Numerical Methods for Partial Differential Equations

CS555 / MATH555 / CSE510

Andreas Kloeckner

March 25, 2021

Outline

Outline

Outline

Outline

What's the point of this class?

PDEs describe lots of things in nature:

- Fluid flow (Navier-Stokes equations)
- Electromagnetism (Maxwell's equations)
- Waves (Elasticity, Acoustics)
- Plasmas (Magnetohydrodynamics)

Idea: Use them to

- Make predictions (and check them, to validate the model: science!)
- Use predictions (for design of cars, airplanes, reactors, ...)

Survey

- Home dept
- Degree pursued
- Longest program ever written
  - in Python?
- Research area

Class web page

https://bit.ly/numpde-s20

- Book Draft
- Notes, Class Outline
- Assignments (submission and return)
- Piazza
- Grading Policies/Syllabus
- Video
- Scribbles
- Demos (binder)

Sources for these Notes

- Adler, James, Hans De Sterck, Scott MacLachlan, and Luke N. Olson. *Numerical Partial Differential Equations*, 2020. (draft)
- Strikwerda, John C. *Finite Difference Schemes and Partial Differential Equations*, Second Edition. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2004.
- LeVeque, Randall J. *Numerical Methods for Conservation Laws*. 2nd ed. Birkhäuser Basel, 1992.
- Braess, Dietrich. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 2007.
- Shu, Chi-Wang. *Lecture Notes for AM257*, Brown University, Fall 2006.
- Heuveline, Vincent. *Lecture Notes for "Numerik für PDEs"*. Universität Karlsruhe, Summer 2005.
- Various prior bits of material by Luke Olson and Stephen Bond.

Open Source <3

These notes (and the accompanying demos) are open-source!

Bug reports and pull requests welcome:
https://github.com/inducer/numpde-notes

Outline

PDEs: Example I

What does this do? $\partial_t u = \partial_x u$

- Slope in x and t matches
- Single profile on an x/t diagonal
- Which one? (left-leaning)
- We'll deal with this a lot.
  - Advection equation, one-way wave equation
  - General solution: $u(x, t) = u_0(x + t)$

PDEs: Example II

What does this do? $\partial_x^2 u + \partial_y^2 u = 0$

- Second derivative measures "bendiness" of a function
- "Bendiness" in x and y need to add up to zero
- Can a function like this have a maximum?

Some good questions

- What is a time-like variable? (Variables labeled t?)
- What if there are boundaries?
  - In space?
  - In time?
- Existence and Uniqueness of Solutions?
  - Depends on where we look (the *function space*)
  - In the case of the two examples? (if there are no boundaries?)

Some general takeaways:

- Don't check common sense at the door.
- Think about what the PDE is "trying" to say.
- Develop physical intuition.

PDEs: An Unhelpfully Broad Problem Statement

Looking for $u : \Omega \to R^n$ where $\Omega \subseteq \mathbb{R}^d$ so that $u \in V$ and

$$F(u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}, \dots, x, y, \dots) = 0$$

Notation

Used as convenient:

$$u_x = \partial_x u = \frac{\partial u}{\partial x}$$

Properties of PDEs

What is the order of the PDE?

> The highest (total, i.e. summing over axes) order of derivative occuring in F.

When is the PDE linear?

> If u and v are solutions, $\alpha u + \beta v$ are, too.

When is the PDE quasilinear?

> The dependency in F on the highest-order partial derivatives is linear in u.

When is the PDE semilinear?

Examples: Order, Linearity?

$$(xu^2)u_{xx} + (u_x + y)u_{yy} + u_x^3 + yu_y = f$$

Second-order quasilinear

$$(x + y + z)u_x + (z^2)u_y + (\sin x)u_z = f$$

First-order semilinear

Properties of Domains

- smooth
- with corners
- with *reentrant* corners
- with cusps

May influence existence/uniqueness of solutions!

Function Spaces: Examples

Name some function spaces with their norms.

$C(\Omega)$	f continuous, $\|f\|_\infty := \sup_{x\in\Omega} \|f(x)\|$
$C^k(\Omega)$	f k-times continuously differentiable
$C^{0,\alpha}(\Omega)$	$\|f\|_\alpha := \|f\|_\infty + \sup_{x\neq y} \frac{\|f(x)-f(y)\|}{\|x-y\|^\alpha}$ $(\alpha \in (0,1))$
$C_L(\Omega)$	$\|f(x) - (y)\| \le L \|x-y\|$
$L_p(\Omega)$	$\|f\|_{p,\Omega} := \sqrt[p]{\int_D \|f(x)\|^p dx} < \infty$
	Why do these only define equivalence classes?
	L_2 special because...?
$W_p^1(\Omega)$	$\|f\|_{W_1^p(\Omega)} := (\|f\|_{p,\Omega} + \|f'\|_{p,\Omega}) < \infty$
$H^1(\Omega)$	equivalent to $W_2^1(\Omega)$, also a Hilbert space

May *also* influence existence/uniqueness of solutions!

Solving PDEs

Closed-form solutions:

- If separation of variables applies to the domain: good luck with your ODE
- If not: Good luck! $\rightarrow$ Numerics

General Idea (that we will follow some of the time)

- Pick $V_h \subseteq V$ finite-dimensional
  - h is often a *mesh spacing*
- Approximate u through $u_h \in V_h$
- Show: $u_h \rightarrow u$ (in some sense) as $h \rightarrow 0$

Example

$u(x) = \sin x$ where V_h is piecewise constant functions with grid spacing h.

About grand big unifying theories

Is there a grand big unifying theory of PDEs?

No. Frustratingly, studying PDEs is a little bit like stamp collecting. For instance, there are broad classes of second-order PDEs that behave mostly alike.

Collect some stamps

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

Discriminant value	Kind	Example
$b^2 - ac < 0$	Elliptic	Laplace $u_{xx} + u_{yy} = 0$
$b^2 - ac = 0$	Parabolic	Heat $u_t = u_{xx}$
$b^2 - ac > 0$	Hyperbolic	Wave $u_{tt} = u_{xx}$

Where do these names come from?

Quadratic forms: $ax^2 + 2bxy + cy^2 +$ lower order terms

PDE Classification in Other Cases

Scalar first order PDEs?

Have characteristics, therefore classified as hyperbolic. (See later.)

First order systems of PDEs?

Can be classified into hyperbolic/elliptic/parabolic as well, using slightly more complicated method, depending on the direction of the characteristics. See for example Loret '08.

Classification in higher dimensions

$$Lu := \sum_{i=1}^{d}\sum_{j=1}^{d} a_{i,j}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \text{lower order terms}$$

Consider the matrix $A(x) = (a_{ij}(x))_{i,j}$. May assume A symmetric. Why?

Schwarz's theorem. So: real-valued eigenvalues.

What cases can arise for the eigenvalues?

Case	Kind
$\lambda_j(x) = 0$ for some λ	parabolic
$\lambda_j(x)$ all have the same sign	elliptic
$\lambda_j(x)$ all but one have the same sign	hyperbolic
$\lambda_j(x) > 1$ eigenvalue per sign, nonsingular	ultra-hyperbolic

Elliptic PDE: Laplace/Poisson Equation

$$\triangle u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} = \nabla \cdot \nabla u(x) \overset{\text{2D}}{=} u_{xx} + u_{yy} = f(x) \quad (x \in \Omega)$$

Called Laplace equation if $f = 0$. With Dirichlet boundary condition

$$u(x) = g(x) \qquad (x \in \partial\Omega).$$

Demo: Elliptic PDE Illustrating the Maximum Principle

Elliptic PDEs: Singular Solution

Demo: Elliptic PDE Radially Symmetric Singular Solution

Given $G(x) = C\log(|x|)$ as the free-space Green's function, can we construct the solution to the PDE with a more general f?

$$u(x) = (G * f)(x) = \int_{\mathbb{R}^d} G(x-y)f(y)dy$$

What can we learn from this?

