norm)

<u>Example</u>: $\mathbb{Q}$ is dense in $\mathbb{R}$, but $\mathbb{R}$ is not dense in $\mathbb{C}$.

__Theorem:__ Let $\Omega \subset \mathbb{R}^n$ be an open set with piecewise smooth boundary, and let $m \geq 0$. Then $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$.

(i.e. $\forall\, u \in H^m(\Omega)$ and $\varepsilon > 0$ $\exists\, v \in C^\infty(\Omega) \cap H^m(\Omega)$ s.t. $\|u - v\|_m < \varepsilon$.)

__pf:__ see e.g. Evans' PDE book.

__def:__ $H_0^m(\Omega)$ = the closure of $C_c^\infty(\Omega)$ in $H^m(\Omega)$ w.r.t. the norm $\|\cdot\|_m$.

__Theorem:__ (Poincaré - Friedrichs inequality)

Suppose Ω is contained in an n-dim'l cube with side length s. Then

$$\|u\|_0 \leq \frac{s}{\sqrt{2}} |u|_1 \quad \text{for all } u \in H_0^1(\Omega).$$

__pf:__ may assume $\Omega \subset Q = \{x \in \mathbb{R}^n : 0 < x_i < s,\ i = 1, \dots, n\}$ by translation and rotation if necessary. ($\|u\|_0$ and $|u|_1$ are invariant under such changes of coordinates)

Let $u \in C_c^\infty(\Omega)$. Extend it to $C_c^\infty(Q)$ via $u = 0$ on $Q \setminus \Omega$

Then $u(x_1,\dots,x_n) = u(0,x_2,\dots,x_n) + \int_0^{x_1} \partial_1 u(t,x_2,\dots,x_n)\,dt$

Cauchy-Schwarz: $|u(x)|^2 \le \int_0^{x_1} 1^2\,dx \int_0^{x_1} |\partial_1 u(t,x_2,\dots,x_n)|^2\,dt$

$$\le x_1 \underbrace{\int_0^s |\partial_1 u(t,x_2,\dots,x_n)|^2\,dt}_{\text{a constant indep. of } x_1}$$

for fixed $x_2,\dots,x_n$:

$$\int_0^s |u(x)|^2\,dx_1 \le \underbrace{\left(\int_0^s x_1\,dx_1\right)}_{s^2/2}\left(\int_0^s |\partial_1 u(t,x_2,\dots,x_n)|^2\,dt\right)$$

integrate over other coords:

$$\|u\|_0^2 = \int_Q |u|^2\,dx \le \frac{s^2}{2}\int_Q |\partial_1 u(t,x_2,\dots,x_n)|^2\,dt\,dx_2\cdots dx_n$$

$$= \frac{s^2}{2}\int_Q |\partial_1 u|^2\,dx \le \frac{s^2}{2}\int_Q |\nabla u|^2\,dx = \frac{s^2}{2}|u|_1^2$$

This establishes the result $\left(\|u\|_0 \le \frac{s}{\sqrt{2}}|u|_1\right)$ for all $u \in C_c^\infty(\Omega)$, which is a dense subset of $H_0^1(\Omega)$.

Now let v be any function in $H_0^1(\Omega)$ and let $\varepsilon > 0$.
Since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, $\exists\, u \in C_c^\infty(\Omega)$ s.t. $\|u-v\|_1 \le \varepsilon$

Then $\|v\|_0 \le \|u\|_0 + \|v-u\|_0 \le \frac{s}{\sqrt{2}}|u|_1 + \varepsilon$

$$\le \frac{s}{\sqrt{2}}\left(|u-v|_1 + |v|_1\right) + \varepsilon \le \frac{s}{\sqrt{2}}|v|_1 + \left(\frac{s}{\sqrt{2}}+1\right)\varepsilon$$

Since ε is arbitrary, $\|v\|_0 \le \frac{s}{\sqrt{2}}|v|_1$ as claimed.

---

corollary: $|\cdot|_1$ is a norm equivalent to $\|\cdot\|_1$ on $H_0^1(\Omega)$ ← important - rules out constant functions

$|u|_1 \le \|u\|_1 = \left(\|u\|_0^2 + |u|_1^2\right)^{1/2} \le \left(\frac{s^2}{2}+1\right)^{1/2}|u|_1$

corollary: $a(u,v) = \int_\Omega \nabla u\cdot\nabla v\,dx$ is coercive on $H_0^1(\Omega)$. $\alpha\|u\|_1^2 \le \overbrace{a(u,u)}^{|u|_1^2}$, $\alpha = \left(1+\frac{s^2}{2}\right)^{-1}$

228B Lec 24

<u>Last time</u> • In 1d, $H^1(a,b) = \{ u:(a,b)\to\mathbb{R} \mid$ ~~u is equal a.e. to an absolutely~~ continuous fcn on $[a,b]$ and $u_x \in L^2(a,b)\}$

• In higher dimensions, $H^1(\Omega)$ contains functions that blow up

Today: Coercivity and the Lax-Milgram theorem

postpone to page 3.

Theorem: $C^\infty(\Omega)\cap H^m(\Omega)$ is dense in $H^m(\Omega)$ for any open set $\Omega \subseteq \mathbb{R}^n$

<u>exercise</u>: fill in the blank: this means: $\forall u \in H^m(\Omega)$ and $\varepsilon>0$, ________

<u>def</u>: $H^m_0(\Omega)$ = the closure of $C^\infty_c(\Omega)$ in $H^m(\Omega)$

$= \{ u \in H^m(\Omega) : \exists\, v_1, v_2, \ldots \in C^\infty_c(\Omega) \text{ s.t. } v_k \to u \text{ in } H^m(\Omega)\}$

<u>exercise</u>: $v_k \to u$ in $H^m(\Omega)$ means: $\forall \varepsilon>0\ \exists N$ s.t. ________

<u>in words</u>:

$H^m_0(\Omega)$ is the set of all functions $u:\Omega\to\mathbb{R}$ such that $\partial^\alpha u \in L^2(\Omega)$ for $|\alpha|\le m$ and $u=0$ on $\partial\Omega$.

i.e. these weak derivatives exist and belong to $L^2(\Omega)$.

$u=0$ is imposed by requiring that you can get arbitrarily close to u by a C^∞ function v with compact support in Ω

(i.e. v is zero outside of a $\underbrace{\text{closed and bounded}}_{\text{compact}}$ set $K \subseteq \Omega$)

The weak formulation of the Dirichlet problem is:

find $u \in H^1_0(\Omega)$ s.t. $a(u,v) = \langle l, v \rangle \quad \forall v \in H^1_0(\Omega)$

where $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$, $\langle l, v \rangle = \int_\Omega f v \, dx$

we need to show that $a(\cdot,\cdot)$ is bounded and coercive and that l is bounded.

To prove boundedness of a and l, we use Cauchy-Schwarz:

Theorem: Let H be a vector space with inner product $(\cdot,\cdot)$.
Then $|(x,y)| \le \|x\| \cdot \|y\| \quad \forall x,y \in H$

pf: if x or y is zero, we have $0 \le 0$ ✓

otherwise let $\lambda = \|x\|$, $\mu = \mathrm{sgn}((x,y))\|y\|$ and check that

$$0 \le (\mu x - \lambda y, \mu x - \lambda y) = \mu^2 (x,x) - 2\lambda\mu (x,y) + \lambda^2 (y,y)$$
$$= 2\|x\|^2\|y\|^2 - 2\|x\| \cdot \|y\| |(x,y)|$$

so $|(x,y)| \le \|x\| \cdot \|y\|$

($a(x,x) \ge 0$ but $a(x,x) = 0$ does not imply $x = 0$)

Corollary: Let $a(\cdot,\cdot)$ be a symmetric, positive semidefinite bilinear form on H.
Then $|a(x,y)| \le \|x\|_a \|y\|_a$ where $\|x\|_a = \sqrt{a(x,x)}$

pf: define the new inner product $((x,y)) = \varepsilon(x,y) + a(x,y)$

Cauchy-Schwarz implies $|((x,y))| \le \sqrt{((x,x))}\sqrt{((y,y))}$

Now take the limit as $\varepsilon \to 0$.

note: $\|\cdot\|_a$ is only a semi-norm if a is not positive definite

Claim: $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$ is bounded on $H^1(\Omega)$ (hence on $H^1_0(\Omega)$)

pf: $|a(u,v)| \le \sqrt{a(u,u)}\sqrt{a(v,v)} \le \sqrt{\|u\|_0^2 + a(u,u)}\sqrt{\|v\|_0^2 + a(v,v)}$

for any $u,v \in H^1(\Omega)$, $= \|u\|_1 \|v\|_1$

so $C=1$ works in $|a(u,v)| \le C\|u\|_1 \|v\|_1$

Claim: $\langle \ell, v\rangle = \int_\Omega f v \, dx$ is bounded on $H^1(\Omega)$ (is a constant)

pf: $|\langle \ell, v\rangle| \le \|f\|_0 \cdot \|v\|_0 \le \|f\|_0 \|v\|_1$ usually strict inequality (unless f

so $C = \|f\|_0$ works in $\langle \ell, v\rangle \le C\|v\|_1$ (so $\|\ell\|_1 \le \|f\|_0$)

Now we want to prove that $a(\cdot,\cdot)$ is coercive on $H^1_0(\Omega)$, i.e.

(*) $\exists \alpha > 0$ s.t. $\alpha\|u\|_1^2 \le a(u,u) \quad \forall u \in H^1_0(\Omega)$

This isn't true on all of $H^1(\Omega)$ as the function $u(x) \equiv 1$

satisfies $\|u\|_1^2 = \int_\Omega u^2 dx = \text{volume}(\Omega) > 0$

while $a(u,u) = \int_\Omega \nabla u \cdot \nabla u \, dx = 0$

Note that the issue in (*) is whether $\|u\|_0$ can be bounded by $|u|_1$:

$$\alpha\|u\|_1^2 = \alpha\left(\|u\|_0^2 + |u|_1^2\right) \overset{?}{\le} a(u,u) = |u|_1$$

need $\|u\|_0^2 \le \left(\frac{1}{\alpha} - 1\right)|u|_1^2 \quad \forall u \in H^1_0(\Omega)$

---

go back to page 1.

Theorem: (Poincaré Friedrichs inequality)

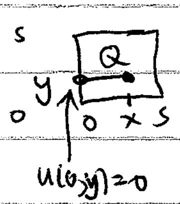
Suppose Ω is contained in an n-dimensional cube with side length s. Then

$$\|u\|_0 \le \frac{s}{\sqrt{2}} |u|_1 \quad \forall u \in H^1_0(\Omega)$$

proof in 2d: (see Lec 23 for n dimensions)

may assume $\Omega \subseteq Q = \{(x,y) : 0 < x < s,\ 0 < y < s\}$ by translation & rotation if necessary. ($\|u\|_0$ and $\|u\|_1$ are invariant under such changes of coordinates)

Let $u \in C_c^\infty(\Omega)$. Extend it to $C_c^\infty(Q)$ via $u=0$ on $Q \setminus \Omega$.



Then $u(x,y) = u(0,y) + \int_0^x u_x(t,y)\,dt$ ← FTOC

Cauchy-Schwarz: $|u(x,y)|^2 \le \int_0^x 1^2\,dt \int_0^x |u_x(t,y)|^2\,dt$

$$\le x \underbrace{\int_0^s |u_x(t,y)|^2\,dt}_{\text{a constant indep. of } x}$$

holding y fixed:

$$\int_0^s |u(x,y)|^2\,dx \le \underbrace{\left(\int_0^s x\,dx\right)}_{s^2/2} \left(\int_0^s |u_x(t,y)|^2\,dt\right)$$

now integrate in y-direction:

$$\|u\|_0^2 = \int_0^s\int_0^s |u(x,y)|^2\,dx\,dy \le \frac{s^2}{2}\int_0^s\int_0^s |u_x(t,y)|^2\,dt\,dy$$

$$= \frac{s^2}{2}\iint_Q |u_x|^2\,dx\,dy \le \frac{s^2}{2}\iint_Q |\nabla u|^2\,dx\,dy = \frac{s^2}{2}|u|_1^2$$

This establishes the result ($\|u\|_0 \le \frac{s}{\sqrt{2}}|u|_1$) for all $u \in C_c^\infty(\Omega)$ which is a dense subset of $H^1_0(\Omega)$

Now let v be any function in $H^1_0(\Omega)$ and let $\varepsilon > 0$.

Since $C_c^\infty(\Omega)$ is dense in $H^1_0(\Omega)$, $\exists\, u \in C_c^\infty(\Omega)$ s.t. $\|u-v\|_1 \le \varepsilon$

Then $\|v\|_0 \le \|u\|_0 + \|v-u\|_0 \le \frac{s}{\sqrt{2}}|u|_1 + \varepsilon$

$$\le \frac{s}{\sqrt{2}}\left(|u-v|_1 + |v|_1\right) + \varepsilon \le \frac{s}{\sqrt{2}}|v|_1 + \left(\frac{s}{\sqrt{2}}+1\right)\varepsilon$$

Since ε is arbitrary, $\|v\|_0 \le \frac{s}{\sqrt{2}}|v|_1$ as claimed.

Corollary: $|\cdot|_1$ is a norm on $H^1_0(\Omega)$ equivalent to $\|\cdot\|_1$

$$|u|_1 \le \|u\|_1 = \left(\|u\|_0^2 + |u|_1^2\right)^{1/2} \le \left(\frac{s^2}{2}+1\right)^{1/2}|u|_1$$

Corollary: $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$ is coercive on $H^1_0(\Omega)$ with with $\alpha = \left(1+\frac{s^2}{2}\right)^{-1}$

Lax-Milgram theorem:
- Let H be a Hilbert space.
- Suppose $a: H \times H \to \mathbb{R}$ is bounded, coercive & bilinear. and $l: H \to \mathbb{R}$ is bounded and linear.
- Then $\exists!\, u \in H$ s.t. $a(u,v) = \langle l, v\rangle \quad \forall v \in H$.

We'll prove it in the special case that $a(\cdot,\cdot)$ is also symmetric. (see e.g. Evans' PDE book for general case) based on Braess p38

idea: J attains its minimum at u iff $a(u,v) = \langle \ell, v\rangle \ \forall v$.
reason:
$J(u+tv) = J(u) + t[a(u,v) - \langle \ell, v\rangle] + \frac{1}{2}t^2 a(v,v)$

used symmetry of a here

Proof: define $J(u) = \frac{1}{2} a(u,u) - \langle \ell, u\rangle$

Note that $a(u,u) \geq \alpha \|u\|^2$

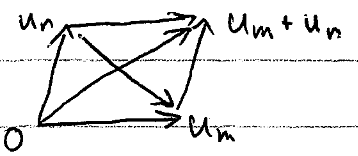
and $\langle \ell, u\rangle \leq \|\ell\| \cdot \|u\|$

so
$$J(u) \geq \frac{1}{2}\alpha \|u\|^2 - \|\ell\|\cdot\|u\|$$
$$= \frac{1}{2\alpha}\left(\alpha\|u\| - \|\ell\|\right)^2 - \frac{\|\ell\|^2}{2\alpha} \geq -\frac{\|\ell\|^2}{2\alpha}$$

$\therefore$ $J(u)$ is bounded from below. Let $J_0 = \inf\{J(u) : u \in H\}$ and let $\{u_k\}_{k=1}^{\infty}$ be a minimizing sequence.

Then $\alpha\|u_m - u_n\|^2 \leq a(u_m - u_n, u_m - u_n)$

Now use parallelogram law



expand it out... cross terms cancel.

$$a(u_m+u_n, u_m+u_n) + a(u_m-u_n, u_m-u_n) = 2a(u_m,u_m) + 2a(u_n,u_n)$$

$$\therefore \alpha\|u_m - u_n\|^2 \leq 2a(u_m,u_m) + 2a(u_n,u_n) - a(u_m+u_n, u_m+u_n)$$
$$= 4J(u_m) + 4J(u_n) - 8J\left(\frac{u_m+u_n}{2}\right)$$
$$\underbrace{+4\langle \ell, u_m\rangle + 4\langle \ell, u_n\rangle - 8\left\langle \ell, \frac{u_m+u_n}{2}\right\rangle}_{0}$$

But $J\left(\frac{u_m+u_n}{2}\right) \geq J_0$ since J_0 is the infemum. So

$$\alpha\|u_m - u_n\|^2 \leq 4J(u_m) + 4J(u_n) - 8J_0 \qquad \text{(subtracting less makes it bigger)}$$

Since $\{u_m\}$ is a minimizing sequence, for every $\varepsilon > 0$ there is an N s.t. $n \geq N \Rightarrow J(u_m) - J_0 < \frac{\varepsilon^2\alpha}{8}$

$\therefore$ if $m,n \geq N$ then $\|u_m - u_n\| \leq \varepsilon$ so $\{u_m\}$ is Cauchy.

Since H is complete, u_m converges to something, say u.

Since J is continuous (ℓ and a are bounded), we have

$$J(u) = J(\lim_n u_n) \overset{\text{continuity}}{=} \lim_n J(u_n) \overset{u_n \text{ is a minimizing sequence.}}{=} J_0$$

Since u minimizes J, it satisfies $a(u,v) = \langle \ell, v \rangle \ \forall v \in H$.

u is unique because two such minimizers u_1, u_2 could be strung together into a sequence $u_1, u_2, u_1, u_2, u_1, u_2, \ldots$ which is also a minimizing sequence, hence is Cauchy and converges. This is a contradiction unless $u_1 = u_2$.