Solutions to the Laplace equation are globally coupled. The value of f at any point influences the solution *everywhere* (if only a little)

Elliptic PDEs: Justifying the Singular Solution

$$u(x) = (G * f)(x) = \int_{\mathbb{R}^d} G(x-y)f(y)dy$$

Why?

$$\triangle u(x) = (\triangle G * f)(x) = \int_{\mathbb{R}^d} (\triangle G(x-y))f(y)dy$$
$$= \int_{\mathbb{R}^d} \delta(x-y)f(y)dy = f(x)$$

Parabolic PDE: Heat Equation · Separation of Variables

$$u_t = u_{xx} \qquad ((x,t) \in [0,1] \times [0,T])$$
$$u(x,0) = g(x) \qquad (x \in [0,1])$$
$$u(0,t) = u(1,t) = 0 \qquad (t \in [0,T])$$

Looking for $u(x,t) = v(t) \cdot w(x)$.
Plug into PDE: $v'(t) \cdot w(x) = v(t) \cdot w''(x)$. Divide:

$$\frac{v'(t)}{v(t)} = C = \frac{w''(x)}{w(x)},$$

where C is constant since it is independent of x and t.

- $w'' = Cw$ with BCs yields $w(x) = \alpha \cdot \sin(m\pi x)$ and $C = -m^2\pi^2$ or any linear combination; Fourier to match g.
- Focus on specific value of m: $v' = Cv$ with ICs yields $v(t) = \exp(-m^2\pi^2 t)$.

Parabolic PDE: Solution Behavior

Demo: Parabolic PDE What can we learn from analytic and numerical solution?

- Heat equation 'washes out' the solution
- Appears to obey a maximum principle
- Appears to smooth the data

Hyperbolic PDE: Wave Equation

$$u_{tt} = c^2 u_{xx} \qquad ((x,t) \in \mathbb{R} \times [0, T])$$
$$u(x,0) = g(x) \qquad (x \in \mathbb{R})$$

with $g(x) = \sin(\pi x)$.

Is this problem well-posed?

No, missing initial condition on u_t.

$$u_t(x,0) = 0 \qquad (x \in \mathbb{R})$$

Can be rewritten in conservation law form:

$$q_t(x) + \nabla \cdot F(q(x)) = s(x)$$

Hyperbolic Conservation Laws

$$q_t(x,t) + \nabla \cdot \boldsymbol{F}(q(x,t)) = s(x)$$

Why is this called a conservation law?

- Balance between a conserved quantity q and a flux f.
- Flux prescribes the 'flow direction'. When is flux divergence < 0?
- s is a source term.

$F :? \to ?$

- $q(x,t) \in \mathbb{R}^n$
- $F : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^d$

Wave Equation as a Conservation Law

Rewrite the wave equation in conservation law form:

Introduce a new variable v and let

$$u_t = cv_x$$
$$v_t = cu_x.$$

Observe $u_{tt} = cv_{xt} = c^2 u_{uxx}$. Define $q := \begin{bmatrix} u & v \end{bmatrix}^T$.

Solving Conservation Laws

Solve

$$u_t = v_x$$
$$v_t = u_x.$$

$$q_t + \begin{bmatrix} 0 & -c \\ -c & 0 \end{bmatrix} q_x = Aq_x = 0$$

Diagonalize: Define $\tilde{q} := V^{-1}q$,

$$\tilde{q}_t + V^{-1}AV\tilde{q}_x = \begin{bmatrix} c & 0 \\ 0 & -c \end{bmatrix} \tilde{q}_x = 0$$

$\rightarrow$ two advection equations

Solution, for some ϕ_ℓ, ϕ_r: $u(t,x) = \phi_\ell(x+ct) + \phi_r(x-ct)$

Demo: Hyperbolic PDE

Hyperbolic: Solution Properties

Properties of the solution for hyperbolic equations:

- Has *conserved quantities*
- q, "energy" ($\to$ HW1)
- Maintains smoothness of IC
- Typical trick: Project to one dimension, diagonalize, understand advection behavior.

Outline

Interpolation and Vandermonde Matrices

Limit the set of functions to a linear combination from an *interpolation basis* φ_i.

$$f(x) = \sum_{j=0}^{N_{\text{func}}} \alpha_j \varphi_j(x)$$

Interpolation becomes solving the linear system:

$$y_i = f(x_i) = \sum_{j=0}^{N_{\text{func}}} \alpha_j \underbrace{\varphi_j(x_i)}_{V_{ij}} \qquad \leftrightarrow \qquad V\boldsymbol{\alpha} = \boldsymbol{y}.$$

Want unique answer: Pick $N_{\text{func}} = N \to V$ square.
V is called the *(generalized) Vandermonde matrix*.

$$V\,(\text{coefficients}) = (\text{values at nodes})\,.$$

Finite Differences Numerically

Demo: Finite Differences
Demo: Finite Differences vs Noise
Demo: Floating point vs Finite Differences

Taking Derivatives Numerically

Why *shouldn't* you take derivatives numerically?

- 'Unbounded'
  A function with small $\|f\|_\infty$ can have arbitrarily large $\|f'\|_\infty$
- Amplifies noise
  Imagine a smooth function perturbed by small, high-frequency wiggles
- Subject to cancellation error
- Inherently less accurate than integration
  - Interpolation: h^n
  - Quadrature: h^{n+1}
  - Differentiation: h^{n-1}
    (where n is the number of points)

Demo: Taking Derivatives with Vandermonde Matrices

Differencing Order of Accuracy Using Taylor

Find the order of accuracy of the finite difference formula $f'(x) \approx [f(x+h) - f(x-h)]/2h$.

$$\begin{aligned}
& f'(x) - \frac{f(x+h) - f(x-h)}{2h} \\
=& f'(x) - \frac{1}{2h}\left[f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + O(h^4)\right] \\
& + \frac{1}{2h}\left[f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x)\right] \\
=& \frac{1}{2h} \cdot \frac{h^3}{6}f'''(x) \quad \text{as } h \to 0.
\end{aligned}$$

Outline

Interpolation Error

If f is n times continuously differentiable on a closed interval I and $p_{n-1}(x)$ is a polynomial of degree at most n that interpolates f at n distinct points $\{x_i\}$ $(i = 1, ..., n)$ in that interval, then for each x in the interval there exists ξ in that interval such that

$$f(x) - p_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!}(x - x_1)(x - x_2)\cdots(x - x_n).$$

Set the error term to be $R(x) := f(x) - p_{n-1}(x)$ and set up an auxiliary function:

$$Y(t) = R(t) - \frac{R(x)}{W(x)}W(t) \quad \text{where} \quad W(t) = \prod_{i=1}^{n}(t - x_i).$$

Note also the introduction of t as an additional variable, independent of the point x where we hope to prove the identity.

Interpolation Error: Proof cont'd

$$Y(t) = R(t) - \frac{R(x)}{W(x)} W(t) \quad \text{where} \quad W(t) = \prod_{i=1}^{n} (t - x_i)$$

- Since x_i are roots of $R(t)$ and $W(t)$, we have $Y(x) = Y(x_i) = 0$, which means Y has at least $n+1$ roots.
- From Rolle's theorem, $Y'(t)$ has at least n roots, then $Y^{(n)}$ has at least one root ξ, where $\xi \in I$.
- Since $p_{n-1}(x)$ is a polynomial of degree at most $n-1$, $R^{(n)}(t) = f^{(n)}(t)$. Thus

$$Y^{(n)}(t) = f^{(n)}(t) - \frac{R(x)}{W(x)} n!.$$

- Plugging $Y^{(n)}(\xi) = 0$ into the above yields the result.

Error Result: Connection to Chebyshev

What is the connection between the error result and Chebyshev interpolation?

- The error bound suggests choosing the interpolation nodes such that the product $|\prod_{i=1}^{n}(x - x_i)|$, is as small as possible. The Chebyshev nodes achieve this.
- Error is zero at the nodes
- If nodes scoot closer together near the interval ends, then

$$(x - x_1)(x - x_2)\cdots(x - x_n)$$

clamps down the (otherwise quickly-growing) error there.

Error Result: Simplified From

Boil the error result down to a simpler form.

Assume $x_1 < \cdots < x_n$.

- $\left|f^{(n)}(x)\right| \le M$ for $x \in [x_1, x_n]$,
- Set the interval length $h = x_n - x_1$.
  Then $|x - x_i| \le h$.