<u>Note:</u> We've actually proved the Riesz representation theorem as a special case:

let $a(u,v) = (u,v)$

given $\ell \in H'$ ← dual space of H

$\exists!\, u$ s.t. $(u,v) = \langle \ell, v \rangle \ \forall v \in H$

Thus the canonical map from H to H' given by

$u \mapsto (u, \cdot)$ is onto (it's an isometry, actually)

(in the complex case, $u \mapsto (\cdot, u)$ is conjugate linear from H to H')

228B Lec 25

Last time: discussed HW problem 4

generalization of Cauchy-Schwarz to symmetric, positive semidefinite bilinear forms

$a(\cdot,\cdot)$ and $\langle \ell, \cdot \rangle$ are bounded

$a(\cdot,\cdot)$ is not coercive on all of $H^1(\Omega)$

definition of $H_0^1(\Omega)$ (closure of $C_c^\infty(\Omega)$ in $H^1(\Omega)$)

started proving Poincaré Friedrichs inequality

Today: finish Poincaré-Friedrichs proof

Lax-Milgram theorem

stability of finite elements

Céa's lemma

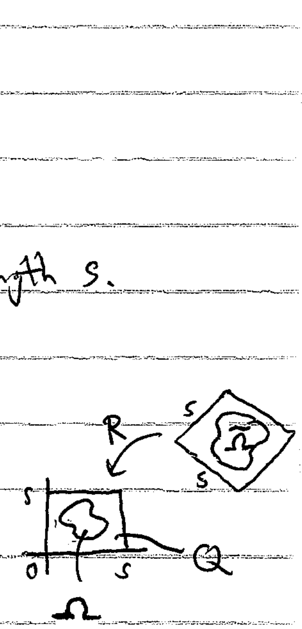
~~assembling the stiffness and mass matrices~~

Notational issues and hint on homework problem 3.

Theorem: (Poincaré-Friedrichs inequality)

Suppose $\Omega \subseteq \mathbb{R}^2$ is contained in a square with side length s.

(*) Then $\|u\|_0 \le \frac{s}{\sqrt{2}} |u|_1$ for all $u \in C_c^\infty(\Omega)$



proof so far: (1) translate and rotate Ω to fit inside $Q = [0,s]\times[0,s]$

(2) show that (*) holds for $u \in C_c^\infty(\Omega)$

(2a) use fundamental theorem of calculus for $u \in C_c^\infty(\Omega)$:

$$u(x,y) = \underbrace{u(0,y)}_{0} + \int_0^x u_x(t,y)\,dt$$

(2b) use Cauchy Schwarz

$$|u(x,y)|^2 \le \int_0^x 1^2\,dt \int_0^x |u_x(t,y)|^2\,dt \le x\int_0^s |u_x(t,y)|^2\,dt$$

(2c) integrate over Ω

$$\|u\|_0^2 = \iint_\Omega |u(x,y)|^2\,dx\,dy = \iint_Q |u(x,y)|^2\,dx\,dy \le \frac{s^2}{2}\iint_Q |u_x(t,y)|^2\,dt\,dy$$

$$\le \frac{s^2}{2}\iint_Q |\nabla u|^2\,dx\,dy = \frac{s^2}{2}|u|_1^2$$

This establishes the result for all $u \in C_c^\infty(\Omega)$, which is a dense subset of $H_0^1(\Omega)$.

(3) final step: extend result to all of $H_0^1(\Omega)$ using a <u>density argument</u>

Let v be any function in $H_0^1(\Omega)$ and let $\varepsilon > 0$. (v is like u here... not a test fn.)

Since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, $\exists u \in C_c^\infty(\Omega)$ s.t. $\|u-v\|_1 \le \varepsilon$.

Then $\|v\|_0 \le \|u\|_0 + \|v-u\|_0 \le \frac{s}{\sqrt{2}}|u|_1 + \varepsilon$

$$\le \frac{s}{\sqrt{2}}\left(|u-v|_1 + |v|_1\right) + \varepsilon \le \frac{s}{\sqrt{2}}|v|_1 + \left(\frac{s}{\sqrt{2}}+1\right)\varepsilon$$

The only way this can be true for all $\varepsilon > 0$ is if

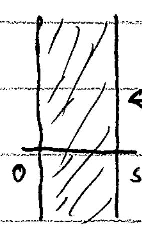
$\|v\|_0 \le \frac{s}{\sqrt{2}}|v|_1$ as claimed.

Corollary: $|\cdot|_1$ is a norm on $H^1_0(\Omega)$ equivalent to $\|\cdot\|_1$

pf:
$$|u|_1 \le \|u\|_1 = \left(\|u\|_0^2 + |u|_1^2\right)^{1/2} \le \left(\frac{s^2}{2}+1\right)^{1/2} |u|_1$$

Corollary: $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx\,dy$ is coercive on $H^1_0(\Omega)$ with $\alpha = \left(1+\frac{s^2}{2}\right)^{-1}$

Remark: this proof would work for Ω contained in a rectangle Q whose shortest side is s. It would even work over an infinite strip of width s.



Ω = infinite strip.

Remark: for a square (or n-cube in n dimensions) we can average the result in each direction and obtain

$$\|u\|_0 \le \frac{s}{\sqrt{2n}} |u|_1 \qquad u \in H^1_0(\Omega)$$

Lax-Milgram theorem: Let H be a Hilbert space. Suppose $a: H \times H \to \mathbb{R}$ is bounded, coercive and bilinear and $\ell: H \to \mathbb{R}$ is bounded and linear. Then $\exists!\, u \in H$ s.t. $a(u,v) = \langle \ell, v \rangle \;\; \forall v \in H$

unique

we'll prove it in the special case that $a(\cdot,\cdot)$ is also symmetric.
(see e.g. Evans' PDE book for the general case)

based on Braess p 38

idea: J attains its minimum at u iff $a(u,v) = \langle \ell, v \rangle \ \forall v$.

reason:

$J(u+tv) = J(u) + t[a(u,v) - \langle \ell, v\rangle] + \frac{1}{2} t^2 a(v,v)$

used symmetry of a here

<u>proof:</u> define $J(u) = \frac{1}{2} a(u,u) - \langle \ell, u \rangle$

Note that $a(u,u) \geq \alpha \|u\|^2$

and $\langle \ell, u \rangle \leq \|\ell\| \cdot \|u\|$

so $J(u) \geq \frac{1}{2} \alpha \|u\|^2 - \|\ell\| \cdot \|u\|$

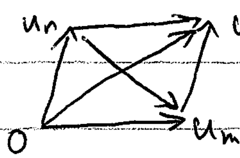
$= \frac{1}{2\alpha} \left(\alpha \|u\| - \|\ell\| \right)^2 - \frac{\|\ell\|^2}{2\alpha} \geq - \frac{\|\ell\|^2}{2\alpha}$

$\therefore J(u)$ is bounded from below. Let $J_0 = \inf \{ J(u) : u \in H \}$

and let $\{u_k\}_{k=1}^{\infty}$ be a minimizing sequence.

Then $\alpha \|u_m - u_n\|^2 \leq a(u_m - u_n, u_m - u_n)$

Now use parallelogram law



expand it out... cross terms cancel.

$a(u_m + u_n, u_m + u_n) + a(u_m - u_n, u_m - u_n) = 2a(u_m, u_m) + 2a(u_n, u_n)$

$\therefore \alpha \|u_m - u_n\|^2 \leq 2a(u_m, u_m) + 2a(u_n, u_n) - a(u_m + u_n, u_m + u_n)$

$= 4J(u_m) + 4J(u_n) - 8J\left(\frac{u_m + u_n}{2}\right)$

$\underbrace{+4\langle \ell, u_m \rangle + 4 \langle \ell, u_n \rangle - 8 \langle \ell, \frac{u_m + u_n}{2} \rangle}_{0}$

But $J\left(\frac{u_m + u_n}{2}\right) \geq J_0$ since J_0 is the infemum. So

$\alpha \|u_m - u_n\|^2 \leq 4J(u_m) + 4J(u_n) - 8J_0$ (subtracting less makes it bigger)

Since $\{u_m\}$ is a minimizing sequence, for every $\varepsilon > 0$ there is an N s.t. $n \geq N \Rightarrow J(u_m) - J_0 < \frac{\varepsilon^2 \alpha}{8}$

$\therefore$ if $m, n \geq N$ then $\|u_m - u_n\| \leq \varepsilon$ so $\{u_m\}$ is Cauchy.

Since H is complete, u_m converges to something, say u.

Since J is continuous (l and a are bounded), we have

$$J(u) = J(\lim_n u_n) \overset{\text{continuity}}{=} \lim_n J(u_n) \overset{u_n \text{ is a minimizing sequence.}}{=} J_0$$

Since u minimizes J, it satisfies $a(u,v) = \langle l, v\rangle \ \forall v \in H$.

u is unique because two such minimizers u_1, u_2 could be strung together into a sequence $u_1, u_2, u_1, u_2, u_1, u_2, \ldots$ which is also a minimizing sequence, hence is Cauchy and converges. This is a contradiction unless $u_1 = u_2$.

<u>Note</u>: We've actually proved the Riesz representation theorem as a special case:

let $a(u,v) = (u,v)$

given $l \in H^*$ ← dual space of H

$\exists!\, u$ s.t. $(u,v) = \langle l, v\rangle \ \forall v \in H$

Thus the canonical map from H to H^* given by

$u \mapsto (u, \cdot)$ is onto (it's an isometry, actually)

(in the complex case, $u \mapsto (\cdot, u)$ is conjugate linear from H to H^*)

finite element space

Note: Lax-Milgram applies equally well to the subspace $S_h \subseteq H_0^1(\Omega)$ so we get existence and uniqueness of the weak formulations of the continuous and discrete problems:

$$(*) \quad \exists!\, u \in H_0^1(\Omega) \text{ s.t. } a(u,v) = \langle \ell, v\rangle \quad \forall v \in H_0^1(\Omega)$$
$$\exists!\, u_h \in S_h \text{ s.t. } a(u_h,v) = \langle \ell, v\rangle \quad \forall v \in S_h$$

Stability: the solutions u and u_h are bounded by the norm of f

$\begin{pmatrix} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{pmatrix}$

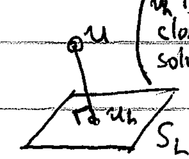
pf: $\alpha \|u\|_1^2 \le a(u,u) = \langle \ell, u\rangle \le \|\ell\|_1 \cdot \|u\|_1$
$\le \|f\|_0 \cdot \|u\|_1$

$\therefore \|u\|_1 \le \frac{1}{\alpha}\|f\|_0$

similarly, $\|u_h\|_1 \le \frac{1}{\alpha}\|f\|_0$

Error analysis (Cea's lemma) the FE solution is within a constant of the best possible approximation in the FE space S_h:

$$\|u - u_h\|_1 \le \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_1$$



(u_h is closest solution in $\|\cdot\|_a$)

pf: from (*) we have $a(u-u_h, v) = 0 \quad \forall v \in S_h$. (Galerkin orthogonality)

Thus $\alpha\|u-u_h\|_1^2 \le a(u-u_h, u-u_h)$ ($v_h \in S_h$ is arbitrary)

$$= a(u-u_h, u-\underbrace{v_h + v_h}_{0} - u_h)$$
$$= a(u-u_h, u-v_h) + \underbrace{a(u-u_h, v_h-u_h)}_{0 \text{ by Galerkin orthog.}}$$
$$\le C\|u-u_h\| \cdot \|u-v_h\|$$

$\therefore \|u-u_h\| \le \frac{C}{\alpha}\|u-v_h\|$ for every $v_h \in S_h$. Now take infimum of RHS.

Notational issues:

$$|(u,v)| \le \|u\| \cdot \|v\| \quad \leftarrow \text{Cauchy Schwarz (inner product)}$$

$$|\langle \ell, v\rangle| \le \|\ell\| \cdot \|v\| \quad \leftarrow \text{definition of norm of } \ell$$

(the brakets are anoth way to write $\ell(v)$)

subscripts:

$$\mathbb{R}^n: \quad \|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$

$$L^p: \quad \|u\|_{L^p} = \left(\iint_\Omega |u(x,y)|^p \, dx\,dy\right)^{1/p}$$

$$H^m: \quad \|u\|_{H^m} = \left(\sum_{|\alpha|\le m} \iint_\Omega |\partial^\alpha u|^2 \, dx\,dy\right)^{1/2}$$

$\|u\|_2$ could mean $\|u\|_{L^2}$ or $\|u\|_{H^2}$

$H^0 = L^2$

Lately I have been writing $\begin{cases} \|u\|_0 = \|u\|_{L^2} \\ \|u\|_2 = \|u\|_{H^2} \end{cases}$

we have:

$$\begin{array}{ccccccc} L^2(\Omega) = H^0(\Omega) & \supseteq & H^1(\Omega) & \supseteq & H^2(\Omega) & \supseteq & \cdots \\ \| & & \cup\!\mid & & \cup\!\mid & & \cup\!\mid \\ H_0^0(\Omega) & \supseteq & H_0^1(\Omega) & \supseteq & H_0^2(\Omega) & \supseteq & \cdots \end{array}$$

Hölder's inequality generalizes Cauchy-Schwarz for L^p spaces.
In $\mathbb{R}^n$, it looks like

$$|b^T x| \le \|b\|_q \|x\|_p, \quad \|b\|_q = \left(\sum_i |b_i|^q\right)^{1/q}$$

equality iff w and z are linearly dependent with $w_i = \mathrm{sgn}(x_i)|x_i|^p$, $z_i = \mathrm{sgn}(b_i)|b_i|^q$

$$\frac{1}{p} + \frac{1}{q} = 1 \quad (p \,\&\, q \text{ are conjugate exponents})$$

dual spaces: If X is a normed $\overbrace{\text{linear}}^{\text{vector}}$ space, its dual space is

$$X^* = \{\text{bounded linear functionals on } X\}$$

the norm on X^* is $\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|}$

X^* is complete even if X is not complete.

continuity and boundedness: a linear mapping $f: X \to \mathbb{R}$ is bounded iff it is continuous.

$\Leftarrow)$ $|f(x) - f(y)| = |f(x-y)| \leq \|f\| \, \|x-y\|$

$\Rightarrow)$ if f is continuous at 0, $\exists \delta > 0$ s.t. $|f(y) - \underbrace{f(0)}_{0}| < \overset{\varepsilon = 1.}{1}$ $\forall \|y\| < \delta$.

so if $x \neq 0$, then $y = \frac{1}{2}\delta \frac{x}{\|x\|}$ satisfies $\|y\| = \delta/2 < \delta$

and $|f(y)| = \frac{1}{2}\delta \frac{|f(x)|}{\|x\|} < 1 \Rightarrow \frac{|f(x)|}{\|x\|} \leq \frac{2}{\delta}$ $\forall x \neq 0$.

Subspaces and extensions. If $M \subseteq X$ and $F \in X^*$

then $f = F|_M \in M^*$ restriction of F to M (i.e. $F(x) = f(x)$ $\forall x \in M$)

and $\|f\| \leq \|F\|$.

The Hahn-Banach theorem says that every linear functional on M is of this form:

HBT: Given $f \in M^*$ $\exists F \in X^*$ s.t. $F|_M = f$ and $\|F\| = \|f\|$.

One of these can't be proved for all f satisfying the hypotheses. In this case you need to give a counterexample.

In problem 3a, we have $M = C^\infty(\Omega) \cap H^1(\Omega)$

$$X = H^1(\Omega)$$

and I gave you a linear functional f on M that is bounded when M is equipped with the L^2 norm:

$$|f(u)| \le C\|u\|_0 \quad \forall u \in M \qquad H^0 = L^2$$

H^1

to apply the HBT you need to show f is bounded with respect to the norm inherited from X, namely $\|\cdot\|_1$

to be shown: $\exists C_1$ s.t. $|f(u)| \le C_1 \|u\|_1 \quad \forall u \in M$

Now HBT $\Rightarrow$ $\exists F \in X^*$ s.t. $F(u) = f(u) \quad \forall u \in M$

In problem 3b, we have $M = C^\infty(\Omega) \cap H^1(\Omega)$

$$X = L^2(\Omega)$$

Now all you know is that $f \in M^*$ when M is equipped with the H^1 norm:

$$|f(u)| \le C\|u\|_1 \quad \forall u \in M$$

and before we can apply the HBT, we have to show that f is bdd w.r.t. the norm inherited from X:

to be shown: $\exists C_1$ s.t. $|f(u)| \le C_1 \|u\|_0 \quad \forall u \in M$

Now HBT $\Rightarrow$ $\exists F \in X^*$ s.t. $F(u) = f(u) \quad \forall u \in M$

Remark: the HBT is not really necessary in this problem since M is dense in X; but it makes the proofs a lot easier.

So how would you ~~prove that~~ produce a counterexample?

here's an example to model your proof on:

Let $M = C[0,1]$, $X = L^1(0,1)$

given $f \in M^*$ when M is equipped with the max norm, i.e.

$$|f(u)| \le C\|u\|_\infty$$

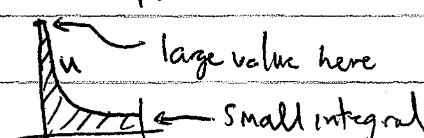
can f be extended continuously to an $F \in X^*$?

<u>Answer</u>: no. Counterexample: $f(u) = u(0)$.

we have $|f(u)| \le \max\limits_{0\le x\le 1} |u(x)|$ so $C=1$ works ~~here~~

but f is not bounded on M when M is equipped with L^1 norm.