Altogether–there is a constant C independent of h so that:

$$\max_x |f(x) - p_{n-1}(x)| \le CMh^n.$$

For the grid spacing $h \to 0$, we have

$$E(h) = O(h^n).$$

This is called *convergence of order n*.

Demo: Interpolation Error

Outline

Outline

1D Advection Equation and Characteristics

$$u_t + au_x = 0, \quad u(0,x) = g(x) \qquad (x \in \mathbb{R})$$

Solution?

Generalize to 1D conservation law: $u_t + f(u)_x = 0$. Find solution.

Characteristic Curve: Define a function $x(t)$ so that $u(x(t),t) = u(x_0, 0)$.

$$\begin{cases} \frac{\mathrm{d}x(t)}{\mathrm{d}t} = f'(u(x(t),t)), \\ x(0) = x_0. \end{cases}$$

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = u_x x'(t) + u_t = u_x f'(u(x(t),t) + u_t = f(u)_x + u_t = 0.$$

So $u(x(t),t) = u(x(0),0) = g(x_0)$.

Solving Advection with Characteristics

$$u_t + au_x = 0, \quad u(0,x) = g(x) \qquad (x \in \mathbb{R})$$

Find the characteristic curve for advection.

Here $x(t) = x_0 + at$.

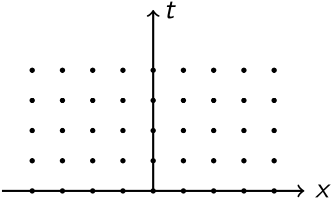
Generalize this to a solution formula.

General solution of advection: $u(t,x) = g(x - at)$. a: Advection speed.

Does the solution formula admit solutions that aren't obviously allowed by the PDE?

Solution formula allows nonsmooth profiles. Unclear: Those are not differentiable.

Finite Difference for Hyperbolic: Idea



$$\{(x_k, t_\ell) : x_k = kh_x, t_\ell = \ell h_t\}$$

If $u(x, t)$ is the exact solution, want

$$u_{k,\ell} \approx u(x_k, t_\ell).$$

Condition at each grid point?

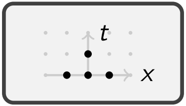
- Pick a derivative stencil for each derivative term in the PDE
- Get system of equations
- Solve

What are explicit/implicit schemes?

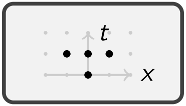
Implicit require solution of a system of equations

Designing Stencils

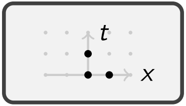
ETCS:



ITCS:



ETFS:



ETBS:

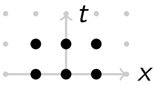
Terminology?

- E Explicit / I Implicit
- T Time / S Space
- F Forward: right
- B Backward: left
- Upwind: left if $a < 0$
- Downwind: right if $a > 0$

Write out ITCS:

$$\frac{u_{k,\ell+1} - u_{k,\ell}}{h_t} + a\frac{u_{k+1,\ell+1} - u_{k-1,\ell+1}}{2h_x} = 0$$

Crank-Nicolson



Crank-Nicolson

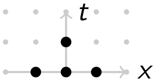
Write out Crank-Nicolson:

$$\frac{u_{k,\ell+1} - u_{k,\ell}}{h_t} + \frac{a}{2}\left[\frac{u_{k+1,\ell+1} - u_{k-1,\ell+1}}{2h_x} + \frac{u_{k+1,\ell} - u_{k-1,\ell}}{2h_x}\right] = 0$$

Lax-Wendroff

What's the core idea behind Lax-Wendroff?

- Write out a Taylor expansion in time
- Use the PDE to replace time ∂ with space ∂
- Allows two-level schemes of any order of accuracy



Lax-Wendroff

Write out Lax-Wendroff.

$u_t = -au_x$ so also $u_{tt} = -a(u_x)_t = -a(u_t)_x = a^2 u_{xx}$.

$$
\begin{aligned}
u_{k,\ell+1} - u_{k,\ell} &\approx h_t u_t(x_k, t_\ell) + \frac{h_t^2}{2} u_{tt}(x_k, t_\ell) \\
&= -h_t a u_x(x_k, t_\ell) + \frac{h_t^2}{2} a^2 u_{xx}(x_k, t_\ell) \\
&\approx -h_t a \frac{u_{k+1,\ell} - u_{k-1,\ell}}{2h_x} + \frac{h_t^2 a^2}{2} \cdot \frac{u_{k+1,\ell} - 2u_{k,\ell} + u_{k-1,\ell}}{h_x^2}
\end{aligned}
$$

Exploring Advection Schemes

Demo: Methods for 1D Advection

- Which of the schemes "work"?
- Any restrictions worth noting?

Outline

A Matrix View of Two-Level Stencil Schemes

Define

$$\boldsymbol{v}_\ell = \begin{bmatrix} u_{1,\ell} \\ \vdots \\ u_{N_x,\ell} \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_{N_t} \end{bmatrix}.$$

Define

$$\boldsymbol{u}_\ell = \begin{bmatrix} u(x_1, t_\ell) \\ \vdots \\ u(x_{N_x}, t_\ell) \end{bmatrix} \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_{N_t} \end{bmatrix}.$$

Definition (Two-Level Finite Difference Scheme)

A finite difference scheme that can be written as

$$P_h \boldsymbol{v}_{\ell+1} = Q_h \boldsymbol{v}_\ell + h_t \boldsymbol{b}_\ell$$

is called a two-level linear finite difference scheme.

- Mostly $\boldsymbol{b}_\ell = 0$, i.e. homogeneous schemes, no source terms.
- P_h and Q_h may depend on both h_x and h_t.
- P_h and Q_h and the spatial grid may also be infinite

Rewriting Schemes in Matrix Form (1/2)

$$P_h \boldsymbol{v}_{\ell+1} = Q_h \boldsymbol{v}_\ell + h_t \boldsymbol{b}_\ell$$

Find P_h and Q_h for ETCS:

ETCS:

$$\frac{u_{k,\ell+1} - u_{k,\ell}}{h_t} + a\frac{u_{k+1,\ell} - u_{k-1,\ell}}{2h_x} = 0.$$

Equivalently:

$$u_{k,\ell+1} = u_{k,\ell} + \frac{ah_t}{2h_x}(-u_{k+1,\ell} + u_{k-1,\ell}).$$

So

$$P_h = I, \qquad Q_h = \text{tridiag}\left(\frac{ah_t}{2h_x}, 1, -\frac{ah_t}{2h_x}\right).$$

Rewriting Schemes in Matrix Form (2/2)

Find P_h and Q_h for Crank-Nicolson:

$$P_h = \text{tridiag}\left(-\frac{ah_t}{4h_x}, 1, \frac{ah_t}{4h_x}\right),$$

and

$$Q_h = \text{tridiag}\left(\frac{ah_t}{4h_x}, 1, -\frac{ah_t}{4h_x}\right).$$

Truncation Error

Definition (Truncation Error)

The local truncation error $\tau_{k,\ell}$ is the error that remains when a finite difference method is applied to a smooth exact solution u at (x_k, t_ℓ).

Demo: Truncation Error Analysis via sympy

Error and Error Propagation

Express truncation error in our two-level framework:

$$P_h \boldsymbol{u}_{\ell+1} = Q_h \boldsymbol{u}_\ell + \boldsymbol{\tau}_\ell h_t.$$

Define $\boldsymbol{e}_\ell = \boldsymbol{u}_\ell - \boldsymbol{v}_\ell$. Understand the error as accumulation of truncation error:

Recall $P_h \boldsymbol{v}_{\ell+1} = Q_h \boldsymbol{v}_\ell$. Subtract from the truncation error definition to find:

$$\begin{aligned} \boldsymbol{e}_0 &= 0 \\ P_h \boldsymbol{e}_{\ell+1} &= Q_h \boldsymbol{e}_\ell + \boldsymbol{\tau}_\ell h_t \\ \boldsymbol{e}_{\ell+1} &= P_h^{-1} Q_h \boldsymbol{e}_l + P_h^{-1} \boldsymbol{\tau}_\ell h_t. \end{aligned}$$

Discrete and Continuous Norms

To measure properties of numerical solutions we need norms. Define a discrete L^∞ norm.

$$\|\boldsymbol{e}\|_\infty = \max_{k,\ell} |e_{k,\ell}|.$$

Define a discrete L^2 norm.

$$\|\boldsymbol{e}\|_2 = \sqrt{\sum_{k,\ell} e_{k,\ell}^2 h_x h_t}.$$

Important features:

- Value of discrete norm should not change wildly if h_x and h_t change (and, along with them, the number of nodes).