Try to reach contradiction: Suppose $\exists\, C_1$ s.t. $|f(u)| \le C_1\|u\|_{L^1}$ $\forall u \in M$.

look for bad u: 

how about $u(x) = \dfrac{1}{\varepsilon + \sqrt{x}}$

Then $|f(u)| = \dfrac{1}{\varepsilon}$ while $\|u\|_{L^1} = \int_0^1 |u(x)|\,dx \le \int_0^1 \frac{1}{\sqrt{x}}\,dx = 2$ $\quad 2x^{1/2}\Big|_0^1$

With $\varepsilon = \frac{1}{2}(C_1+1)^{-1}$ we have $|f(u)| = 2C_1 + 2$

while $C_1\|u\|_{L^1} \le 2C_1$ contradicting

so f can't extend to $F \in X^*$ since the proposed F isn't even bounded on M.

In your homework, ~~Math~~ consider $f(u) = u(1) - u(0)$. ← why is it bounded in $H^1(0,1)$?
(or in the homework notation, $\langle \ell, u\rangle = u(1) - u(0)$.)

228B Lec 26

Last time: finished proving Poincaré Friedrichs inequality $\|u\|_0 \le \frac{S}{\sqrt{2}}|u|_1$, $\forall u \in H_0^1(\Omega)$

proved Lax-Milgram theorem in the symmetric case $a(u,v) = a(v,u)$

established connection between minimizing $J(u) = \frac{1}{2}a(u,u) - \langle l,u\rangle$

and solving $a(u,v) = \langle l,v\rangle \quad \forall v \in H$.

Today: stability of finite elements

Céa's lemma

implementation issues

Recap: Lax-Milgram: let H be a Hilbert space, $a: H\times H \to \mathbb{R}$ a bounded, coercive bilinear form, $l: H \to \mathbb{R}$ a bounded linear functional. Then $\exists!\, u \in H$ s.t. $a(u,v) = \langle l,v\rangle \quad \forall v \in H$.

pf: 1. $J(u) = \frac{1}{2}a(u,u) - \langle l,u\rangle$

we used symmetry of a here.

$$J(u+tv) = J(u) + t\underbrace{\left[a(u,v) - \langle l,v\rangle\right]}_{\text{first variation of } J \text{ in } v \text{ direction:}} + \frac{1}{2}t^2 a(v,v)$$

u minimizes J iff $\uparrow = 0 \quad \forall v \in H$.

$$DJ(u)v = a(u,v) - \langle l,v\rangle = \iint \frac{\delta J}{\delta u} v\, dx\, dy, \qquad \boxed{\frac{\delta J}{\delta u} = -\Delta u - f}$$

2. coercivity & parallelogram law $\Rightarrow$ any minimizing sequence for J is a Cauchy sequence

3. completeness of H $\Rightarrow$ this sequence converges to something, say u.

4. u minimizes J, hence satisfies $a(u,v) = \langle l,v\rangle \quad \forall v \in H$

5. u is unique since two minimizers u_1 & u_2 can be strung together to form a minimizing sequence (which is then Cauchy)

Example: coercivity is important! $H = \ell^2$, $\|x\| = \sum_{k=1}^{\infty} x_k^2$.

$$a(x,y) := \sum_{k=1}^{\infty} 2^{-k} x_k y_k$$

is positive and continuous, but not coercive.

$$\langle f, x\rangle := \sum_{k=1}^{\infty} 2^{-k} x_k$$

is a bounded linear functional since $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots) \in \ell^2$.

But $J(x) = \frac{1}{2}a(x,x) - \langle f, x\rangle$ does not attain a minimum in ℓ^2: the only way for

$$a(x,y) = \langle f, y\rangle \quad \forall y \in \ell^2$$

is for $x = (1,1,1,\dots)$, which does not belong to ℓ^2.

Remember, the notation $\langle f, \cdot\rangle$ is just a fancy way of writing $f(\cdot)$. Here f is just a linear functional on H (we don't always have to use the letter l in this notation) finite element space

Note: Lax-Milgram applies equally well to the subspace $S_h \subseteq H_0^1(\Omega)$ so we get existence and uniqueness of the weak formulations of the continuous and discrete problems:

$$(*) \quad \begin{aligned} &\exists!\, u \in H_0^1(\Omega) \ \text{s.t.} \ a(u,v) = \langle l, v\rangle \quad \forall v \in H_0^1(\Omega) \\ &\exists!\, u_h \in S_h \ \text{s.t.} \ a(u_h,v) = \langle l, v\rangle \quad \forall v \in S_h \end{aligned}$$

Stability: the solutions u and u_h are bounded by the norm of f

$$\begin{pmatrix} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{pmatrix}$$

pf: $\alpha \|u\|_1^2 \le a(u,u) = \langle l, u\rangle \le \|l\|_1 \cdot \|u\|_1$

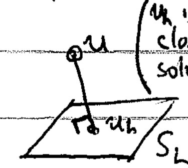
$$\le \|f\|_0 \cdot \|u\|_1$$

$\therefore \ \|u\|_1 \le \frac{1}{\alpha}\|f\|_0$

similarly, $\|u_h\|_1 \le \frac{1}{\alpha}\|f\|_0$

Error analysis (Cea's lemma) the FE solution is within a constant of the best possible approximation in the FE space S_h:

$$\|u - u_h\|_1 \le \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_1$$



(u_h is closest solution in $\|\cdot\|_a$)

pf: from (*) we have $a(u-u_h, v) = 0 \quad \forall v \in S_h$. (Galerkin orthogonality)

Thus $\alpha\|u-u_h\|_1^2 \le a(u-u_h, u-u_h)$ ($v_h \in S_h$ is arbitrary)

$$= a(u-u_h, u \underbrace{- v_h + v_h}_{0} - u_h)$$

$$= a(u-u_h, u-v_h) + \underbrace{a(u-u_h, v_h-u_h)}_{0 \text{ by Galerkin orthog.}}$$

$$\le C\|u-u_h\| \cdot \|u-v_h\|$$

$\therefore \ \|u-u_h\| \le \frac{C}{\alpha}\|u-v_h\|$ for every $v_h \in S_h$. Now take infimum of RHS.

So how do we actually solve the discrete system: find $u_h \in S_h$ s.t

$$(*) \qquad a(u_h, v) = \langle \ell, v \rangle \quad \forall v \in S_h \,?$$

Choose a basis for S_h, say $\varphi_1, \varphi_2, \ldots, \varphi_N$. Then $(*)$ is equivalent to

$$a(u_h, \varphi_i) = \langle \ell, \varphi_i \rangle \qquad i = 1 \ldots N$$

writing $u_h = \sum_{j=1}^{N} u_j \varphi_j$ we obtain the system of equations

$$Au = b \quad , \quad A_{ij} = a(\varphi_j, \varphi_i) \quad , \quad b_i = \langle \ell, \varphi_i \rangle$$

A is symmetric, positive definite (hence invertible):

$$u^T A u = \sum_{i,j} u_i A_{ij} u_j = a\Big(\sum_j u_j \varphi_j, \sum_i u_i \varphi_i\Big)$$
$$= a(u_h, u_h) \geq \alpha \|u_h\|_1^2 > 0 \text{ unless } u = 0.$$

Which basis should we choose?

right now we're doing conforming FE, so we require the $\varphi_i(x,y)$ to belong to $H_0^1(\Omega)$.

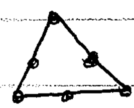
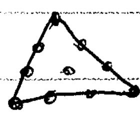
one may show that a piecewise polynomial fcn. belongs to $H^1(\Omega)$ iff it is continuous across the edges of the mesh. (discontinuities in slope are OK: but the value can't jump)

A convenient basis to use is a <u>nodal basis</u>: $\varphi_i(\vec{x}_j) = \delta_{ij}$

$\uparrow$ nodes of the mesh

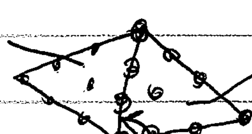
the support of these basis functions is limited to the elements touching the relevant node $\varphi_i \leftrightarrow \vec{x}_i$

for triangles, there's a 1-1 correspondence between polynomials of degree p and their values at uniformly spaced points:

	# nodes = # of polynomials				
	1		$p=0$	constant fcns	1
P_1	3		$p=1$	linear fcns	$1 \quad x \quad y$
P_2	6		$p=2$	quadratic fcns	$1 \; x \; y \; x^2 \; xy \; y^2$
P_3	10		$p=3$	cubic fcns	$1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3$

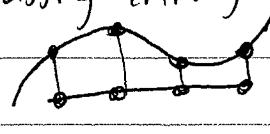
For linear and higher elements, the value of the function along an edge is uniquely determined by its values at the nodes on that edge. So continuity across edges (between nodes) is automatic.

cubic fcn of (x,y)

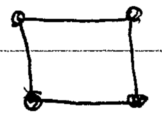


cubic fcn of (x,y)

along this edge, the cubic functions of two variables on either side match up with the unique cubic function of 1 variable passing through those 4 nodal values

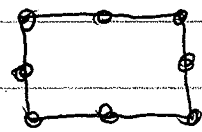


for quadrilateral elements, the # of degrees of freedom don't match so nicely:

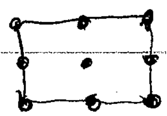


bilinear quadrilateral element (Q_1)

$1, x, y, xy$ (so not all 2nd order polynomials are used)



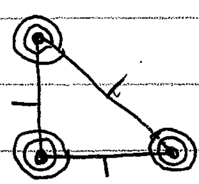
serendipity element $1 \quad x \quad y \quad xy \quad x^2 \quad y^2 \quad x^2y \quad xy^2$



biquadratic element (Q_2)

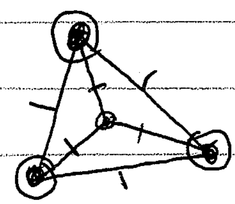
$1 \quad x \quad y \quad xy \quad x^2 \quad y^2 \quad x^2y \quad xy^2 \quad x^2y^2$

These are the most commonly used C^0 elements. For some problems, we need C^1 elements (so the basis functions belong to $H^2(\Omega)$) Example: biharmonic equation $\Delta^2 u = f$



Argyris triangle, 21 d.o.f., (all 5th order polynomials)

- ' means match normal derivative
- • match value
- ⊙ match gradients (u_x, u_y)
- ◯ match 2nd derivs (u_{xx}, u_{xy}, u_{yy})



Clough-Tocher 12 dof

macro-element (cubic on each sub-element)

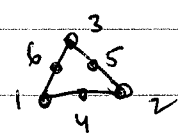
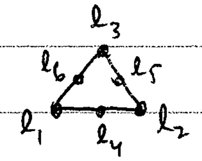
$30 = 3(6) + 3 + 3 + 2(3)$

Rather than compute $A_{ij} = a(\varphi_j, \varphi_i)$ basis function by basis function (i.e. node by node), it's more efficient to do it element by element.

$$A_{ij} = \sum_{T \in \mathcal{T}} a_T(\varphi_j, \varphi_i)$$ $\mathcal{T}$ = set of triangles in mesh

$$a_T(u,v) = \iint_T \nabla u \cdot \nabla v \, dx\,dy$$

On each triangle we compute a local stiffness matrix (with nodes numbered 1..np) and then <u>add</u> these entries to the global stiffness matrix in the appropriate rows and columns

local numbering  global numbering 

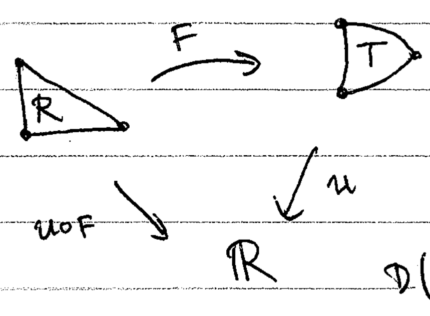
$$A^{loc}_{ij} = a_T(\varphi_i, \varphi_j)$$

for i=1..6
 for j=1..6
  $A_{\ell_i \ell_j} \mathrel{+}= A^{loc}_{ij}$

we add because several triangles T are likely to affect each entry of A

A^{loc} is a full np×np matrix but A is very sparse (so be sure to use a sparse matrix for A)

Assembly of the local stiffness matrix is usually done using the change of variables formula. in the reference triangle



$$\iint_T uv \, dx\,dy = \iint_R uv \underbrace{|\det DF|}_{\text{Jacobian}} d\xi\,d\eta$$ i.e. $u \circ F$, $v \circ F$

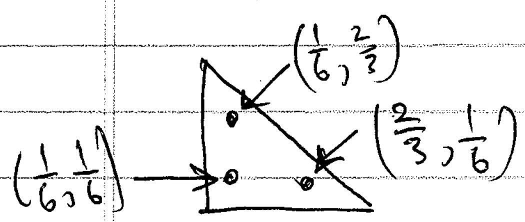
$$\iint_T \nabla_x u \cdot \nabla_x v \, dx\,dy = \iint_R \left(\nabla_\xi u \circ (DF)^{-1}\right) \cdot \left(\nabla_\xi v \cdot (DF)^{-1}\right) |\det DF| \, d\xi\,d\eta$$

$D(u \circ F) = Du \circ DF \rightarrow \nabla_\xi u = \nabla_x u \cdot DF$

Numerical quadrature

to actually do the integrals over the reference triangle, we use Gaussian quadrature:

example 3 point G.Q. rule:

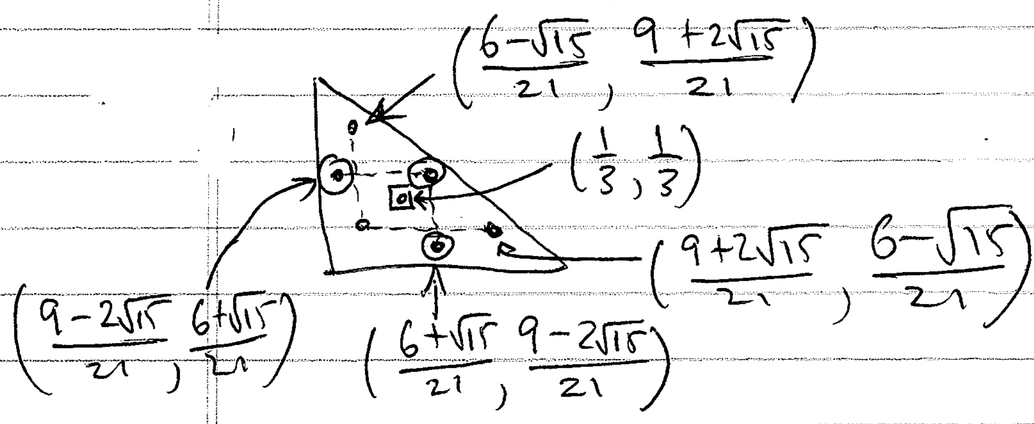


$A = \frac{1}{2}$ for ref. tri.

equal weights $w_i = \frac{A}{3} = \frac{1}{6}$

integrates polynomials of deg ≤ 2 exactly.

example from book: 7 pt G.Q. rule



$wt_{\boxed{\cdot}} = 9/180$

$wt_{\odot} = \frac{155+\sqrt{15}}{2400}$

$wt_{\circ} = \frac{155-\sqrt{15}}{2400}$

sum = $\frac{1}{2}$

Integrates polynomials of deg ≤ 5 exactly

in hw7 directory, I give you several G.Q. rules:

n	d
3	2
7	5
16	8
37	13
73	19

gauss02
05
08
13
19

these high order ones aren't so easy to find in the literature!

Last time: coercivity is important in Lax-Milgram (example)

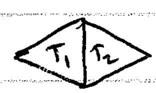
stability: FE sol'n u_h is bdd in terms of data f: $\|u_h\|_1 \le \frac{1}{\alpha}\|f\|_0$

Céa's lemma: $\|u-u_h\|_1 \le \frac{1}{\alpha} \inf_{v_h \in S_h} \|u-v_h\|_1$ — best possible approx in S_h

choose a basis, get positive definite linear system $Au=b$

today: implementation details

def: a function $u: \Omega_h \to \mathbb{R}$ is piecewise smooth if its restriction to each triangle is a C^∞ function with derivatives that extend continuously to the boundary of the triangle. The limit need not be the same when approached from a different triangle.

e.g. 

$$u(x,y) = \begin{cases} x^2 + 7y & (x,y) \in T_1 \\ 3e^y & (x,y) \in T_2 \end{cases}$$

is piecewise smooth. It doesn't matter how you define the function on the boundaries of the triangles. What does matter is that once you pick a triangle T, you can re-define u and its derivatives on ∂T to get continuous functions on all of T (including ∂T)

theorem: A piecewise smooth function $u: \Omega_h \to \mathbb{R}$ belongs to $H^m(\Omega_h)$ iff $u \in C^{m-1}(\Omega_h)$, i.e. the limit of $\partial^\alpha u$ on the boundary of adjacent triangles is the same when approached from either triangle for $|\alpha| \le m-1$.

in particular, if u is piecewise smooth, then $u \in H^1(\Omega) \iff u \in C^0(\Omega)$

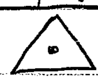
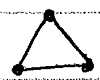
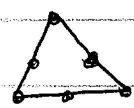
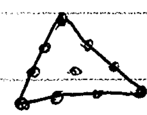
i.e. u is continuous across triangles but its derivatives can jump.