Consistency and Convergence

Assume $u, (\partial_x^{q_x})u, (\partial_t^{q_t})u \in L^2(\mathbb{R} \times [0, t^*])$.

Definition (Consistency)

A two-level scheme is consistent in the L^2-norm with order q_t in time and q_x in space if

$$\max_{\ell, \ell h_t \le t^*} \|\boldsymbol{\tau}_\ell\| = O(h_x^{q_x} + h_t^{q_t}) \qquad \text{as } (h_x, h_t) \to (0, 0).$$

Definition (Convergence)

A two-level scheme is convergent in the L^2-norm with order q_t in time and q_x in space if

$$\max_{\ell, \ell h_t \le t^*} \|\boldsymbol{e}_\ell\| = O(h_x^{q_x} + h_t^{q_t}) \qquad \text{as } (h_x, h_t) \to (0, 0).$$

Analyzing ETFS

$$\frac{u_{k,\ell+1} - u_{k,\ell}}{h_t} + a\frac{u_{k+1,\ell} - u_{k,\ell}}{h_x} = 0$$

Let's understand more precisely what happens for this scheme.

Rewrite as

$$u_{k,\ell+1} = u_{k,\ell} - \frac{ah_t}{h_x}(u_{k+1,\ell} - u_{k,\ell}) = (1+\lambda)u_{k,\ell} - \lambda u_{k+1,\ell}$$

for $\lambda = ah_t/h_x$.

ETFS Part 2

$$u_{k,\ell+1} = (1+\lambda)u_{k,\ell} - \lambda u_{k+1,\ell}$$

Consider $u(x,0) = 1_{[-1,0]}(x)$. Predict solution behavior.

$$\begin{aligned} u_{0,0} &= 1 & u_{1\ldots,0} &= 0 \\ u_{0,1} &= (1+\lambda) & u_{1\ldots,1} &= 0 \\ u_{0,2} &= (1+\lambda)^2 & u_{1\ldots,2} &= 0 \end{aligned}$$

So the right half never "sees" the traveling bump; this can't be convergent. Meanwhile,

$$u(0,t) \approx u_{0,t/h_t} = \left(1 + \frac{ah_t}{h_x}\right)^{t/h_t} = \left(1 + \frac{a/h_x}{1/h_t}\right)^{t/h_t} = \exp\left(\frac{at}{h_x}\right)$$

Demo: Methods for 1D Advection (Revisit ETFS)

Stability

$$P_h \boldsymbol{v}_{\ell+1} = Q_h \boldsymbol{v}_\ell$$

Write down a matrix product to bring $\boldsymbol{v}_0$ to $\boldsymbol{v}_\ell$:

$$\boldsymbol{v}_\ell = (P_h^{-1} Q_h)^\ell \boldsymbol{v}_0$$

Definition (Stability)

A two-level scheme is stable in the L^2-norm if there exists a constant $c > 0$ independent of h_t and h_x so that

$$\left\| (P_h^{-1} Q_h)^\ell P_h^{-1} \right\| \leq c$$

for all ℓ and h_t such that $\ell h_t \leq t^*$.

Lax Convergence Theorem

Theorem (Lax Convergence)

If a two-level FD scheme is

- *consistent in the L^2-norm with order q_t in time and q_x in space, and*
- *stable in the L^2-norm, then*

it is convergent in the L^2-norm with order q_t in time and q_x in space.

A stronger result holds: The above is actually "if and only if".
(called the Lax Equivalence Theorem or Lax-Richtmyer Theorem)
Think of this as an important 'meta-theorem' of numerical analysis (or "fundamental theorem of NA"):

$$\text{Consistent} + \text{Stable} \Rightarrow \text{Convergent}$$

A related result holds for ODEs, due to Dahlquist.

Lax Convergence: Proof (1/2)

Recall error propagation:

$$P_h \boldsymbol{e}_{\ell+1} = Q_h \boldsymbol{e}_\ell + \boldsymbol{\tau}_\ell h_t$$

So:

$$\boldsymbol{e}_{\ell+1} = P_h^{-1} Q_h \boldsymbol{e}_l + P_h^{-1} \boldsymbol{\tau}_\ell h_t.$$

Since $\boldsymbol{e}_0 = 0$,

$$\begin{aligned} \boldsymbol{e}_1 &= h_t P_h^{-1} \boldsymbol{\tau}_0, \\ \boldsymbol{e}_2 &= h_t (P_h^{-1} Q_h) P_h^{-1} \boldsymbol{\tau}_0 + h_t P_h^{-1} \boldsymbol{\tau}_1. \end{aligned}$$

By induction,

$$\boldsymbol{e}_\ell = h_t \sum_{m=1}^{\ell} (P_h^{-1} Q_h)^{\ell-m} P_h^{-1} \boldsymbol{\tau}_{m-1}.$$

Lax Convergence: Proof (2/2)

$$\boldsymbol{e}_\ell = h_t \sum_{m=1}^{\ell} (P_h^{-1} Q_h)^{\ell-m} P_h^{-1} \boldsymbol{\tau}_{m-1}.$$

Let $\ell h_t \le t^*$. Taking the norm of both sides,

$$\begin{aligned}
\|\boldsymbol{e}_\ell\| &\le h_t \sum_{m=1}^{\ell} \left\| (P_h^{-1} Q_h)^{\ell-m} P_h^{-1} \boldsymbol{\tau}_{m-1} \right\| \\
&\le h_t \sum_{m=1}^{\ell} \underbrace{\left\| (P_h^{-1} Q_h)^{\ell-m} P_h^{-1} \right\|}_{\le c \text{ (stab.)}} \|\boldsymbol{\tau}_{m-1}\| \\
&\le h_t \ell c \cdot \max_{\ell : \ell h_t \le t^*} \|\boldsymbol{\tau}_{m-1}\| \le c t^* \max_{\ell : \ell h_t \le t^*} \|\boldsymbol{\tau}_{m-1}\| \\
&\overset{\text{cons.}}{=} O(h_x^{q_x} + h_t^{q_t}).
\end{aligned}$$

Conditions for Stability

$$\left\|(P_h^{-1}Q_h)^\ell P_h^{-1}\right\| \le c$$

Give a simpler, sufficient condition:

$$\left\|(P_h^{-1}Q_h)^\ell\right\| \le 1, \qquad P_h^{-1}\right\| \le c.$$

Also called Lax-Richtmyer stability.

How can we show bounds on these matrix norms?

- Observe: bounds have to hold for all h_t and h_x.
- Generally: cumbersome.
- Possibly easiest: approach via singular values.
- Bound singular values: For example using Gershgorin.

Stability of ETBS (1/3)

Theorem (Gershgorin)

For a matrix $A \in \mathbb{C}^{N\times N} = (a_{i,j})$,

$$\sigma(A) \subset \bigcup_{j=1}^{N} \bar{B}\left(a_{j,j}, \sum_{k\neq j} |a_{j,k}|\right).$$

ETBS:

$$\frac{u_{k,\ell+1} - u_{k,l}}{h_t} + a\frac{u_{k,\ell} - u_{k-1,\ell}}{h_x} = 0$$

Analyze stability of ETBS:

Let $\lambda = ah_t/h_x$. Then $u_{k,\ell+1} = \lambda u_{k-1,\ell} + (1-\lambda)u_{k,\ell}$.
So $P_h = I$ and $Q_h = \text{tridiag}(\lambda, 1-\lambda, 0)$. $\left\|P_h^{-1}\right\| \leq 1$ trivially.