A convenient basis to use is a <u>nodal basis</u>: $\varphi_i(\vec{x}_j) = \delta_{ij}$

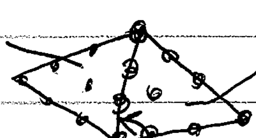
nodes of the mesh

the support of these basis functions is limited to the elements touching the relevant node $\varphi_i \leftrightarrow \vec{x}_i$

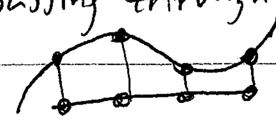
for triangles, there's a 1-1 correspondence between polynomials of degree p and their values at uniformly spaced points:

	# nodes = # of polynomials				
	1		$p=0$	constant fcns	1
$\mathcal{P}_1$	3		$p=1$	linear fcns	$1 \quad x \quad y$
$\mathcal{P}_2$	6		$p=2$	quadratic fcns	$1 \quad x \quad y \quad x^2 \quad xy \quad y^2$
$\mathcal{P}_3$	10		$p=3$	cubic fcns	$1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3$

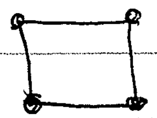
For linear and higher elements, the value of the function along an edge is uniquely determined by its values at the nodes on that edge. So continuity across edges (between nodes) is automatic.



along this edge, the cubic functions of two variables on either side match up with the unique cubic function of 1 variable passing through those 4 nodal values

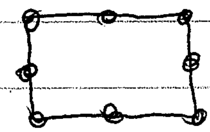


for quadrilateral elements, the # of degrees of freedom don't match so nicely:

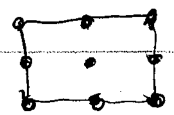


bilinear quadrilateral element (Q_1)

$1, x, y, xy$ (so not all 2nd order polynomials are used)



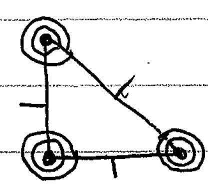
serendipity element $\quad 1 \quad x \quad y \quad xy \quad x^2 \quad y^2 \quad x^2y \quad xy^2$



biquadratic element (Q_2)

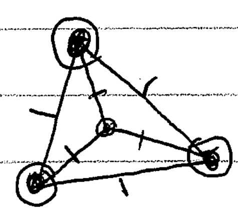
$1 \quad x \quad y \quad xy \quad x^2 \quad y^2 \quad x^2y \quad xy^2 \quad x^2y^2$

These are the most commonly used C^0 elements. For some problems, we need C^1 elements (so the basis functions belong to $H^2(\Omega)$) Example: biharmonic equation $\Delta^2 u = f$



Argyris triangle, 21 d.o.f., (all 5th order polynomials)

- ′ means match normal derivative
- • match value
- ⊙ match gradients (u_x, u_y)
- ◯ match 2nd derivs (u_{xx}, u_{xy}, u_{yy})



Clough-Tocher 12 dof

macro-element (cubic on each sub-element)

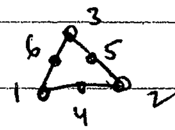
$30 = 3(6) + 3 + 3 + 2(3)$

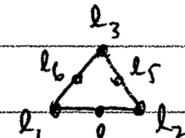
Rather than compute $A_{ij} = a(\varphi_j, \varphi_i)$ basis function by basis function (i.e. node by node), it's more efficient to do it element by element.

$$A_{ij} = \sum_{T \in \mathcal{T}} a_T(\varphi_j, \varphi_i) \qquad \mathcal{T} = \text{set of triangles in mesh}$$

$$a_T(u,v) = \iint_T \nabla u \cdot \nabla v \, dx\,dy$$

On each triangle we compute a local stiffness matrix (with nodes numbered $1..np$) and then add these entries to the global stiffness matrix in the appropriate rows and columns

local numbering 

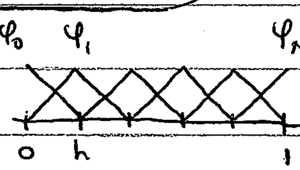
global numbering 

np × np matrix → $A^{loc}_{ij} = a_T(\varphi_{l_j}, \varphi_{l_i})$

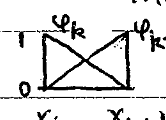
for $i=1..6$
 for $j=1..6$
  $A_{l_i l_j} \mathrel{+}= A^{loc}_{ij}$

we add because several triangles T are likely to affect each entry of A

A^{loc} is a full $np \times np$ matrix but A is very sparse (so be sure to use a sparse matrix for A)

Example : 1d linear elements 

Let $I = [x_k, x_{k+1}]$ ← in 1d the elements are intervals instead of triangles

There are two basis functions with support on this interval: 

$\frac{\partial \varphi_k}{\partial x} = -\frac{1}{h}$

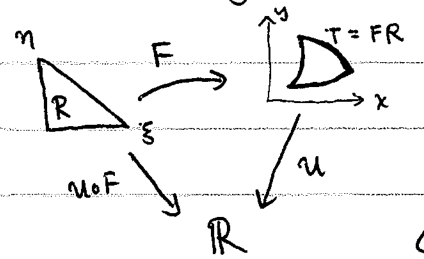
$\frac{\partial \varphi_{k+1}}{\partial x} = \frac{1}{h}$

$$A^{loc}_{ij} = a_I(\varphi_{l_j}, \varphi_{l_i}) = \int_{x_k}^{x_{k+1}} \frac{\partial \varphi_{l_j}}{\partial x}\frac{\partial \varphi_{l_i}}{\partial x}\,dx \to A^{loc} = \frac{1}{h}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

local to global mapping: $l_1 = k$, $l_2 = k+1$

sum the contributions from each interval → $A = \frac{1}{h}\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & & -1 & 1 \end{pmatrix}$

In practice, it's often convenient to compute the local stiffness matrix using a reference element



composition $u\circ F(\xi,\eta) = u(F(\xi,\eta))$

$$\iint_T u \, dx\,dy = \iint_R \overbrace{u\circ F}\ \underbrace{|\det DF|}_{\text{Jacobian}}\, d\xi\, d\eta$$

change of variables formula

$$A^{loc}_{ij} = a_T(\varphi_{\ell_j}, \varphi_{\ell_i}) = \iint_T \underbrace{\nabla_x \varphi_{\ell_j} \cdot \nabla_x \varphi_{\ell_i}}_{u \text{ in the formula above}} \, dx\,dy$$

Let $\psi_i(\xi,\eta) = \varphi_{\ell_i}(F(\xi,\eta))$ $\qquad \xi \leftrightarrow \xi_1 \quad \eta \leftrightarrow \xi_2 \qquad x \leftrightarrow x_1 \quad y \leftrightarrow x_2$

3 flavors of chain rule: $(\psi = \varphi\circ F)$

$$\begin{cases} \dfrac{\partial \psi}{\partial \xi_j} = \sum_k \dfrac{\partial \varphi}{\partial x_k}\dfrac{\partial x_k}{\partial \xi_j} \\ D\psi = D\varphi \cdot DF \\ (\nabla_\xi \psi)^T = (\nabla_x \varphi)^T \cdot DF \\ \underline{\quad} = \underline{\quad} \cdot \boxed{} \end{cases}$$

$\vec{x} = F(\vec{\xi})$

$$DF = \begin{pmatrix} \dfrac{\partial x}{\partial \xi} & \dfrac{\partial x}{\partial \eta} \\ \dfrac{\partial y}{\partial \xi} & \dfrac{\partial y}{\partial \eta} \end{pmatrix}$$

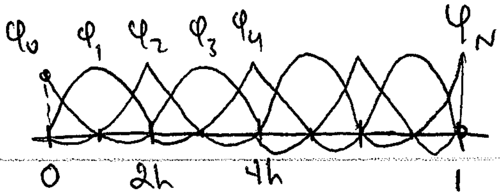
The last one gives $(\nabla_x \varphi)^T = (\nabla_\xi \psi)^T \cdot (DF)^{-1}$ or $\nabla_x \varphi = (DF)^{-T} \nabla_\xi \psi$

$$\therefore\ A^{loc}_{ij} = \iint_R \left[(DF)^{-T} \nabla_\xi \psi_j\right] \cdot \left[(DF)^{-T}\nabla_\xi \psi_i\right] |\det DF| \, d\xi\, d\eta$$

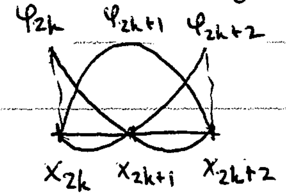
dot product

The integrand is now a function of $\vec{\xi}$ only, and the basis functions $\psi_1, \dots, \psi_{np}$ do not change from triangle to triangle... only the mapping F from R to T changes.

Example: 1d quadratic elements



$I = [x_{2k}, x_{2k+2}]$



$x = F(\xi) = x_{2k} + 2h\xi$

$DF(\xi) = 2h$

$\xi = 0 \quad \frac{1}{2} \quad 1$

$$\psi_1(\xi) = \frac{(\frac{1}{2}-\xi)(1-\xi)}{(\frac{1}{2})(1)} = \frac{\frac{1}{2} - \frac{3}{2}\xi + \xi^2}{\frac{1}{2}} = 2\xi^2 - 3\xi + 1$$

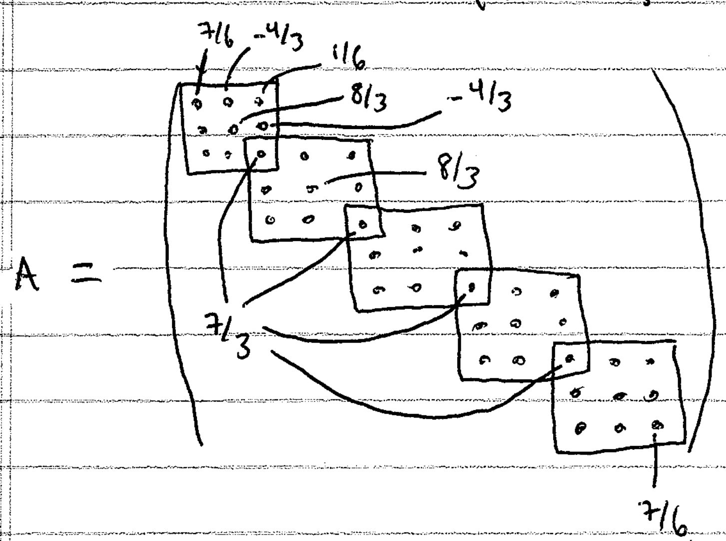
$$\psi_2(\xi) = \frac{(\xi)(1-\xi)}{(\frac{1}{2})(\frac{1}{2})} = -4\xi^2 + 4\xi$$

$$\psi_3(\xi) = \frac{(\xi)(\xi - \frac{1}{2})}{(1)(\frac{1}{2})} = 2\xi^2 - \xi$$

$$A_{ij}^{loc} = \int_0^1 \left[(2h)^{-1}\frac{\partial \psi_j}{\partial \xi}\right]\left[(2h)^{-1}\frac{\partial \psi_i}{\partial \xi}\right] |2h| \, d\xi = \frac{1}{2h}\int_0^1 \frac{\partial \psi_j}{\partial \xi}\frac{\partial \psi_i}{\partial \xi} \, d\xi$$

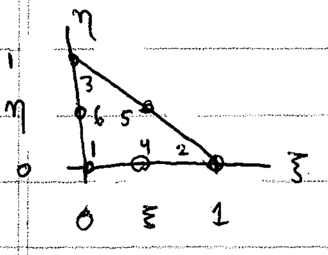
e.g. $$A_{13}^{loc} = \frac{1}{2h}\int_0^1 (4\xi - 1)(4\xi - 3)\, d\xi = \frac{1}{2h}\int_0^1 16\xi^2 - 16\xi + 3 \, d\xi = \frac{1}{6h}$$

result: $$A^{loc} = \frac{1}{h}\begin{pmatrix} 7/6 & -4/3 & 1/6 \\ -4/3 & 8/3 & -4/3 \\ 1/6 & -4/3 & 7/6 \end{pmatrix}$$



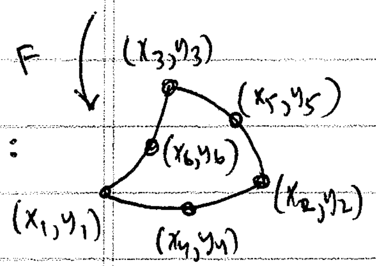
Example: 2d isoparametric elements

R:



$$\psi_1(\xi,\eta) = (1-\xi-\eta)(1-2\xi-2\eta)$$
$$\psi_2(\xi,\eta) = (\xi)(2\xi-1)$$
$$\psi_3(\xi,\eta) = (\eta)(2\eta-1)$$
$$\psi_4(\xi,\eta) = (2\xi)(2-2\xi-2\eta)$$
$$\psi_5(\xi,\eta) = 4\xi\eta$$
$$\psi_6(\xi,\eta) = (2\eta)(2-2\xi-2\eta)$$

T:



$$\begin{pmatrix} x \\ y \end{pmatrix} = F\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \sum_{k=1}^{6} \begin{pmatrix} x_k \\ y_k \end{pmatrix} \psi_k(\xi,\eta)$$

$$DF = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix} = \begin{pmatrix} \sum_k x_k \frac{\partial \psi_k}{\partial \xi}(\xi,\eta) & \sum_k x_k \frac{\partial \psi_k}{\partial \eta}(\xi,\eta) \\ \sum_k y_k \frac{\partial \psi_k}{\partial \xi}(\xi,\eta) & \sum_k y_k \frac{\partial \psi_k}{\partial \eta}(\xi,\eta) \end{pmatrix}$$

2×2 matrix depending on ξ and η

break into sum of scalar integrals

Note that $A^{loc}_{ij} = \iint_T \nabla_x \varphi_{\ell_i} \cdot \nabla_x \varphi_{\ell_j}\, dx\, dy = \underbrace{\iint_T (\partial_x \varphi_{\ell_i})(\partial_x \varphi_{\ell_j})\, dx\,dy}_{A^{loc,1}_{ij}} + \underbrace{\iint_T (\partial_y \varphi_{\ell_i})(\partial_y \varphi_{\ell_j})\, dx\,dy}_{A^{loc,2}_{ij}}$

$$A^{loc,k}_{ij} = \iint_R \left[\mathrm{row}_k(DF^{-T}) \nabla_\xi \psi_i\right]\left[\mathrm{row}_k(DF^{-T}) \nabla_\xi \psi_j\right] |\det DF|\, d\xi\, d\eta$$

This integral is most easily evaluated using Gaussian quadrature:

quadrature points

quadrature weights

$$A^{loc,k}_{ij} = \sum_{m=1}^{g} \left[\mathrm{row}_k\left(DF(\xi_m,\eta_m)^{-T}\right) \nabla_\xi \psi_i(\xi_m,\eta_m)\right] \cdot \left[\mathrm{row}_k\left(DF(\xi_m,\eta_m)^{-T}\right)\nabla_\xi \psi_j(\xi_m,\eta_m)\right] \left|\det DF(\xi_m,\eta_m)\right| W_m$$

$$= \left(E^{(k)T} E^{(k)}\right)_{ij}$$

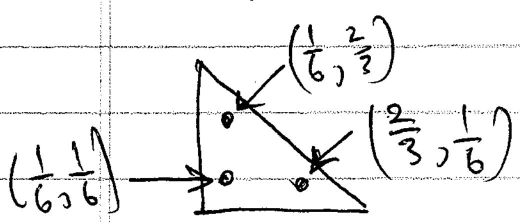
where the $g \times np$ matrix $E^{(k)}$ is $E^{(k)}_{mj} = \left[\mathrm{row}_k\left(DF(\xi_m,\eta_m)^{-T}\right)\nabla_\xi \psi_j(\xi_m,\eta_m)\right]\sqrt{|\det DF(\xi_m,\eta_m)|\, W_m}$

$O(g \cdot np)$ work to form $E^{(k)}$, $O(g\cdot(np)^2)$ to compute $A^{loc,k}$ from $E^{(k)}$ (at Level 3 Blas speed)

Numerical quadrature

to actually do the integrals over the reference triangle, we use Gaussian quadrature:

example 3 point G.Q. rule:

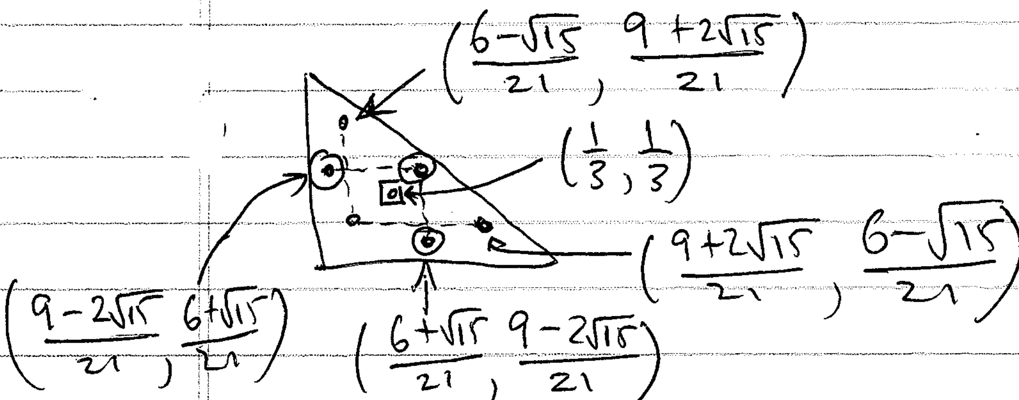


$A=\frac{1}{2}$ for ref. tri.

equal weights $w_i = \frac{A}{3} = \frac{1}{6}$

integrates polynomials of deg ≤ 2 exactly.

example from book: 7 pt G.Q. rule



$wt_{\boxed{\cdot}} = 9/180$

$wt_{\odot} = \frac{155+\sqrt{15}}{2400}$

$wt_{\circ} = \frac{155-\sqrt{15}}{2400}$

sum = $\frac{1}{2}$

Integrates polynomials of deg ≤ 5 exactly

in hw7 directory, I give you several G.Q. rules:

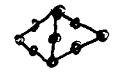
n	d
3	2
7	5
16	8
37	13
73	19

gauss02
05
08
13
19

these high order ones aren't so easy to find in the literature!