Stability of ETBS (2/3)

$P_h = I$ and $Q_h = \text{tridiag}(\lambda, 1-\lambda, 0)$.

$$\|Q_h\| = \sqrt{\rho(Q_h^T Q_h)},$$

where $Q_h^T Q_h = \text{tridiag}(\lambda(1-\lambda), (1-\lambda)^2+\lambda^2, \lambda(1-\lambda))$. If $0 \leq \lambda \leq 1$, then $\lambda(1-\lambda) \geq 0$.

$$\begin{aligned} 2\lambda^2 - 2\lambda \leq &\quad \Lambda - (1-\lambda)^2 - \lambda^2 &\leq 2\lambda - 2\lambda^2, \\ 1 - 4\lambda + 4\lambda^2 \leq &\quad \Lambda &\leq 1, \\ 0 \leq (1-2\lambda)^2 \leq &\quad \Lambda &\leq 1. \end{aligned}$$

So $|\Lambda| \leq 1$, which implies $\left\|Q_h^T Q_h\right\| \leq 1$, which means $\|Q_h\| \leq 1$.
If $\lambda > 1$, analogously:
$|\Lambda| \geq 1$, which implies $\left\|Q_h^T Q_h\right\| \geq 1$, which means $\|Q_h\| \geq 1$.

Stability of ETBS (3/3)

Summarize ETBS stability:

We learn that ETBS is stable if $0 \leq \lambda \leq 1$. Rewriting, we obtain

$$\frac{a h_t}{h_x} < 1 \quad \Leftrightarrow \quad h_t \leq \frac{h_x}{a}.$$

This type of stability is called conditional stability, and the condition we found a Courant-Friedrichs-Lewy (CFL) condition.

Comments?

Way cumbersome to prove. Is there something easier that gives necessary/sufficient conditions?

Outline

Discrete Time Fourier Transform

Assume $\boldsymbol{x}$ infinitely long. Define:

$$\hat{\boldsymbol{x}}(\theta) = \sum_k x_k e^{-i\theta k}$$

When is this well-defined?

$$|\hat{\boldsymbol{x}}(\theta)| = \left| \sum_k x_k e^{-i\theta k} \right| \leq \sum_k |x_k| ,$$

Well-defined if $\sum |x_k|$ is absolutely convergent.

Inverting the Fourier Transform

To recover $\boldsymbol{x}$:

$$x_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta) e^{i\theta k} d\theta.$$

Proof?

$$x_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_j x_j e^{-i\theta j} e^{i\theta k} d\theta = \frac{1}{2\pi}\sum_j x_j \int_{-\pi}^{\pi} e^{i\theta(k-j)} d\theta = \sum_j x_j \delta_{j,k}.$$

Getting to L^2

- Fourier Transform well defined for $\boldsymbol{x} \in \ell^1$.
- Problem: We care about L^2, not ℓ^1.

Theorem (Parseval)

If $\|\boldsymbol{x}\|_2 < \infty$, *then*

$$\|\boldsymbol{x}\|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{\boldsymbol{x}}(\theta)|^2 \, d\theta < \infty.$$

Impact?

Can extend definition of Fourier transform to L^2.

Toeplitz Operators

Definition (Toeplitz Operator)

An operator T is a Toeplitz operator if $(T\boldsymbol{x})_j = \sum_k x_k p_{j-k}$. In this case, $\boldsymbol{p}$ is called the Toeplitz vector.

Example: ETCS

Let $\lambda = ah_t/2h_x$. Then

$$u_{k,\ell+1} = \lambda u_{k-1,\ell} + u_{k,l} - \lambda u_{k+1,\ell}.$$

Is ETCS Toeplitz?

Is ETCS Toeplitz?

$$(P_h \boldsymbol{u}_{\ell+1})_j = u_{j,\ell+1} \stackrel{!}{=} \sum_k u_{k,\ell+1} p_{j-k}$$

$$p_{j-k} = \begin{cases} 1 & k = j, \\ 0 & \text{otherwise.} \end{cases} \qquad p_\ell = \delta_{0,\ell}.$$

$$(Q_h \boldsymbol{u}_\ell)_j = \lambda u_{k-1,\ell} + u_{k,l} - \lambda u_{k+1,\ell} \stackrel{!}{=} \sum_k u_{k,\ell} q_{j-k}$$

$$q_{j-k} = \begin{cases} \lambda & k = j-1, \\ 1 & k = j, \\ -\lambda & k = j+1, \\ 0 & \text{otherwise.} \end{cases} \qquad q_\ell = \begin{cases} \lambda & \ell = 1, \\ 1 & \ell = 0, \\ -\lambda & \ell = -1, \\ 0 & \text{otherwise.} \end{cases}$$

Both P_h and Q_h are Toeplitz.

Fourier Transforms of Toeplitz Operators (1/3)

$$y_j = \sum_k x_k p_{j-k}$$

$$\begin{aligned}
\hat{\boldsymbol{y}}(\theta) &= \sum_j \sum_k x_k p_{j-k} e^{-i\theta j} \\
&= \sum_j \sum_k \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta) e^{i\varphi k} d\varphi \right) p_{j-k} e^{-i\theta j} \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta) \sum_j \sum_k e^{i\varphi k} p_{j-k} e^{-i\theta j} d\theta \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta) \sum_j \left(\sum_k e^{i\varphi(k-j)} p_{j-k} \right) e^{i(\varphi-\theta)j} d\theta.
\end{aligned}$$

Fourier Transforms of Toeplitz Operators (2/3)

$$\hat{\boldsymbol{y}}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta) \sum_j \left(\sum_k e^{i\varphi(k-j)} p_{j-k} \right) e^{i(\varphi-\theta)j} d\theta.$$

Consider

$$\sum_k e^{i\varphi(k-j)} p_{j-k} = \sum_k e^{-i\varphi(j-k)} p_{j-k} \overset{\ell=j-k}{=} \hat{\boldsymbol{p}}(\varphi).$$

So

$$\hat{\boldsymbol{y}}(\theta) = \int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta)\hat{\boldsymbol{p}}(\varphi)\frac{1}{2\pi}\sum_j e^{i(\varphi-\theta)j} d\theta.$$

Fourier Transforms of Toeplitz Operators (3/3)

$$\hat{\boldsymbol{y}}(\theta) = \int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta)\hat{\boldsymbol{p}}(\varphi)\frac{1}{2\pi}\sum_j e^{i(\varphi-\theta)j}\,d\theta.$$

Define $w_j = (1/2\pi)e^{i\varphi j}$. Then $\hat{\boldsymbol{w}}(\theta) = \frac{1}{2\pi}\sum_k e^{i(\varphi-\theta)k}$. So

$$\hat{\boldsymbol{y}}(\theta) = \int_{-\pi}^{\pi} \hat{\boldsymbol{x}}(\theta)\hat{\boldsymbol{p}}(\varphi)\hat{\boldsymbol{w}}(\theta)\,d\theta.$$

To determine $\hat{\boldsymbol{w}}(\theta)$, consider

$$(1/2\pi)e^{i\varphi j} = w_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{\boldsymbol{w}}(\theta)e^{i\theta j}\,d\theta.$$

Observe that $\hat{\boldsymbol{w}}(\theta) = \delta(\varphi - \theta)$ would do the trick.
Therefore $\hat{\boldsymbol{y}}(\theta) = \hat{\boldsymbol{x}}(\theta)\hat{\boldsymbol{p}}(\theta)$.

Fourier Transforms of Inverse Toeplitz Operators

Fourier transform $P_h^{-1} Q_h \boldsymbol{y}$?

$$\frac{\hat{\boldsymbol{q}}(\theta)}{\hat{\boldsymbol{p}}(\theta)} \hat{\boldsymbol{y}}(\theta).$$

Bounding the Operator Norm

Bound $\left\|P_h^{-1}Q_h\right\|_2^2$ using Fourier:

$$\left\|P_h^{-1}Q_h\right\|_2^2 = \sup_{\boldsymbol{x}\neq 0} \frac{\left\|P_h^{-1}Q_h\boldsymbol{x}\right\|_2^2}{\|\boldsymbol{x}\|_2^2} = \sup_{\boldsymbol{x}\neq 0} \frac{\frac{h_x}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\hat{\boldsymbol{q}}(\theta)}{\hat{\boldsymbol{p}}(\theta)}\hat{\boldsymbol{x}}(\theta)\right|^2 d\theta}{\frac{h_x}{2\pi}\int_{-\pi}^{\pi}|\hat{\boldsymbol{x}}(\theta)|^2\, d\theta}$$

$$\leq \sup_{\boldsymbol{x}\neq 0} \frac{\max_{\varphi\in[-\pi,\pi]}\left|\frac{\hat{\boldsymbol{q}}(\varphi)}{\hat{\boldsymbol{p}}(\varphi)}\right|\int_{-\pi}^{\pi}|\hat{\boldsymbol{x}}(\theta)|^2\, d\theta}{\int_{-\pi}^{\pi}|\hat{\boldsymbol{x}}(\theta)|^2\, d\theta} = \max_{\varphi\in[-\pi,\pi]}\left|\frac{\hat{\boldsymbol{q}}(\varphi)}{\hat{\boldsymbol{p}}(\varphi)}\right|.$$

Similarly,

$$\left\|P_h^{-1}\right\|_2^2 \leq \max_{\varphi\in[-\pi,\pi]}|\hat{\boldsymbol{p}}(\varphi)|.$$

Is the upper bound attained?

If $\hat{\boldsymbol{x}}(\theta) = \delta(\theta-\varphi^*)$, where φ^* maximizes $|\hat{\boldsymbol{q}}(\theta)/\hat{\boldsymbol{p}}(\theta)|$, then yes. (So $x_k = (1/2\pi)e^{i\varphi^* k}$.)

von Neumann Stability

Two-level finite difference scheme

$$P_h \boldsymbol{v}_{\ell+1} = Q_h \boldsymbol{v}_\ell + h_t \boldsymbol{b}_\ell,$$

where P_h and Q_h are Toeplitz operators with vectors $\boldsymbol{p}$ and $\boldsymbol{q}$.

Definition (Symbol of a Two-Level Finite Difference Scheme)

Let

$$\hat{\boldsymbol{p}}(\theta) = \sum_k p_k e^{-i\varphi k}, \qquad \hat{\boldsymbol{q}}(\theta) = \sum_k q_k e^{-i\varphi k}.$$

Then the symbol of the two-level FD method is $s(\varphi) = \hat{\boldsymbol{q}}(\varphi)/\hat{\boldsymbol{p}}(\theta)$.

Definition (Von Neumann Stability)

If

$$\max_\varphi |s(\varphi)| \leq 1, \qquad \max_\varphi \left| \frac{1}{\hat{\boldsymbol{p}}(\varphi)} \right| \leq c$$

for some constant $c > 0$, we say the scheme is von Neumann stable.

Comparison with Lax-Richtmyer Stability

Need $\left\|(P_h^{-1}Q_h)^\ell P_h^{-1}\right\| \le c$.

Implied by von Neumann stability.

Why is bounding the symbol the most salient part?

If there doesn't exist a c so that $\left\|P_h^{-1}\right\| \le c$, then $\left\|P_h^{-1}Q_h\right\|$ often also encounters problems.

Main restriction of von Neumann stability?

- Only works on infinite/periodic grids.
- Have BCs? Analysis gets more difficult.

von Neumann Stability: ETBS (1/2)

ETBS: Let $\lambda = ah_t/h_x$. $u_{k,\ell+1} = \lambda u_{k-1,\ell} + (1-\lambda)u_{k,\ell}$.

$$P_h = I, \qquad Q_h = \text{tridiag}(\lambda, 1-\lambda, 0).$$

Auxiliary result: Fourier transform of $r_k = \delta_{k,j}$.

$$\hat{\boldsymbol{r}}(\varphi) = \sum_k r_k e^{-i\varphi k} = \sum_k \delta_{k,j} e^{-i\varphi k} = e^{-i\varphi j}.$$

Recall: $\boldsymbol{r}$ Toeplitz vector indices are 'flipped' compared to matrix entries $\to$ index sign flip

$$\hat{\boldsymbol{p}}(\varphi) = 1, \qquad \hat{\boldsymbol{q}}(\varphi) = \lambda e^{-i\varphi} + (1-\lambda) = 1 - \lambda(1 - e^{-i\varphi}).$$

$$\begin{aligned} |s(\varphi)|^2 &= \left|\frac{\hat{\boldsymbol{q}}(\varphi)}{\hat{\boldsymbol{p}}(\varphi)}\right|^2 = (1-\lambda(1-e^{-i\varphi}))(1-\lambda(1-e^{i\varphi})) \\ &= 1 + 2(\lambda - \lambda^2)(\cos\varphi - 1). \end{aligned}$$

von Neumann Stability: ETBS (2/2)

Found: $|s(\varphi)|^2 = 1 + 2(\lambda - \lambda^2)(\cos\varphi - 1)$.

Maximize: take derivative w.r.t. φ, set to 0:

$$\frac{d}{d\varphi}\left(1 + 2(\lambda - \lambda^2)(\cos\varphi - 1)\right) = -2(\lambda - \lambda^2)\sin\varphi = 0$$

if and only if $\varphi \in \mathbb{Z}\pi$.

For $m \in \mathbb{Z}$, $s(m\pi) = 1 + 2(\lambda - \lambda^2)((-1)^m - 1)$. For m even, $s(m\pi) = 1$.

For m odd, $s(m\pi) = 1 - 4(\lambda - \lambda^2) = (1 - 2\lambda)^2$.

Thus $|s(\varphi)|^2 \le 1$ if and only if

$$|1 - 2\lambda| \le 1 \quad \Leftrightarrow \quad 0 \le \lambda \le 1 \quad \Leftrightarrow \quad 0 \le h_t \le \frac{h_x}{a}.$$

Found: conditionally von Neumann stable with CFL as before.

von Neumann Stability: ETCS

Let $\lambda = ah_t/h_x$. Then

$$u_{k,\ell+1} = \frac{\lambda}{2}u_{k-1,\ell} + u_{k,\ell} - \frac{\lambda}{2}u_{k+1,\ell}.$$

$$P_h = I, \qquad Q_h = \text{tridiag}(\lambda/2, 1, -\lambda/2).$$

So $\hat{\boldsymbol{p}}(\varphi) = 1$, and

$$\hat{\boldsymbol{q}}(\varphi) = \frac{\lambda}{2}e^{-i\varphi} + 1 - \frac{\lambda}{2}e^{-i\varphi(-1)} = 1 - \lambda\sin(\varphi)i.$$

So

$$\max_{\varphi}|s(\varphi)|^2 = \max_{\varphi}\left|\frac{\hat{\boldsymbol{q}}(\varphi)}{\hat{\boldsymbol{p}}(\varphi)}\right|^2 = 1 + \lambda^2\sin(\varphi) \geq 1.$$

Not von Neumann stable $\Rightarrow$ not Lax-Richtmyer stable.

von Neumann Stability: Crank-Nicolson

Let $\lambda = ah_t/(4h_x)$

$$-\lambda u_{k-1,\ell+1} + u_{k,\ell+1} + \lambda u_{k+1,\ell+1} = \lambda u_{k-1,\ell} + u_{k,\ell} - \lambda u_{k+1,\ell}.$$

$$P_h = \text{tridiag}(-\lambda, 1, \lambda), \quad Q_h = \text{tridiag}(\lambda, 1, -\lambda).$$

$$\begin{aligned} \hat{\boldsymbol{p}}(\varphi) &= -\lambda e^{-i\varphi} + 1 + \lambda e^{i\varphi} = 1 + 2\lambda i \sin(\varphi), \\ \hat{\boldsymbol{q}}(\varphi) &= \lambda e^{-i\varphi} + 1 - \lambda e^{i\varphi} = 1 - 2\lambda i \sin(\varphi). \end{aligned}$$

$$|s(\varphi)|^2 = \frac{1 + 4\sin^2(\varphi)}{1 + 4\sin^2(\varphi)} = 1.$$

Crank-Nicolson is unconditionally von Neumann stable.

Outline

Studying Solutions of the PDE

Saw numerically: interesting dispersion/dissipation behavior.
Want: theoretical understanding.