228B Lec 28

Last time: ① if u is piecewise smooth, then $u \in H^m(\Omega_h) \iff u \in C^{m-1}(\Omega_h)$

② nodal basis $\Rightarrow$ C^0 elements ( matching values at common nodes guarantees continuity across entire edge)

③ C^1 elements are tricky (have to avoid jump in normal derivative)

④ element by element assembly

today: ① reference element for computing local stiffness matrix

② numerical quadrature

③ interpolating f (use of a mass matrix)

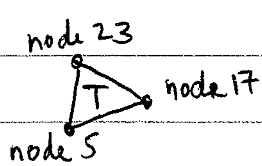
④ non-zero dirichlet data

<u>Recap</u>: element by element assembly

global stiffness matrix: $A_{ij} = a(\varphi_i, \varphi_j) = \sum_{T \in \mathcal{T}} \iint_T \nabla\varphi_i \cdot \nabla\varphi_j \, dx\,dy$

or $A_{ij} = \sum_T A_{ij}^{(T)}$, $A_{ij}^{(T)} = \iint_T \nabla\varphi_i \cdot \nabla\varphi_j \, dx\,dy$

Note: $A_{ij}^{(T)}$ is zero unless nodes i & j both belong to triangle T



columns 5, 17, 23

$$A^{(T)} = \begin{pmatrix} \cdot & & \cdot\ \cdot \\ \\ \cdot & & \cdot\ \cdot \\ \cdot & & \cdot\ \cdot \end{pmatrix} \begin{array}{l} \leftarrow \text{row } 5 \\ \\ \leftarrow \text{row } 17 \\ \leftarrow \text{row } 23 \end{array}$$

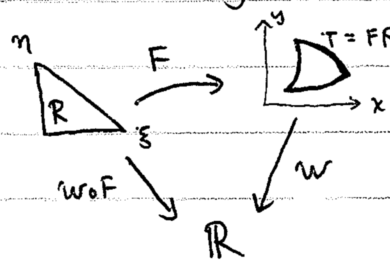
all other entries of $A^{(T)}$ are zero

To represent $A^{(T)}$, we just need the node numbers l_1, l_2, l_3 (in general $l_1, \dots, l_{np}$) and the 3×3 local stiffness matrix A_{ij}^{loc}:

$$A_{ij}^{loc} = A_{l_i l_j}^{(T)} = \iint_T \nabla\varphi_{l_i} \cdot \nabla\varphi_{l_j} \, dx\,dy.$$

global assembly: $A = \text{sparse}(n,n)$
foreach $T \in \mathcal{T}$
for $i = 1 \dots np$
for $j = 1 \dots np$
$A_{l_i l_j} = A_{l_i l_j} + A_{ij}^{loc}$

In practice, it's often convenient to compute the local stiffness matrix using a reference element



composition $w \circ F(\xi,\eta) = w(F(\xi,\eta))$

$$\iint_T w \, dx\, dy = \iint_R \overbrace{w \circ F}^{} \underbrace{|\det DF|}_{\text{Jacobian}} \, d\xi\, d\eta$$

change of variables formula

$$A^{loc}_{ij} = a_T(\varphi_{\ell_i}, \varphi_{\ell_j}) = \iint_T \underbrace{\nabla_x \varphi_{\ell_i} \cdot \nabla_x \varphi_{\ell_j}}_{w \text{ in the formula above}} \, dx\, dy$$

Let $\psi_i(\xi,\eta) = \varphi_{\ell_i}(F(\xi,\eta))$ $\qquad \xi \leftrightarrow \xi_1 \quad \eta \leftrightarrow \xi_2 \qquad x \leftrightarrow x_1 \quad y \leftrightarrow x_2$

3 flavors of chain rule: $(\psi = \varphi \circ F)$

$$\left\{ \begin{array}{l} \dfrac{\partial \psi}{\partial \xi_j} = \sum\limits_k \dfrac{\partial \varphi}{\partial x_k} \dfrac{\partial x_k}{\partial \xi_j} \\ D\psi = D\varphi \circ DF \\ (\nabla_\xi \psi)^T = (\nabla_x \varphi)^T \circ DF \\ \_\_\_ = \_\_\_ \cdot \boxed{} \end{array} \right.$$

$\vec{x} = F(\vec{\xi})$

$$DF = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix}$$

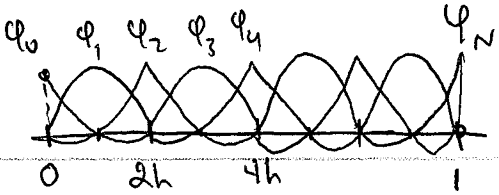
The last one gives $(\nabla_x \varphi)^T = (\nabla_\xi \psi)^T \circ (DF)^{-1}$ or $\boxed{\nabla_x \varphi = (DF)^{-T} \nabla_\xi \psi}$ ☆

$$\therefore \ A^{loc}_{ij} = \iint_R \left[(DF)^{-T} \nabla_\xi \psi_j \right] \cdot \left[(DF)^{-T} \nabla_\xi \psi_i \right] |\det DF| \, d\xi\, d\eta$$

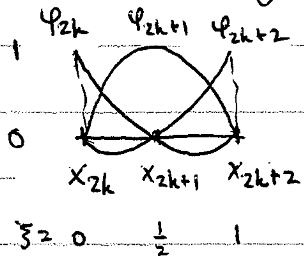
dot product

The integrand is now a function of $\vec{\xi}$ only, and the basis functions $\psi_1, \dots, \psi_{np}$ do not change from triangle to triangle... only the mapping F from R to T changes.

Example: 1d quadratic elements



$I = [x_{2k}, x_{2k+2}]$



$x = F(\xi) = x_{2k} + 2h\xi$

$DF(\xi) = 2h$

node numbers of this element

$\boxed{l_1 = 2k,\ l_2 = 2k+1,\ l_3 = 2k+2}$

basis functions in reference frame

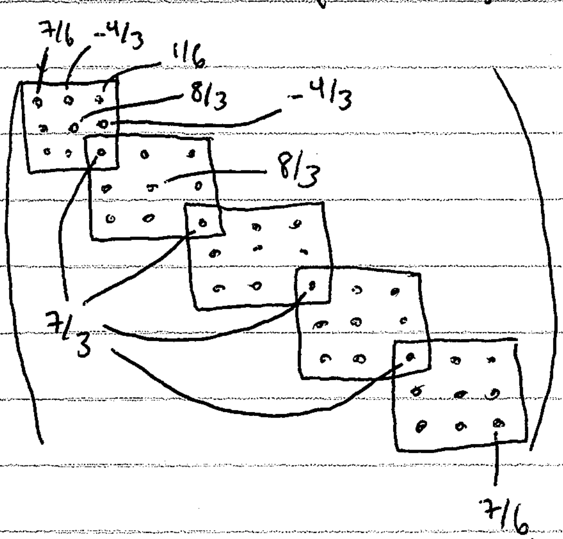
$$\left\{\begin{aligned}
\psi_1(\xi) &= \frac{(\frac{1}{2}-\xi)(1-\xi)}{(\frac{1}{2})(1)} = \frac{\frac{1}{2} - \frac{3}{2}\xi + \xi^2}{\frac{1}{2}} = 2\xi^2 - 3\xi + 1 \\
\psi_2(\xi) &= \frac{(\xi)(1-\xi)}{(\frac{1}{2})(\frac{1}{2})} = -4\xi^2 + 4\xi \\
\psi_3(\xi) &= \frac{(\xi)(\xi-\frac{1}{2})}{(1)(\frac{1}{2})} = 2\xi^2 - \xi
\end{aligned}\right.$$

$$A^{loc}_{ij} = \int_0^1 \left[(2h)^{-1}\frac{\partial\psi_j}{\partial\xi}\right]\left[(2h)^{-1}\frac{\partial\psi_i}{\partial\xi}\right]|2h|\, d\xi = \frac{1}{2h}\int_0^1 \frac{\partial\psi_j}{\partial\xi}\frac{\partial\psi_i}{\partial\xi}\, d\xi$$

e.g. $$A^{loc}_{13} = \frac{1}{2h}\int_0^1 (4\xi-1)(4\xi-3)\, d\xi = \frac{1}{2h}\int_0^1 16\xi^2 - 16\xi + 3\, d\xi = \frac{1}{6h}$$

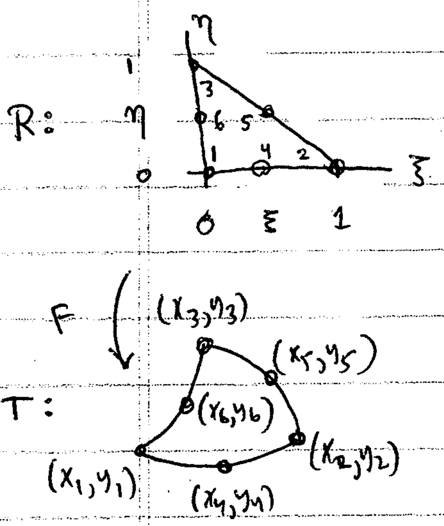
result: $$A^{loc} = \frac{1}{h}\begin{pmatrix} 7/6 & -4/3 & 1/6 \\ -4/3 & 8/3 & -4/3 \\ 1/6 & -4/3 & 7/6 \end{pmatrix}$$

$A =$



each box holds $A^{(T)}$ from page 1 as T (or better, I) ranges over all the elements

Example: 2d isoparametric elements



$$\psi_1(\xi,\eta) = (1-\xi-\eta)(1-2\xi-2\eta)$$
$$\psi_2(\xi,\eta) = (\xi)(2\xi-1)$$
$$\psi_3(\xi,\eta) = (\eta)(2\eta-1)$$
$$\psi_4(\xi,\eta) = (2\xi)(2-2\xi-2\eta)$$
$$\psi_5(\xi,\eta) = 4\xi\eta$$
$$\psi_6(\xi,\eta) = (2\eta)(2-2\xi-2\eta)$$

$$\begin{pmatrix} x \\ y \end{pmatrix} = F\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \sum_{k=1}^{6} \begin{pmatrix} x_k \\ y_k \end{pmatrix} \psi_k(\xi,\eta)$$

$$DF = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix} = \begin{pmatrix} \sum_k x_k \frac{\partial \psi_k}{\partial \xi}(\xi,\eta) & \sum_k x_k \frac{\partial \psi_k}{\partial \eta}(\xi,\eta) \\ \sum_k y_k \frac{\partial \psi_k}{\partial \xi}(\xi,\eta) & \sum_k y_k \frac{\partial \psi_k}{\partial \eta}(\xi,\eta) \end{pmatrix}$$

2×2 matrix depending on ξ and η

in the homework, $(x_4,y_4) = \frac{1}{2}\left[(x_1,y_1)+(x_2,y_2)\right]$ and $(x_6,y_6) = \frac{1}{2}\left[(x_1,y_1)+(x_3,y_3)\right]$

so
$$\begin{pmatrix} x \\ y \end{pmatrix} = F\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} x_1 + (x_2-x_1)\xi + (x_3-x_1)\eta + \left(x_5 - \frac{x_2+x_3}{2}\right)4\xi\eta \\ y_1 + (y_2-y_1)\xi + (y_3-y_1)\eta + \left(y_5 - \frac{y_2+y_3}{2}\right)4\xi\eta \end{pmatrix}$$

because of the curved boundary, the Jacobian depends on ξ and η:

$$DF = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix} = \begin{pmatrix} (x_2-x_1) + \left(x_5 - \frac{x_2+x_3}{2}\right)4\eta & (x_3-x_1)+\left(x_5-\frac{x_2+x_3}{2}\right)4\xi \\ (y_2-y_1) + \left(y_5 - \frac{y_2+y_3}{2}\right)4\eta & (y_3-y_1)+\left(y_5-\frac{y_2+y_3}{2}\right)4\xi \end{pmatrix}$$

$$DF = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad DF^{-T} = \frac{1}{ad-bc}\begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$
← also a function of ξ and η

Now that we have DF^{-T} and $\psi_1, \psi_2, \dots, \psi_6$ we can compute

$$A^{loc}_{ij} = \iint_R \left[(DF)^{-T} \nabla_\xi \psi_i\right] \cdot \left[(DF)^{-T} \nabla_\xi \psi_j\right] |\det DF| \, d\xi \, d\eta$$

There are ~~36~~ 21 integrals to be performed here. It turns out a lot of the work can be re-used:

step 1: get rid of the dot product (do the x and y derivatives separately)

$$A^{loc}_{ij} = A^{loc,1}_{ij} + A^{loc,2}_{ij} \; ; \quad A^{loc,k}_{ij} = \iint_T (\partial_k \varphi_{\ell_i})(\partial_k \varphi_{\ell_j}) \, dx\, dy$$

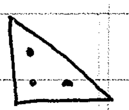
remember $\nabla_x \varphi = (DF)^{-T} \nabla_\xi \psi$

so $k=1 \leftrightarrow \partial_x \varphi$

$k=2 \leftrightarrow \partial_y \varphi$

$$A^{loc,k}_{ij} = \iint_R \underbrace{\left[\text{row}_k(DF^{-T}) \nabla_\xi \psi_i\right]\left[\text{row}_k(DF^{-T}) \nabla_\xi \psi_j\right] |\det DF|}_{w(\xi,\eta)} \, d\xi \, d\eta$$

step 2: use Gaussian quadrature to do the integrals over R



$$\iint_R w(\xi,\eta) \, d\xi \, d\eta \approx \sum_{m=1}^{g} w(\xi_m, \eta_m) \, W_m$$

quadrature points

" weights

In our case we find that $A^{loc,k}_{ij} = \sum_{m=1}^{g} E^{(k)}_{mi} E^{(k)}_{mj}$

where

$$E^{(k)}_{mj} = \left[\text{row}_k\left(DF(\xi_m,\eta_m)^{-T}\right) \underbrace{\nabla_\xi \psi_j(\xi_m,\eta_m)}\right] \sqrt{|\det DF(\xi_m,\eta_m)| \, W_m}$$

$1 \le m \le g$

$1 \le j \le np = 6$

these numbers can be computed once and for all (derivatives of the basis functions at the gauss points)

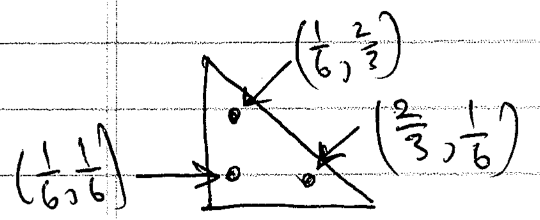
$O(g \cdot np)$ work to form $E^{(k)}$

$O(g \cdot (np)^2)$ work to compute $A^{loc,k} = \underbrace{E^{(k)T} E^{(k)}}_{\text{matrix multiplication}}$ (at Level 3 BLAS speed)

Numerical quadrature

to actually do the integrals over the reference triangle, we use Gaussian quadrature:

example 3 point G.Q. rule:

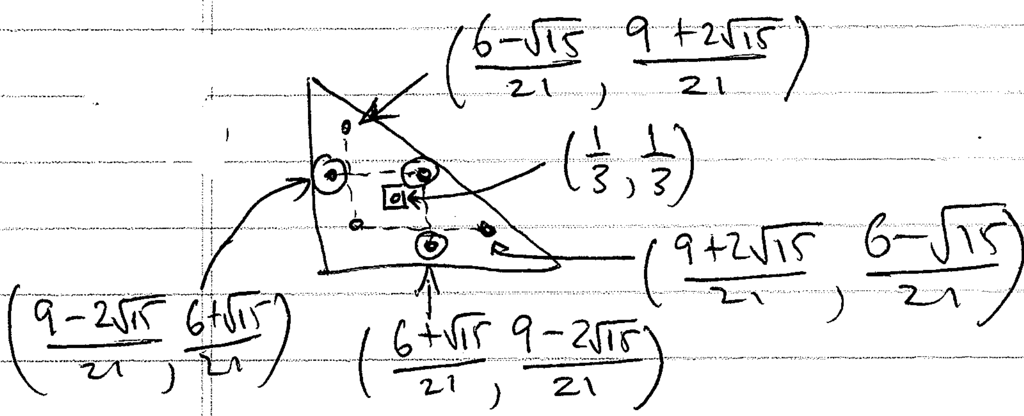


$A = \frac{1}{2}$ for ref. tri.

equal weights $w_i = \frac{A}{3} = \frac{1}{6}$

integrates polynomials of deg ≤ 2 exactly.

example from book: 7 pt G.Q. rule



$wt_{\boxed{\cdot}} = 9/180$

$wt_{\odot} = \frac{155+\sqrt{15}}{2400}$

$wt_{\circ} = \frac{155-\sqrt{15}}{2400}$

sum $= \frac{1}{2}$

Integrates polynomials of deg ≤ 5 exactly

in hw7 directory, I give you several G.Q. rules:

g	d	
3	2	gauss02
7	5	05
16	8	08
37	13	13
73	19	19

these high order ones aren't so easy to find in the literature!

e.g. this scheme integrates all polynomials of degree ≤ 19 exactly

<u>Interpolating f</u>: We need to solve $a(u_h, v_h) = \langle \ell, v_h \rangle \ \forall v_h \in V$

where $a(u_h, v_h) = \iint_{\Omega_h} \nabla u_h \cdot \nabla v_h \, dx dy$

$\langle \ell, v_h \rangle = \iint_{\Omega_h} f v_h \, dx dy$

In practice, we usually replace f by $\tilde{f}(x,y) = \sum_k f(x_k, y_k) \varphi_k(x,y)$

i.e. we interpolate f from its values at the nodes using the same basis functions we use to represent the solution.