Consider *linear*, *continuous* (not yet discrete) differential operators

$$
\begin{aligned}
L_1 u &= u_t + a u_x, \\
L_2 u &= u_t - D u_{xx} + a u_x \qquad \text{(D>0)} \\
L_3 u &= u_t + a u_x - \mu u_{xxx}.
\end{aligned}
$$

What could we use as a prototype solution?

A Prototype Solution of the PDE

Observation: all these operators are diagonalized by complex exponentials. Come up with a 'prototype complex exponetial solution'.

Let $z(x,t) = z_0 e^{i(kx - \omega t)}$.

What type of function is this?

- For k, ω real: traveling wave with speed $c = \omega/k$.
  $z(x - ct, 0) = z_0 e^{i(k(x-ct))} = z(x,t)$.
- For k imaginary: an evanescent wave in x.
- For $\operatorname{Im}\omega < 0$: a wave decaying in time.

Wave-like Solutions of the PDE

$$z(x,t) = z_0 e^{i(kx - \omega t)}$$

Observations in connection with L?

- $Lz = \lambda(\omega, k) z$.
- $z(x,t)$ is a solution iff $Lz = 0$ iff $\lambda(\omega, k) = 0$.

What is the dispersion relation?

The equation $\lambda(\omega, k) = 0$ is called the dispersion relation for the PDE L.

Picking Apart the Dispersion Relation

Consider $\omega(k) = \alpha(k) + i\beta(k)$. Rewrite the wave solution with this.

$$
\begin{aligned}
z(x,t) &= z_0 e^{i(kx-\omega t)} \\
&= z_0 e^{i(kx-\alpha(k)t-i\beta(k)t)} \\
&= z_0 e^{\beta(k)t} e^{i(kx-\alpha(k)t)}.
\end{aligned}
$$

How can we recognize dissipation?

If $\beta(k) < 0$, we call the PDE dissipative.

What is the phase speed? How can we recognize dispersion?

- The phase speed of $z(x,t)$ is $v_{\text{ph}} = \alpha(k)/k$.
- If v_{ph} is a constant ($\Leftrightarrow \alpha(k)$ is linear in k), all waves move at the same speed.

Dispersion Relation: Examples

In each case, find the dispersion relation and identify properties.

$L_1 u = u_t + a u_x$

- $\lambda(\omega, k) = i(ak - \omega) = 0$, i.e. $\omega = ak$.
- Neither dissipative nor dispersive.

$L_2 u = u_t - D u_{xx} + a u_x \ (D > 0)$

- $\lambda(\omega, k) = -i\omega + iak + Dk^2$, i.e. $\omega = ak - iDk^2$.
- Dissipative, but not dispersive.

$L_3 u = u_t + a u_x - \mu u_{xxx}$

- $\lambda(\omega, k) = -i\omega + iak + i\mu k^3$, i.e. $\omega = ak + \mu k^3$.
- Dispersive, but not dissipative.

Numerical Dissipation/Dispersion Analysis

Goal: Want discrete finite difference scheme to match dissipation/dispersion behavior of continuous PDE.

Define a discrete wave-like function:

$$z_{j,\ell} = z_0 e^{i(kjh_x - \omega \ell h_t)}$$

We want $\boldsymbol{z}$ to solve $P_h \boldsymbol{z}_{\ell+1} = Q_h \boldsymbol{z}_{\ell}$. How can we connect the operators to the wave solution?

P_h and Q_h consist of Toeplitz operators.

Toeplitz and Waves

$$z_{j,\ell} = z_0 e^{i(kjh_x - \omega \ell h_t)}.$$

Theorem (Waves Diagonalize Toeplitz Operators)

Let T be a Toeplitz operator. Then $T\boldsymbol{z}_\ell = \lambda(k)\boldsymbol{z}_\ell = \hat{\boldsymbol{t}}(kh_x)\boldsymbol{z}_\ell$.

$$\begin{aligned}
(T\boldsymbol{z}_\ell)_j &= \sum_m z_{m,\ell} t_{j-m} = \sum_m z_0 e^{i(kmh_x - \omega \ell h_t)} t_{j-m} \\
&= \sum_m z_0 e^{i(k(m-j)h_x)} e^{i(kjh_x - \omega \ell h_t)} t_{j-m} \\
&= \left(\sum_{m'} e^{-ikm'h_x} t_{m'} \right) z_0 e^{i(kjh_x - \omega \ell h_t)}.
\end{aligned}$$

$$\Rightarrow \lambda(k) = \sum_m e^{-ikmh_x} t_m = \hat{\boldsymbol{t}}(kh_x).$$

Waves and Two-Level Schemes

Since P_h and Q_h are Toeplitz, we must have

$$P_h \boldsymbol{z}_{\ell+1} = \lambda_P(k)\boldsymbol{z}_{\ell+1}, \qquad Q_h \boldsymbol{z}_\ell = \lambda_Q(k)\boldsymbol{z}_\ell.$$

What does that mean?

$$\begin{aligned}
\lambda_P(k)\boldsymbol{z}_{\ell+1} &= \lambda_Q(k)\boldsymbol{z}_\ell \\
\lambda_P(k) z_0 e^{i(kjh_x - \omega(\ell+1)h_t)} &= \lambda_Q(k) z_0 e^{i(kjh_x - \omega\ell h_t)} \\
e^{-i\omega h_t} &= \frac{\lambda_Q(k)}{\lambda_P(k)} = \frac{\hat{\boldsymbol{q}}(kh_x)}{\hat{\boldsymbol{p}}(kh_x)} = s(kh_x),
\end{aligned}$$

which is the symbol of of the finite difference method.

Seen before?

Used in von Neumann stability analysis.

Discrete Dispersion Relation (1/2)

So $\boldsymbol{z}_\ell$ is a solution of the finite difference scheme if $\omega = \omega(kh_x)$ satisfies

$$e^{-i\omega(\kappa)h_t} = s(\kappa),$$

where we let $\kappa = kh_x$. Interpret κ.

A number proportional to the number of wavelengths per point.

Let $s(\kappa) = |s(\kappa)|\, e^{i\varphi(\kappa)} = e^{\log|s(\kappa)| + i\varphi(\kappa)}$. $\omega(\kappa)$?

$$\omega(\kappa) = \frac{-\varphi(\kappa) + i \log|s(\kappa)|}{h_t}.$$

Discrete Dispersion Relation (2/2)

$$\omega(\kappa) = \frac{-\varphi(\kappa) + i\log|s(\kappa)|}{h_t}.$$

Plug that into the wave-like solution:

$$\begin{aligned} z_{j,\ell} &= z_0 e^{i(kjh_x - \omega \ell h_t)} \\ &= z_0 e^{i\left(kjh_x - \frac{-\varphi(\kappa)+i\log|s(\kappa)|}{h_t}\ell h_t\right)} \\ &= z_0 e^{\log|s(\kappa)|\ell} e^{ik\left(jh_x - \frac{-\varphi(\kappa)}{kh_t}\ell h_t\right)} \end{aligned}$$

Criterion for stability?

$|s(\kappa)| \leq 1$ (as before)

Numerical Dispersion/Dissipation

Finite difference scheme $P_h \boldsymbol{u}_{\ell+1} = Q_h \boldsymbol{u}_\ell$ with symbol $s(k)$.

$$z_{j,\ell} = z_0 e^{\log|s(\kappa)|\ell} e^{ik\left(jh_x - \frac{-\varphi(\kappa)}{kh_t}\ell h_t\right)}$$

When is the scheme dissipative?

If $|s(kh_x)| < 1$, the scheme is called dissipative. Dissipation occurs exponentially in time, with factor $s(kh_x)$.

What is the phase speed?

The scheme has phase speed $v_{\mathrm{ph}} = \frac{-\varphi(kh_x)}{kh_t}$.

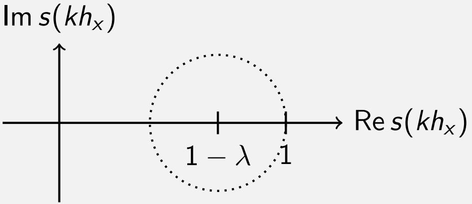
Dispersion?

If v_{ph} is independent of k, all waves move with the same speed. If not, the scheme is called dispersive.