<u>error</u>: $a(u_h - \tilde{u}_h, v_h) = \langle \ell - \tilde{\ell}, v_h \rangle = \iint_{\Omega_h} (f - \tilde{f}) v_h \, dx dy$

choosing $v_h = u_h - \tilde{u}_h$ we get $\boxed{\|u_h - \tilde{u}_h\|_1 \le \frac{1}{\alpha} \|f - \tilde{f}\|_0}$ (as in Lec 25)

linear elements: if $f \in H^2(\Omega_h)$ then $\|f - \tilde{f}\| \le C \|f\|_2 h^2$

quadratic " : if $f \in H^3(\Omega_h)$ then $\|f - \tilde{f}\| \le C \|f\|_3 h^3$

so the error committed by interpolating f is one order higher in h than the error estimates we'll derive for u_h next week.

$\boxed{\therefore \text{ interpolating } f \text{ does not significantly affect convergence of the FE method.}}$

<u>mass matrix</u>: $u_h(x,y) = \sum_k u_k \varphi_k(x,y)$, $\tilde{f}(x,y) = \sum_k f_k \varphi_k(x,y)$

$$a(u_h, \varphi_i) = \langle \tilde{\ell}, \varphi_i \rangle \qquad 1 \le i \le n$$

$$\sum_k a(\varphi_k, \varphi_i) u_k = \sum_k \left(\iint_{\Omega_h} \varphi_k \varphi_i \, dx dy \right) f_k$$

$$Au = Mf \ , \quad A_{ij} = a(\varphi_i, \varphi_j) \ , \ M_{ij} = \iint \varphi_i \varphi_j \, dx dy$$

M should be assembled element by element as well, but the formulas are simpler since there are no derivatives involved.

$$M = \sum_T M_{ij}^{(T)}, \qquad M_{\ell_i \ell_j}^{(T)} = M_{ij}^{loc} = \iint_T \varphi_{\ell_i} \varphi_{\ell_j}\, dx\,dy$$

$$= \iint_R \psi_i \psi_j \, |\det DF| \, d\xi\, d\eta$$

basis fcns in ref. triangle

$$= \sum_{m=1}^{g} E_{mi} E_{mj} = (E^T E)_{ij}$$

$$E_{mj} = \psi_j(\xi_m, \eta_m) \sqrt{|\det DF(\xi_m, \eta_m)|\, w_m}$$

ξ_m, η_m, w_m quadrature points and weights

Nonzero b/c's: want to solve $-\Delta u = f$ in Ω

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

idea: pick any function u_0 that equals g on $\partial\Omega$. $(u_0 \in H^1(\Omega))$

decompose $u = u_0 + u_1$, $u_1 \in H_0^1(\Omega)$

Then $-\Delta u = -\Delta u_0 - \Delta u_1 = f$

so u_1 should satisfy

$$-\Delta u_1 = f + \Delta u_0 \quad \text{in } \Omega$$
$$u_1 = 0 \quad \text{on } \partial\Omega$$

hit with test fcn, integrate by parts:

$$a(u_1, v) = \iint_\Omega f v \, dx\,dy - a(u_0, v) \qquad \forall v \in H_0^1(\Omega)$$

the RHS is a bounded linear functional on $H_0^1(\Omega)$, so Lax-Milgram gives a unique solution $u_1 \in H_0^1(\Omega)$ such that $u = u_0 + u_1$ solves the original problem.

The finite element approach is identical.

Let $S^h \subset H^1(\Omega)$ contain all nodal basis functions on the mesh (even those corresponding to nodes on the boundary)

and let $S_0^h = H_0^1(\Omega) \cap S_h$ be the linear span of the basis functions corresponding to interior nodes.

We use the basis functions on the boundary to represent $u_{0,h}$

$$u_{0,h}(x,y) = \sum_{k \in K_{bdry}} g_k \, \varphi_k(x,y)$$

$K_{bdry} = \{k : (x_k, y_k) \in \partial\Omega_h\}$ = set of boundary node numbers

$g_k = g(x_k, y_k)$ = prescribed values on boundary

We decompose $u_h = u_{0,h} + u_{1,h}$ and solve

$$a(u_{1,h}, v_h) = \langle \tilde{\ell}, v_h \rangle - a(u_{0,h}, v_h) \qquad \forall v_h \in S_0^h$$

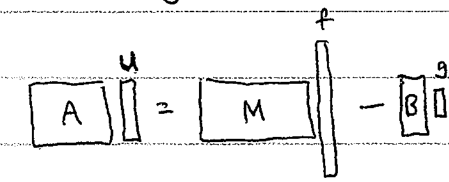
Note that $u_{1,h}$ and v_h only involve interior basis functions while $\tilde{f}$ (the interpolated version of f) and $u_{0,h}$ also involve boundary nodes.

$$u_{1,h} = \sum_{k\in K_{int}} u_k \varphi_k , \quad \tilde{f} = \sum_{k\in K_{all}} f_k \varphi_k , \quad u_{0,h} = \sum_{k\in K_{bdry}} g_k \varphi_k$$

$K_{all} = K_{bdry} \cup K_{int}$ ← interior nodes

$$\sum_{k\in K_{int}} a(\varphi_i, \varphi_k) u_k = \sum_{k\in K_{all}} \left(\iint \varphi_i \varphi_k \, dx\, dy \right) f_k - \sum_{k \in K_{bdry}} a(\varphi_i, \varphi_k) g_k$$

$$Au = Mf - Bg$$



A = usual stiffness matrix $(N_{int} \times N_{int})$

M = mass matrix $(N_{int} \times N)$ $\qquad N = N_{int} + N_{bdry}$

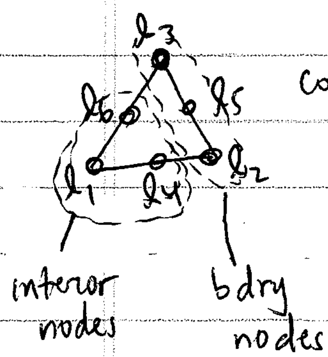
B = boundary stiffness matrix $(N_{int} \times N_{bdry})$

To implement this, I would have each node carry: $x, y, f, u, flag, eqn$

g on bdry (u); flag: 1 bdry, 0 interior

eqn: interior or bdry equation number (mapping to rows and columns of B)

A and B can be updated simultaneously from the local stiffness matrix:



Compute A^{loc} as before then add to A and B like this, for example →

```
for i=1..6
  if flag(ℓi)==0          // bdry nodes irrelevant to rows
    for j=1..6
      if flag(ℓj)==0
        A(eqn(ℓi),eqn(ℓj)) += A_ij^loc
      else
        B(eqn(ℓi),eqn(ℓj)) += A_ij^loc
```

<u>Last time</u>: (i) reference element for computing local stiffness matrix

(a) change of variables formula for the integral

(b) chain rule to convert derivatives from $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$ into $\frac{\partial}{\partial \xi}, \frac{\partial}{\partial \eta}$

(c) numerical quadrature is used to do the integrals
(most of the work boils down to matrix-matrix multiplication)

<u>Today</u>: (i) finish discussing implementation issues

(a) interior and boundary nodes

(b) interpolating f

(c) non-zero drichlet data

(d) computing errors in the homework

---

comment on Dirichlet conditions:

I've been writing $A_{ij} = \iint_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx\,dy$

but actually the stiffness matrix only involves <u>interior nodes</u> i & j.

<u>2 options</u>: (1) number the nodes so that interior nodes come first

or (2) maintain a separate numbering of the interior nodes
(sweep through the mesh once to set up this numbering)

```
k=1 , b=1
for i=1..N
  if flag(i)==0
     eqn(i) = k++   <- interior numbering
  else
     eqn(i) = b++   <- boundary numbering
```

to avoid excessive notation, let's pretend that the interior nodes come first in the list, i.e.

$\text{eqn}(i) = i \qquad 1 \le i \le N_{int}$

$\text{eqn}(i) = i - N_{int} \qquad N_{int}+1 \le i \le N$

__Interpolating f__: We need to solve $a(u_h, v_h) = \langle \ell, v_h \rangle \quad \forall v_h \in V$

where $a(u_h, v_h) = \iint_{\Omega_h} \nabla u_h \cdot \nabla v_h \, dx\,dy$

$\langle \ell, v_h \rangle = \iint_{\Omega_h} f v_h \, dx\,dy$

In practice, we usually replace f by $\tilde{f}(x,y) = \sum_k f(x_k, y_k)\varphi_k(x,y)$

i.e. we interpolate f from its values at the nodes using the same basis functions we use to represent the solution.

__error__: $a(u_h - \tilde{u}_h, v_h) = \langle \ell - \tilde{\ell}, v_h \rangle = \iint_{\Omega_h} (f - \tilde{f}) v_h \, dx\,dy$

choosing $v_h = u_h - \tilde{u}_h$ we get $\boxed{\|u_h - \tilde{u}_h\|_1 \le \frac{1}{\alpha}\|f - \tilde{f}\|_0}$ (as in Lec 25)

linear elements: if $f \in H^2(\Omega_h)$ then $\|f - \tilde{f}\|_0 \le C|f|_2 h^2$

quadratic " : if $f \in H^3(\Omega_h)$ then $\|f - \tilde{f}\|_0 \le C|f|_3 h^3$

so the error committed by interpolating f is one order higher in h than the error estimates we'll derive for u_h shortly

$\boxed{\therefore \text{ interpolating } f \text{ does not significantly affect convergence of the FE method.}}$

__mass matrix__: $u_h(x,y) = \sum_{k=1}^{N_{int}} u_k \varphi_k(x,y)$, $\quad \tilde{f}(x,y) = \sum_{k=1}^{N} f_k \varphi_k(x,y)$

$$a(u_h, \varphi_i) = \langle \tilde{\ell}, \varphi_i \rangle \qquad 1 \le i \le N_{int}$$

$$\sum_k a(\varphi_k, \varphi_i) u_k = \sum_k \left(\iint_{\Omega_h} \varphi_k \varphi_i \, dx\,dy \right) f_k$$

$$Au = Mf, \qquad A_{ij} = a(\varphi_i, \varphi_j), \quad M_{ij} = \iint \varphi_i \varphi_j \, dx\,dy$$

$1 \le i, j \le N_{int}$ $\qquad$ $1 \le i \le N_{int}, \ 1 \le j \le N$

M should be assembled element by element as well, but the formulas are simpler since there are no derivatives involved.

$$M = \sum_T M_{ij}^{(T)}, \qquad M_{\ell_i \ell_j}^{(T)} = M_{ij}^{loc} = \iint_T \varphi_{\ell_i}\varphi_{\ell_j}\,dx\,dy$$

$$= \iint_R \psi_i \psi_j \,|\det DF|\, d\xi\, d\eta$$

basis fcns in ref. triangle

$$= \sum_{m=1}^{9} E_{mi} E_{mj} = (E^T E)_{ij}$$

$$E_{mj} = \psi_j(\xi_m, \eta_m)\sqrt{|\det DF(\xi_m,\eta_m)|\, w_m}$$

ξ_m, η_m, w_m quadrature points and weights

Nonzero b/c's: want to solve $-\Delta u = f$ in Ω

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

idea: pick <u>any</u> function u_0 that equals g on $\partial\Omega$. $(u_0 \in H^1(\Omega))$

decompose $u = u_0 + u_1, \quad u_1 \in H^1_0(\Omega)$

Then $-\Delta u = -\Delta u_0 - \Delta u_1 = f$

so u_1 should satisfy

$$-\Delta u_1 = f + \Delta u_0 \quad \text{in } \Omega$$

$$u_1 = 0 \quad \text{on } \partial\Omega$$

hit with test fcn, integrate by parts:

$$a(u_1, v) = \iint_\Omega f v\, dx\, dy - a(u_0, v) \qquad \forall v \in H^1_0(\Omega)$$

the RHS is a bounded linear functional on $H_0^1(\Omega)$, so Lax-Milgram gives a unique solution $u_1 \in H_0^1(\Omega)$ such that $u = u_0 + u_1$ solves the original problem.

The finite element approach is identical.

Let $S^h \subset H^1(\Omega)$ contain all nodal basis functions on the mesh (even those corresponding to nodes on the boundary)

and let $S_0^h \subset H_0^1(\Omega) \cap S_h$ be the linear span of the basis functions corresponding to interior nodes.

We use the basis functions on the boundary to represent $u_{0,h}$

$$u_{0,h}(x,y) = \sum_{k \in K_{bdry}} g_k \varphi_k(x,y)$$

$K_{bdry} = \{k : (x_k, y_k) \in \partial\Omega_h\}$ = set of boundary node numbers

$g_k = g(x_k, y_k)$ = prescribed values on boundary

We decompose $u_h = u_{0,h} + u_{1,h}$ and solve

$$a(u_{1,h}, v_h) = \langle \tilde{\ell}, v_h \rangle - a(u_{0,h}, v_h) \qquad \forall v_h \in S_0^h$$

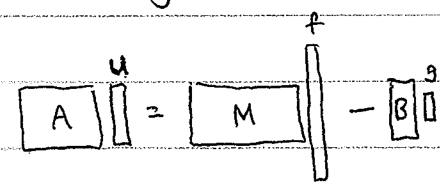
Note that $u_{1,h}$ and v_h only involve interior basis functions while $\tilde{f}$ (the interpolated version of f) and $u_{0,h}$ also involve boundary nodes.

$$u_{1,h} = \sum_{k \in K_{int}} u_k \varphi_k \,, \quad \tilde{f} = \sum_{k \in K_{all}} f_k \varphi_k \,, \quad u_{0,h} = \sum_{k \in K_{bdry}} g_k \varphi_k$$

$K_{all} = K_{bdry} \cup K_{int} \leftarrow$ interior nodes

$$\sum_{k \in K_{int}} a(\varphi_i, \varphi_k) u_k = \sum_{k \in K_{all}} \left(\iint \varphi_i \varphi_k \, dx\, dy \right) f_k - \sum_{k \in K_{bdry}} a(\varphi_i, \varphi_k) g_k$$

$$Au = Mf - Bg$$



A = usual stiffness matrix $(N_{int} \times N_{int})$

M = mass matrix $(N_{int} \times N)$ $\qquad N = N_{int} + N_{bdry}$

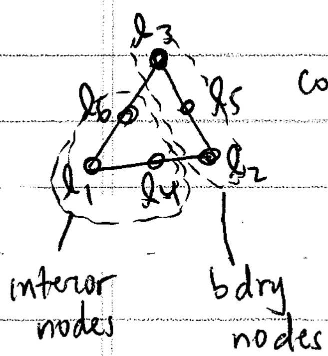
B = boundary stiffness matrix $(N_{int} \times N_{bdry})$

To implement this, I would have each node carry: $x, y, f, u, \text{flag}, \text{eqn}$

g on bdry (for u); flag: $\begin{cases} 1 & \text{bdry} \\ 0 & \text{interior} \end{cases}$

eqn: interior or bdry equation number (mapping to rows and columns of B)

A and B can be updated simultaneously from the local stiffness matrix:



Compute A^{loc} as before then add to A and B like this, for example

```
for i=1..6
  if flag(l_i)==0          // bdry nodes irrelevant to rows
    for j=1..6
      if flag(l_j)==0
        A(eqn(l_i),eqn(l_j)) += A^loc_ij
      else
        B(eqn(l_i),eqn(l_j)) += A^loc_ij
      M(eqn(l_i), l_j) += M^loc_ij
```

computing errors in the homework. Once you get the solution, you need to compute

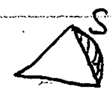
$$\|u_h - u\|_0 = \sqrt{\iint_\Omega (u_h(x,y) - u(x,y))^2 \, dx\,dy}$$

and

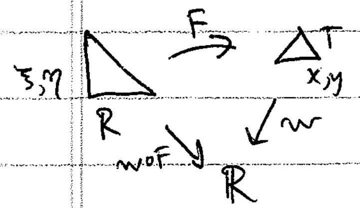
$$|u_h - u|_1 = \sqrt{\iint_\Omega \left|\nabla(u_h(x,y) - u(x,y))\right|^2 dx\,dy}$$

These integrals may be broken up as $\underline{\iint_\Omega \cdots dx dy = \sum_T \iint_T \cdots dxdy + \sum_S \iint_S \cdots dxdy}$

↑ slivers outside mesh.

over each sliver , $u_h(x,y)$ is zero and you should integrate $u(x,y)^2$ or $|\nabla u(x,y)|^2$ by hand.

Over each triangle, we use the reference element to do the integration



$$w(x,y) = u_h(x,y) - u(x,y)$$

$$\iint_T w^2 \, dx\,dy = \iint_R (w \circ F)^2 |\det DF| \, d\xi\, d\eta$$

$$= \sum_m w(F(\xi_m,\eta_m))^2 \, |\det DF(\xi_m,\eta_m)| \, W_m$$

(so you just have to evaluate the errors w at the images of the gauss points in T)

depends on ξ,η only in the isoparametric case.

quadrature weights

$$\boxed{\text{note that } w(F(\xi_m,\eta_m)) = \underbrace{\sum_{i=1}^{6} u_{l_i} \varphi_i(\xi_m,\eta_m)}_{u_h(F(\xi_m,\eta_m))} - u(F(\xi_m,\eta_m))}$$

similarly

$$\iint_T |\nabla_x w|^2 \, dx\,dy = \iint_R |\nabla_x w \circ F|^2 |\det DF| \, d\xi\, d\eta = \sum_m |\nabla_x w(F(\xi_m,\eta_m))|^2 \, |\det DF(\xi_m,\eta_m)| \, W_m$$

where

$$\nabla_x w(F(\xi_m,\eta_m)) = \sum_{i=1}^{6} u_{l_i} \underbrace{DF(\xi_m,\eta_m)^{-T} \nabla_\xi \varphi_i(\xi_m,\eta_m)}_{\nabla_x \varphi_{l_i}(F(\xi_m,\eta_m))} - \underbrace{\nabla_x u(F(\xi_m,\eta_m))}_{\binom{u_x}{u_y} \text{ evaluated at } F(\xi_m,\eta_m)}$$

Last time: ① two ways of setting up the data structure for solving the Dirichlet problem

ⓐ eliminate the boundary nodes to get an $N_{int} \times N_{int}$ system

or ⓑ zero out the rows and columns of A corresponding to boundary nodes and put 1's on the diagonal there

② nonzero b/c's: decompose $u_h = \underbrace{\sum_{k \in K_{int}} u_k \varphi_k}_{u_h^{(1)}} + \underbrace{\sum_{k \in K_{bdry}} u_k \varphi_k}_{u_h^{(0)}}$

solve

$$a(u_h^{(1)}, v_h) = \langle \ell, v_h \rangle - a(u_h^{(0)}, v_h) \quad \forall v_h \in S_0^h \leftarrow \text{test functions still zero on bdry}$$

or $$Au^{(1)} = Mf - Bu^{(0)}$$

in the ⓑ strategy above, you've simply corrected the error you made by zeroing out the boundary node columns:

$$\begin{pmatrix} * & * & & * \\ & * & & \\ 0\,0 & 1 & 0\,0\,0\,0 & \\ & * & & \\ * & * & & * \\ & * & & \end{pmatrix} \begin{pmatrix} u \end{pmatrix} = Au = Mf = \begin{pmatrix} * & * & * \\ & * & \\ 0\,0 & 0 & 0\,0\,0 \\ & * & \\ * & * & * \end{pmatrix} \begin{pmatrix} f \end{pmatrix}$$

want this column to be zero so stiffness matrix is symmetric. you know the value of u_j for this column since (x_j, y_j) is on the boundary. So just move it to the RHS

$$\begin{pmatrix} * & 0 & * \\ & 0 & \\ 0\,0 & 1 & 0\,0\,0\,0 \\ & 0 & \\ * & 0 & * \\ & 0 & \end{pmatrix} \begin{pmatrix} u \end{pmatrix} = Mf - \begin{pmatrix} * \\ * \\ 0 \\ * \\ * \\ * \end{pmatrix} u_j \quad \left(\text{also need to zero out that row of } M\right)$$

doing this for all the bdry nodes gives a system like $Au^{(1)} = Mf - Bu^{(0)}$ for the unknown interior values.

③ computing errors on the mesh.

<u>Remark</u>: method ⓐ is not difficult to implement either, and is particularly useful for problems in fluid mechanics where you use quadratic elements for velocity and linear elements for pressure.

Today: last steps of the error analysis.

we know $\|u_h - u\|_1 \le \frac{1}{\alpha} \inf_{v_h \in S_h} \|v_h - u\|_1 \le \frac{1}{\alpha} \|I_h u - u\|_1$

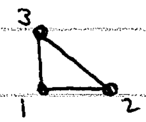
so our final task is to estimate the interpolation error $\|u - I_h u\|_1$

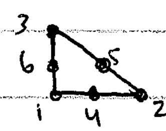
main steps:

(i) on the reference triangle, for all integers $t \ge 2$ $\exists c_1$ depending on t such that

$$\boxed{\|u - I u\|_{t,R} \le C_1(t)\, |u - Iu|_{t,R} = C_1(t)\, |u|_{t,R}}$$

where Iu is the polynomial of degree $t-1$ that agrees with u at the $\frac{t(t+1)}{2}$ uniformly spaced points on the triangle.

examples: $t=2$  $Iu(\xi,\eta) = u_2 \xi + u_3 \eta + u_1(1-\xi-\eta)$

$t=3$  $Iu(\xi,\eta) = \sum_{k=1}^{6} u_k \psi_k(\xi,\eta)$

$(u_k = u(\xi_k, \eta_k))$

this is a Poincaré-Friedrichs type of result. It says that the lower derivatives of any function that is zero at the interpolation points are controlled by the highest derivatives (those of order t):

$v = u - Iu$ is zero at the nodes. (so Iv is the zero polynomial)

$$\|v\|_{t,R}^2 = \underbrace{\sum_{|\alpha|<t} \iint_R |\partial^\alpha v(\xi,\eta)|^2 d\xi\, d\eta}_{\|v\|_{t-1,R}^2} + \underbrace{\sum_{|\alpha|=t} \iint_R |\partial^\alpha v|^2 d\xi\, d\eta}_{|v|_{t,R}^2}$$

the claim is that $\boxed{\|v\|_{t-1,R}^2 \le [C_1(t)^2 - 1]\, |v|_{t,R}^2 \quad \forall v \in H^t(R) \text{ s.t. } Iv = 0}$

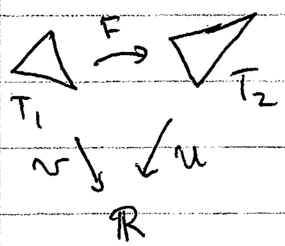
just as the constant functions prevent $\|u\|_0 \le C|u|_1 \ \forall u \in H^1(\Omega)$ without restricting u to be zero on the boundary, the polynomials prevent

$$\|u\|_{t-1} \le C|u|_t \quad \forall u \in H^t(\Omega).$$

Pinning down the values of u to be zero at the nodes removes these lower order polynomials from the space.

<u>step 2.</u> • Suppose T_1 and T_2 are any two triangles in the plane.

• Let F be an affine mapping of T_1 onto T_2.



(i.e. $F(\xi,\eta) = \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} + \begin{pmatrix} a_1 \\ b_1 \end{pmatrix}\xi + \begin{pmatrix} a_2 \\ b_2 \end{pmatrix}\eta$ for suitable constants $a_0, b_0, a_1, b_1, a_2, b_2$)

• Let $u : T_2 \to \mathbb{R}$ and $v : T_1 \to \mathbb{R}$ satisfy $v = u \circ F$

Then there are constants $C_2(t)$ for $t = 0, 1, 2, \ldots$ such that

$$\boxed{|v|_{t,T_1} \le C_2(t)\, \|DF\|^t\, |\det DF|^{-1/2}\, |u|_{t,T}}$$

for all $u \in H^t(T_2)$ and $v = u \circ F$ (which automatically belongs to $H^t(T_1)$)

To prove this, it's useful to think of higher derivatives as multilinear maps:

$u : \mathbb{R}^2 \to \mathbb{R}$ function

$Du(\vec{x}) : \mathbb{R}^2 \to \mathbb{R}$ linear operator

components, not powers

$$Du(\vec{x})\vec{w} = \left.\frac{d}{ds}\right|_{s=0} u(\vec{x} + s\vec{w}) = \frac{\partial u}{\partial x}(\vec{x})\, w^1 + \frac{\partial u}{\partial y}(\vec{x})\, w^2$$

$D^2u(\vec{x}) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ symmetric, bilinear operator

$$D^2u(\vec{x})(\vec{w}_1, \vec{w}_2) = \left.\frac{d}{ds_1}\right|_{s_1=0} \left.\frac{d}{ds_2}\right|_{s_2=0} u(\vec{x} + s_1\vec{w}_1 + s_2\vec{w}_2) = w_1^T \begin{pmatrix} \frac{\partial^2 u}{\partial x^2} & \frac{\partial^2 u}{\partial x \partial y} \\ \frac{\partial^2 u}{\partial x \partial y} & \frac{\partial^2 u}{\partial y^2} \end{pmatrix} w_2$$

or $D^2u(\vec{x})(\vec{w}_1,\vec{w}_2) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial^2 u}{\partial x_i \partial x_j}(\vec{x})\, w_1^i w_2^j$

in general: $D^m u(\vec{x}) : (\mathbb{R}^2)^m \to \mathbb{R}$ is a symmetric, multilinear operator

$$D^m u(\vec{x})(\vec{w}_1,\dots,\vec{w}_m) = \sum_{i_1=1}^{2}\cdots\sum_{i_m=1}^{2} \frac{\partial^m u}{\partial x_{i_1}\cdots\partial x_{i_m}}(\vec{x})\, w_1^{i_1}\cdots w_m^{i_m}$$

so a fancy way of writing $\frac{\partial^m u}{\partial x_{i_1}\cdots\partial x_{i_m}}(\vec{x})$ is $D^m u(\vec{x})(\vec{e}_{i_1},\vec{e}_{i_2},\dots,\vec{e}_{i_m})$

$\vec{e}_1 = \binom{1}{0}$, $\vec{e}_2 = \binom{0}{1}$

chain rule: $v(\vec{\xi}) = u(\overbrace{F(\vec{\xi})}^{\vec{x}})$

$$\frac{\partial v}{\partial \xi_j} = \frac{\partial u}{\partial x_k}\frac{\partial x_k}{\partial \xi_j} \quad \left(\text{summation implied}\right)$$

$$\frac{\partial^2 v}{\partial \xi_i \partial \xi_j} = \frac{\partial^2 u}{\partial x_l \partial x_k}\frac{\partial x_k}{\partial \xi_j}\frac{\partial x_l}{\partial \xi_i} + \frac{\partial u}{\partial x_k}\overbrace{\frac{\partial^2 x_k}{\partial \xi_i \partial \xi_j}}^{\text{zero since } F \text{ is affine}}$$

$$\frac{\partial^2 v}{\partial \xi_a \partial \xi_i \partial \xi_j} = \frac{\partial^2 u}{\partial x_b \partial x_l \partial x_k}\frac{\partial x_k}{\partial \xi_j}\frac{\partial x_l}{\partial \xi_i}\frac{\partial x_b}{\partial \xi_a} + 0$$

terms involving higher derivatives of F

more simply:

$$\boxed{D^m v(\vec{\xi})(\vec{w}_1,\dots,\vec{w}_m) = D^m u(F(\vec{\xi}))(DF\,\vec{w}_1,\dots,DF\,\vec{w}_m)}$$

the matrix DF is constant since F is affine

Next we compute

$$|v|^2_{t,T_1} = \iint_{T_1} \sum_{j=0}^{t} \left|\partial_\xi^j \partial_\eta^{t-j} v(\xi,\eta)\right|^2 d\xi\, d\eta$$

$$= \iint_{T_1} \sum_{j=0}^{t} \Big|D^t v(\xi,\eta)(\underbrace{\vec{e}_1,\dots,\vec{e}_1}_{j \text{ times}},\underbrace{\vec{e}_2,\dots,\vec{e}_2}_{t-j \text{ times}})\Big|^2 d\xi\, d\eta$$

$$= \iint_{T_2} \sum_{j=0}^{t} \left|D^t u(x,y)(Df\,\vec{e}_1,\dots,Df\,\vec{e}_2)\right|^2 \left|\det DF^{-1}\right| dx\,dy$$

$$\overset{*}{\le} \iint_{T_2} (t+1)\|DF\|^{2t} \|D^t u(x,y)\|^2 \left|\det DF^{-1}\right| dx\,dy \qquad \begin{pmatrix}\det(DF^{-1}) \\ = (\det DF)^{-1}\end{pmatrix}$$

$$\le (t+1)\|DF\|^{2t} |\det DF|^{-1} \iint_{T_2} \|D^t u(x,y)\|^2 dx\,dy$$

$$\overset{**}{\le} c_2(t)^2 \|DF\|^{2t} |\det DF|^{-1} |u|_{t,T_2} \qquad \boxed{c_2(t) = \sqrt{\frac{(t+1)!}{\left\lfloor\frac{t}{2}\right\rfloor!\left\lceil\frac{t}{2}\right\rceil!}}}$$

$*$ follows from the definition of norm of a multilinear functional:

$$B = D^t u(x,y), \quad \|B\| = \sup_{\substack{\|\vec{w}_1\|=1 \\ \vdots \\ \|\vec{w}_t\|=1}} |B(\vec{w}_1,\dots,\vec{w}_t)|$$

so $|B(\vec{w}_1,\dots,\vec{w}_t)| \le \|B\| \cdot \|\vec{w}_1\| \cdots \|\vec{w}_t\|$

in our case, $\vec{w}_j = DF\vec{e}_j$ so $\|\vec{w}_j\| \le \|DF\| \cdot \overbrace{\|\vec{e}_j\|}^{1}$

hence $|B(\underbrace{DF\vec{e}_1,\dots}_{j \text{ times}},\underbrace{\dots,DF\vec{e}_2}_{t-j \text{ times}})| \le \|B\| \cdot \|DF\|^t$

(**) follows from multilinearity of $B = D^t u(x,y)$:

$$\left|B(\underbrace{\vec{w}_1}_{w_1^1\vec{e}_1+w_1^2\vec{e}_2},\dots,\vec{w}_t)\right| = \left| w_1^1 B(\vec{e}_1,\vec{w}_2,\dots,\vec{w}_t) + w_1^2 B(\vec{e}_2,\vec{w}_2,\dots,\vec{w}_t)\right|$$

$$\overset{\text{cauchy-schwarz}}{\leq} \sqrt{(w_1^1)^2+(w_1^2)^2}\sqrt{\sum_{i=1}^{2}|B(\vec{e}_i,\vec{w}_1,\dots,\vec{w}_t)|^2}$$

By induction:

$$|B(\vec{w}_1,\dots,\vec{w}_t)| \leq \|\vec{w}_1\|\,\|\vec{w}_2\|\cdots\|\vec{w}_t\|\sqrt{\sum_{i_1=1}^{2}\cdots\sum_{i_t=1}^{2}|B(\vec{e}_{i_1},\dots,\vec{e}_{i_t})|^2}$$

which gives an upper bound on $\|B\|$

Finally, we use symmetry of the derivative:

$$\|D^t u(x,y)\|^2 \leq \sum_{i_1=1}^{2}\cdots\sum_{i_t=1}^{2}\left|\frac{\partial^t u}{\partial x_{i_1}\cdots\partial x_{i_t}}(x,y)\right|^2$$

$$= \sum_{j=0}^{t}\binom{t}{j}\left|\partial_x^j\partial_y^{t-j}u(x,y)\right|^2$$

$$\leq \left(\max_{0\leq j\leq t}\frac{t!}{j!(t-j)!}\right)\sum_{j=0}^{t}\left|\partial_x^j\partial_y^{t-j}u(x,y)\right|^2$$

$$\therefore \iint_{T_2}\|D^t u(x,y)\|^2\,dx\,dy \leq \frac{t!}{\lfloor\frac{t}{2}\rfloor!\lceil\frac{t}{2}\rceil!}\iint_{T_2}\sum_{j=0}^{t}\left|\partial_x^j\partial_y^{t-j}u(x,y)\right|^2 dx\,dy$$

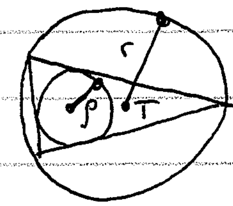
$$(t+1)\iint_{T_2}\|D^t u(x,y)\|^2\,dx\,dy \leq C_2(t)^2\,|u|_{t,T_2}$$

$$C_2(t)^2 = \frac{(t+1)!}{\lfloor\frac{t}{2}\rfloor!\lceil\frac{t}{2}\rceil!}$$

t	$C_2(t)$	
0	1	1.00
1	$\sqrt{2}$	1.41
2	$\sqrt{6}$	2.45
3	$\sqrt{12}$	3.46
4	$\sqrt{30}$	5.48
5	$\sqrt{60}$	7.75
6	$\sqrt{140}$	11.8

<u>step 3</u> derive a bound on $\|DF\|$ in terms of the mesh quality, K.

for a triangle T, define two radii:

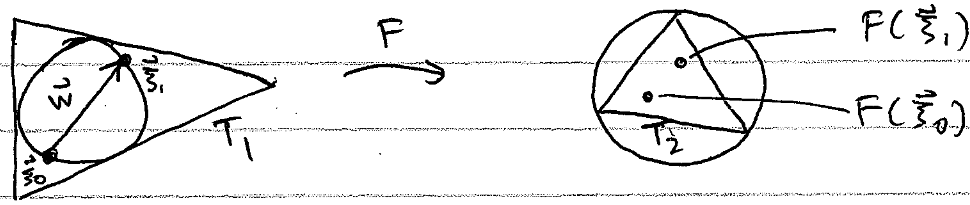


$2r = h$

Suppose $F : T_1 \to T_2$ is an affine map

pick any $\vec{w} \in \mathbb{R}^2$ s.t. $\|w\| = 2\rho_1$

Pick $\vec{\xi}_0, \vec{\xi}_1 \in T_1$ on inscribed circle so $\vec{w} = \vec{\xi}_1 - \vec{\xi}_0$



Then $\|DF\vec{w}\| \underset{\substack{\uparrow \\ F \text{ affine}}}{=} \|F(\vec{\xi}_1) - F(\vec{\xi}_0)\| \leq 2r_2 = \dfrac{2r_2}{2\rho_1}\|\vec{w}\|$

$$\therefore \|DF\| \leq \frac{2r_2}{2\rho_1}$$

reverse T_1, T_2:

$$\|DF^{-1}\| \leq \frac{2r_1}{2\rho_2}$$

condition number of DF: $\|DF\| \cdot \|DF^{-1}\| \leq \dfrac{r_2}{\rho_1} \cdot \dfrac{r_1}{\rho_2}$

<u>def</u>: The mesh quality parameter K is defined via

$$K = \max_{T \in \mathcal{T}_h} \frac{r_T}{\rho_T}. \quad \text{(a smaller } K \text{ is better)}$$

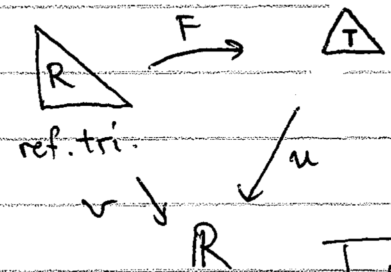
Step 4: For $t \geq 2$ $\exists\, c_3$ depending on t s.t.

$$\|u - I_h u\|_m \leq c_3(t)\, K^m h^{t-m} |u|_t \qquad \forall u \in H^t(\Omega_h)$$

$$0 \leq m \leq t$$

interpolation by piecewise polynomials of deg $t-1$.

proof: it suffices to establish the estimate on each triangle $T \in \mathcal{T}$



Pick j between 0 and m

From step 2 in the reverse direction, we have

$$\boxed{|u - I_h u|_{j,T} \leq c_2(j)\, \|DF^{-1}\|^j\, |\det DF|^{1/2}\, |v - Iv|_{j,R}}$$

From step 1, we have $\boxed{|v - Iv|_{j,R} \leq c_1(t)\, |v|_{t,R}}$

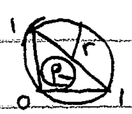
from step 2 in the forward direction, we learn:

$$\boxed{|v|_{t,R} \leq C_2(t)\, \|DF\|^t\, |\det DF|^{-1/2}\, |u|_{t,T}}$$

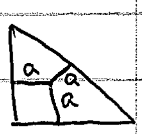
stringing these together, we obtain:

$$|u - I_h u|_{j,T} \leq c_1(t)\, c_2(j)\, c_2(t) \left(\|DF\| \cdot \|DF^{-1}\|\right)^j \|DF\|^{t-j} |u|_{t,T}$$

ref tri:



$r_R = \frac{\sqrt{2}}{2}$ $\quad \rho_R = \frac{2-\sqrt{2}}{2}$ $\quad \frac{r_R}{\rho_R} = \frac{\sqrt{2}}{2-\sqrt{2}} = \sqrt{2}+1$



step 3

$$\|DF\| \cdot \|DF^{-1}\| \leq \frac{r_R}{\rho_T} \cdot \frac{r_T}{\rho_R} \leq (\sqrt{2}+1) K$$

$a^2 = 2\left(\frac{1}{2} - a\right)^2$

$= \frac{1}{2} - 2a + 2a^2$

$a^2 - 2a + \frac{1}{2} = 0$

$a = \frac{2 \pm \sqrt{2}}{2}$

$$\|DF\| \leq \frac{r_T}{\rho_R} = \frac{2 r_T}{2-\sqrt{2}} \leq \frac{h}{2-\sqrt{2}} = \left(1 + \frac{\sqrt{2}}{2}\right) h$$

r_T: radius, h_T: diameter

$$|u - I_h u|_{j,T} \leq C_1(t)\, C_2(j)\, C_2(t)\, (1+\sqrt{2})^j \left(1+\tfrac{\sqrt{2}}{2}\right)^{t-j} K^j h^{t-j} |u|_{t,T}$$

sum from $j=0$ to m

$$\|u - I_h u\|_{m,T} \leq C_3(m,t)\, K^m h^{t-m} |u|_{t,T}$$

$$C_3(m,t) = \sum_{j=0}^{m} C_1(t)\, C_2(j)\, C_2(t)\, (1+\sqrt{2})^j \left(1+\tfrac{\sqrt{2}}{2}\right)^{t-j}$$

for simplicity, we set $m=t$ and drop the m dependence of $C_3(t)$

so now we have the error estimate

$$\|u_h - u\|_1 \overset{\text{Cea}}{\leq} \frac{1}{\alpha} \|u - I_h u\|_1 \leq \frac{1}{\alpha} C_3(t)\, K\, h^{t-1} |u|_t$$

which bounds the H^1 norm of the error in the F.E. solution in terms of the H^t seminorm of the solution.

Regularity theorem:

1. If Ω is convex, $\exists\, C_4(\Omega)$ s.t. $\|u\|_2 \leq C_4 \|f\|_0$
2. If $\partial\Omega$ is C^t with $t \geq 2$, $\exists\, C_4(\Omega, t)$ s.t. $\|u\|_t \leq C_4 \|f\|_{t-2}$
3. If Ω is a rectangle and $f=0$ in a neighborhood of the corners, then $\exists\, C_4(\Omega, \tilde{\Omega}, t)$ s.t. $\|u\|_{t,\Omega} \leq C_4 \|f\|_{t-2,\tilde{\Omega}}$ $\forall f \in H_0^{t-2}$)

$\tilde{\Omega} \subset\subset \Omega$ ↑ a compact set that stays away from the corners

We haven't quite done enough to analyze isoparametric elements (since we assumed F was affine) but at least we have proved the following:

Theorem: Suppose Ω is a convex polygon and we use linear or higher-order elements. Then

$$\|u_h - u\|_1 \le C_5(\Omega)\, \kappa\, h\, \|f\|_0 \qquad C_5(\Omega) = \frac{1}{\alpha} C_3(2) C_4(\Omega)$$

(corner singularities prevent improved estimates)

coercivity constant depends on Ω.

Theorem: Suppose Ω is a rectangle and $\tilde{\Omega} \subset\subset \Omega$ "compact subset of"

Then if we use triangular elements of degree p (linear: $p=1$, quadratic: $p=2$, cubic: $p=3$, ⋮) then

$$\|u_h - u\|_1 \le C_5(\Omega, \tilde{\Omega}, p)\, \kappa\, h^p\, \|f\|_{p-1} \quad \forall f \in H_0^{p-1}(\tilde{\Omega})$$

$$C_5 = \frac{1}{\alpha} C_3(p+1) C_4(\Omega, \tilde{\Omega}, p+1)$$

The isoparametric theorem should look something like:

Theorem: Suppose $\partial\Omega$ is C^t and we use isoparametric elements of degree $p \ge t-1$. Then

$$\|u_h - u\|_1 \le C_5(\Omega, t)\, \kappa\, h^p\, \|f\|_{p-1} \quad \forall f \in H^{p-1}(\Omega)$$

2 issues I haven't worked out: (1) κ needs to take into account higher derivatives of F since F is no longer affine

(2) we need to estimate the errors in the slivers between Ω and Ω_h. Also, Ω_h is probably not strictly contained in Ω, so we need to study non-conforming elements.

228B Lec 31

Last time: error analysis, how the H^m seminorm transforms under an affine map
interpolation on the reference element
shape regular meshes (bounding $\|DF\|$, $\|DF^{-1}\|$ in terms of K)
interpolation on the mesh
elliptic regularity theorem

today: $\begin{cases} L^2 \text{ error estimates} \\ \text{Neumann problem} \end{cases}$

☆ $\|u - I_h u\|_m \le C_3(t) K^m h^{t-m} |u|_t$ $\forall u \in H^t(\Omega_h)$, $0 \le m \le t$, $t \ge 2$,

L^2 error estimates

recall the method of proof for H^1 errors:

$$\|u-u_h\|_1 \le \frac{1}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_1 \le \frac{1}{\alpha}\|u - I_h u\|_1 \le \frac{C_3}{\alpha} K h |u|_2$$
$$\le \frac{C_3 C_4}{\alpha} K h \|f\|_0$$

☆ still works for L^2:

$$\|u - I_h u\|_0 \le C_3(2) K^0 h^{2-0} |u|_2 = C_3 h^2 |u|_2$$

intuitively expect more accuracy in u than Du

~~but the bilinear form $a(\cdot,\cdot)$ is~~ not continuous (or even defined) on L^2
~~so Cea's lemma breaks down.~~

It turns out you do get an extra power of h in the L^2 error.

proof (Nitsche's trick, a duality argument)

define $e = u - u_h$

Galerkin orthogonality

as in proof of Cea, $a(e,v) = 0 \quad \forall v \in S_h$

Let φ solve $\quad -\Delta\varphi = e \quad$ in Ω

$\varphi = 0 \quad$ on $\partial\Omega$

Ω convex $\Rightarrow \|\varphi\|_2 \leq C_4 \|e\|_0$. $\quad C_4$ indep. of e

Green's formula: $(e,e) = -(e,\Delta\varphi) = a(e,\varphi)$

Note that $I_h\varphi \in S_h$, so $a(e, I_h\varphi) = 0$

$\therefore \quad (e,e) = a(e, \varphi - I_h\varphi)$

$\|e\|_0^2 = (e,e) = a(e,\varphi - I_h\varphi) \leq \|e\|_1 \|\varphi - I_h\varphi\|_1$

finally we get to use the interpolation theorem:

$$\|\varphi - I_h\varphi\|_1 \leq C_3 K h |\varphi|_2 \leq \underbrace{C_3 C_4 K}_{C} h \|e\|_0$$

divide through by $\|e\|_0$:

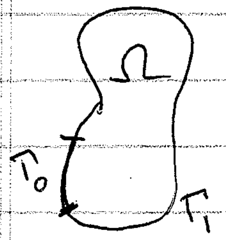
$$\|e\|_0 \leq Ch\|e\|_1 \leq Ch^2 |u|_2 \leq Ch^2 \|f\|_0$$

different C's

second order in L^2

first order in H^1

Mixed bounday conditions



$$-\Delta u = f$$

$$u = 0 \text{ on } \Gamma_0 \quad \leftarrow \text{ assume } \Gamma_0 \text{ has positive length for now}$$

$$\frac{\partial u}{\partial n} = g \text{ on } \Gamma_1$$

closure in H^1 norm

define Hilbert space $V = \overline{\{ u \in H^1(\Omega) \cap C^\infty(\Omega) : u = 0 \text{ in a neighborhood of } \Gamma_0 \}}$

in other words, V is the set of all $H^1(\Omega)$ functions that can be gotten to arbitrarily closely by C^∞ functions that vanish near Γ_0.

note that $H^1_0(\Omega) \subset V \subset H^1(\Omega)$

Any classical soln of this problem satisfies

$$\iint_\Omega \overbrace{-v\Delta u}^{vf}\, dx\,dy = \int_\Gamma -v \nabla u \cdot n \, ds + \iint_\Omega \nabla u \cdot \nabla v \, dx\,dy$$

for $v \in V$, the integral over Γ_0 is zero since $v = 0$ there.
" " " Γ_1 is $\int_{\Gamma_1} vg\, ds$ since $\nabla u \cdot n = g$ there.

def: a weak soln of the mixed b.v.p. is a function $u \in V$ s.t.

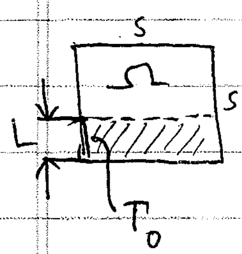
$$a(u,v) = \iint_\Omega fv \, dx\,dy + \int_{\Gamma_1} gv \, ds$$

To apply Lax-Milgram, we need to show that $a(\cdot,\cdot)$ is coercive on V and that $\int_{\Gamma_1} g v \, ds$ is bounded on V.

<u>coercivity</u>: Claim: Poincaré-Friedrichs holds as long as Γ_0 has positive length. (i.e. $\exists\, C>0$ s.t.

$$\|u\|_0 \leq C|u|_1 \quad \forall u \in V)$$

<u>proof</u>: this is hard to prove in general, but suppose Ω is a square and Γ_0 is a line segment of length L.



Let $R = (0,s)\times(0,L)$ be the shaded region shown.

Let $u \in C^\infty(\Omega)\cap H^1(\Omega)$ s.t. $u=0$ near Γ_0.

Choose $(x,y)\in\Omega$ and $y' \in (0,L)$.

Then $u(x,y) = u(x,y') + \int_{y'}^{y} u_y(x,t)\,dt$ F.T.O.C.

for any numbers, $(a+b)^2 \leq 2a^2 + 2b^2$ Young's inequality

$$\therefore \quad u(x,y)^2 \leq 2u(x,y')^2 + 2\left(\int_{y'}^{y} u_y(x,t)\,dt\right)^2$$

Cauchy-Schwarz:

$$\left(\int_{y'}^{y} u_y \, dt\right)^2 \leq \int_{y'}^{y} 1^2\,dt \int_{y'}^{y} u_y^2\,dt$$

$$\leq \quad s\int_0^s u_y^2\,dt$$

so $$u(x,y)^2 \le 2u(x,y')^2 + 2s\int_0^s u_y(x,t)^2\,dt$$

now let's integrate with respect to x, y and y' (3 integrals)

$\int_0^L \cdots dy'$: $$L\,u(x,y)^2 \le 2\int_0^L u(x,y')^2\,dy' + 2sL\int_0^s u_y(x,t)^2\,dt$$

$\iint_\Omega \cdots dx\,dy$: $$L\iint_\Omega u(x,y)^2\,dx\,dy \le 2s\int_0^s\int_0^L u(x,y')^2\,dy'\,dx + 2s^2L\int_0^s\int_0^s u_y(x,t)^2\,dt\,dx$$

$$L\|u\|_{0,\Omega}^2 \le 2s\|u\|_{0,R}^2 + 2s^2L\,|u|_{1,\Omega}^2$$

$$\|u\|_{0,\Omega}^2 \le 2\frac{s}{L}\|u\|_{0,R}^2 + 2s^2|u|_{1,\Omega}^2$$

over the rectangle R, the previous proof works:

$$\|u\|_{0,R}^2 \le \frac{s^2}{2}|u|_{1,R}^2 \le \frac{s^2}{2}|u|_{1,\Omega}^2$$

see lec 25

so $$\|u\|_{0,\Omega}^2 \le 2\frac{s}{L}\left(\frac{s^2}{2}|u|_{1,\Omega}^2\right) + 2s^2|u|_{1,\Omega}^2$$

$$\le \left(\frac{s}{L}+2\right)s^2|u|_{1,\Omega}^2$$

and $$\|u\|_{0,\Omega} \le \sqrt{2+\frac{s}{L}}\; s\,|u|_{1,\Omega}$$

density of ($C^\infty \cap H^1$ vanishing near Γ_0) in V gives the result for all $u \in V$.

$$\|u\|_1^2 = \|u\|_0^2 + |u|_1^2 \le \left[1 + \left(2+\frac{s}{L}\right)s^2\right] a(u,u) \quad \forall u \in V.$$

$\therefore$ a is coercive. $\alpha = [\cdots]^{-1}$

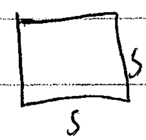
Last step: the linear functional $\langle l, v \rangle = \int_{\Gamma_1} v g \, ds$ is bounded on V.

<u>Theorem:</u> (trace theorem): there is a bounded linear operator

$\gamma : H^1(\Omega) \to L^2(\Gamma)$ such that $\gamma v = v|_{\Gamma}$ for $v \in C^1(\bar{\Omega})$

trace operator

restriction operator

<u>proof</u>: again for the rectangle. $\Omega =$ 

Let $u \in C^1(\bar{\Omega})$.

$$u(x,0) = u(x,y) - \int_0^y u_y(x,t)\,dt$$

$$u(x,0)^2 \le 2\, u(x,y)^2 + 2\left(\int_0^y u_y^2\, dt\right)^2 \le 2\, u(x,y)^2 + 2s\int_0^s u_y(x,t)^2\, dt$$

$$s\int_0^1 u(x,0)^2\,dx \le 2\, \|u\|_0^2 + 2s^2 |u|_1^2 \le \underbrace{\max(2, 2s^2)}_{\le 2(1+s^2)} \|u\|_1^2$$

repeat on other 3 sides

$$\int_{\partial\Omega} u^2\, ds \le 8\left(s + \tfrac{1}{s}\right)\|u\|_1^2$$

$$\|u\|_{0,\Gamma} \le \sqrt{8\left(s+\tfrac{1}{s}\right)}\, \|u\|_1$$

LHS is $\|\gamma u\|_{0,\Gamma}$. This shows $\gamma : C^1(\bar{\Omega}) \to L^2(\Gamma)$ is bounded when $C^1(\bar{\Omega})$ is given the H^1 norm.

Since $C^1(\bar{\Omega})$ is dense in $H^1(\Omega)$ and $L^2(\Gamma)$ is complete, γ extends continuously to $H^1(\Omega)$ without increasing its norm.

so $\langle l, v\rangle = \int_{T_1} v g\, ds$ is a composition of two bounded operators: $H^1(\Omega) \xrightarrow{\gamma} L^2(T^1) \xrightarrow{\int_{T_1} \cdot\, g ds} \mathbb{R}$

$\therefore$ l is bounded.

$\therefore \exists!\ u \in V$ s.t. $a(u,v) = \iint_\Omega fv\, dxdy + \int_{T_1} gv\, ds$

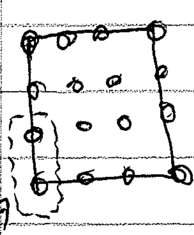
for all $v \in V$.

$\therefore$ if there is a classical solution, we have found it.

(classical solution are weak solutions and weak solution are unique.)

In the finite element framework, Neumann b.c.'s are imposed "naturally", i.e. you just leave them as variables like the interior unknowns and the solution ends up having the right slope in the mesh refinement limit.

(all of our convergence theorems work the same for the mixed problem, except there could be singularities, preventing $\|u\|_2 \leq C\|f\|_0$. where T_0 meets T_1



$$a(u_h, v_h) = \iint_\Omega f v_h\, dxdy + \int_T g v_h\, ds$$

principle of virtual work in mechanics.

Dirichlet conditions

Remove them from the system. (impose their values directly)