Dispersion/Dissipation Analysis of ETBS

Let $\lambda = ah_t/h_x$. Shown earlier: $s(kh_x) = 1 - \lambda(1 - e^{-ikh_x})$.

$|s(kh_x)| = 1$ holds on a circle:



For small λ, the circle moves towards $s(kh_x) = 1$, implying lower dissipation per step.

Overall, we obtain

$$e^{-i\omega(\kappa)h_t} = 1 - \lambda(1 - e^{-ikh_x}).$$

Dispersion/Dissipation Analysis of ETBS: Fine Grid

$$e^{-i\omega(\kappa)h_t} = 1 - \lambda(1 - e^{-ikh_x})$$

If kh_x is small, $e^{-ikh_x} \approx 1 - ikh$, so that

$$s(kh_x) \approx (1-\lambda) + \lambda(1 - ikh_x) = 1 - i\lambda kh_x.$$

For small $\omega(kh_x)$, approximate $e^{-i\omega(kh_x)h_t} = 1 - i\omega(kh_x)h_t$.
Setting the two (approximately) equal yields

$$1 - i\omega(kh_x)h_t \approx 1 - i\lambda kh_x \qquad \Rightarrow \qquad \omega(kh_x)h_t \approx \lambda kh_x = \frac{ah_t}{h_x}kh_x,$$

i.e. $\omega(kh_x) \approx ak$, or $v_{\text{ph}} \approx (-ak)/(kh_t) = -a/h_t$, which is independent of k. Thus we expect little dispersion for waves with low number of wavelengths per point.

Dispersion/Dissipation: Demo

- **Demo:** Experimenting with Dispersion and Dissipation
- **Demo:** Dispersion and Dissipation

Outline

Heat Equation

Heat equation ($D > 0$):

$$\begin{aligned} u_t &= D u_{xx}, \qquad (x,t) \in \mathbb{R} \times (0,\infty), \\ u(x,0) &= g(x) \qquad x \in \mathbb{R}. \end{aligned}$$

Fundamental solution ($g(x) = \delta(x)$):

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}.$$

Why is this a weird model?

Infinite speed of propagation of information

Schemes for the Heat Equation

Cook up some schemes for the heat equation.

Explicit Euler:

$$\frac{u_{k,\ell+1} - u_{k,\ell}}{h_t} - D\frac{u_{k+1,\ell} - 2u_{k,\ell} + u_{k-1,\ell}}{h_x^2} = 0$$

Implicit Euler:

$$\frac{u_{k,\ell+1} - u_{k,\ell}}{h_t} - D\frac{u_{k+1,\ell+1} - 2u_{k,\ell+1} + u_{k-1,\ell+1}}{h_x^2} = 0$$

Von Neumann Analysis of Explicit Euler for Heat (1/2)

Let $\lambda = Dh_t/h_x^2$.

$$u_{k,\ell+1} = u_{k,\ell} + \lambda(u_{k+1,\ell} - 2u_{k,\ell} + u_{k-1,\ell}).$$

$$P_h = I, \qquad Q_h = \text{tridiag}(\lambda, 1-2\lambda, \lambda).$$

Thus

$$\begin{aligned} \hat{\boldsymbol{p}}(\varphi) &= 1, \\ \hat{\boldsymbol{q}}(\varphi) &= \lambda e^{-i\varphi} + (1-2\lambda) + \lambda e^{i\varphi} = 1 - 2\lambda + 2\lambda\cos(\varphi). \end{aligned}$$

We want $|s(\varphi)| \leq 1$, thus we need

$$\begin{aligned} -1 \leq & \quad 1 + 2\lambda(\cos(\varphi) - 1) & \leq 1 \\ \Leftrightarrow \quad -2 \leq & \quad 2\lambda(\cos(\varphi) - 1) & \leq 0. \end{aligned}$$

Von Neumann Analysis of Explicit Euler for Heat (2/2)

$$-2 \le 2\lambda(\cos(\varphi) - 1) \le 0.$$

Since $|\cos(\varphi)| \le 1$, also $-2 \le \cos(\varphi) - 1 \le 0$. For the lower bound,

$$-2 \le -4\lambda \quad \Leftrightarrow \quad \frac{1}{2} \ge \frac{D h_t}{h_x^2} \quad \Leftrightarrow \quad h_t \le \frac{h_x^2}{2D}.$$

Observe $h_t = O(h_x^2)$, which is often prohibitively small.

Comment on the stability region found regarding speeds of propagation.

- Saw: heat equation has infinite speed of information propagation
- Explicit Euler has finite speed of information propagation (how fast?)

Von Neumann Analysis of Implicit Euler for Heat

Let $\lambda = Dh_t/h_x^2$.

$$u_{k,\ell+1} - \lambda(u_{k+1,\ell+1} - 2u_{k,\ell+1} + u_{k-1,\ell+1}) = u_{k,\ell}$$

$$P_h = \operatorname{tridiag}(-\lambda, 1+2\lambda, -\lambda), \qquad Q_h = I.$$

$$\hat{\boldsymbol{p}}(\varphi) = 1 + 2\lambda(1-\cos(\varphi)), \quad \hat{\boldsymbol{q}}(\varphi) = 1.$$

To obtain $|s(\varphi)| \le 1$, consider $1 \le |1 + 2\lambda(1-\cos(\varphi))|$, which is always true.

Does the type of system we need to solve for implicit+parabolic correspond to another PDE?

- Yes, elliptic.
- Focus on solving those later.

Outline

Outline

Conservation Laws: Recap

$$u_t + f(u)_x = 0,$$

where u is a function of x and $t \in \mathbb{R}_0^+$.

Rewrite in integral form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b u(x,t)\mathrm{d}x + f(u(b,t)) - f(u(a,t)) = 0 \qquad \text{for any } a,\ b.$$

Recall: Characteristic Curve: a function $x(t)$ so that $u(x(t),t) = u(x_0, 0)$.

$$\begin{cases} \frac{\mathrm{d}x(t)}{\mathrm{d}t} = f'(u(x(t),t)), \\ x(0) = x_0. \end{cases}$$

What assumption underlies all this?

Smooth Solution.

Burger's Equation

Consider Burgers' Equation:

$$\begin{cases} u_t + \left(\frac{u^2}{2}\right)_x = 0, \\ u(x,0) = g(x) = \sin(x). \end{cases}$$

Interpret Burger's equation.

$f(u) = u^2/2$. So $f'(u) = u$.
Characteristic speed is given by 'how much stuff there is'/'the density'

Consider the characteristics at $\pi/2$ and $3\pi/2$.

$f(u) = u^2/2$. So $f'(u) = u$.

- $x = \pi/2$: $f'(\sin x) = 1$.
- $x = 3\pi/2$: $f'(\sin x) = -1$.

They intersect!

Weak Solutions

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b u(x,t)\mathrm{d}x = f(u(a,t)) - f(u(b,t))$$

Define a weak solution:

- If u satisfies the integral form for almost all (a, b) then u is called a weak solution. (physically meaningful, correct)
- If for any $\varphi \in C_0^1(\mathbb{R} \times [0, \infty))$ (compact support),

$$-\int_0^\infty \int_{-\infty}^\infty (u\varphi_t + f(u)\varphi_x)\mathrm{d}x\mathrm{d}t - \int_{-\infty}^\infty u^0(x)\varphi(x,0)\mathrm{d}x = 0,$$

  then in u is called a weak solution. (more meaningful mathematically)

Turns out: equivalent. (not shown)

Rankine-Hugoniot Condition (1/2)

Consider: Two C^1 segments separated by a curve $x(t)$ with no regularity.

$$(d/dt)\Big(\underbrace{\int_a^{x(t)} u(x,t)dx}_{G_a(x(t),t):=} + \underbrace{\int_{x(t)}^{b} u(x,t)dx}_{G_b(x(t),t):=}\Big) + f(u(b,t)) - f(u(a,t)) = 0.$$

$$\begin{aligned}
\frac{d}{dt}G_a(x(t),t) &= \frac{\partial G_a(x(t),t)}{\partial x}\cdot\frac{dx(t)}{dt} + \frac{\partial G_a}{\partial t}\\
&= u(x(t),t)x'(t) + \int_a^{x(t)} u_t(x,t)dx\\
&= u(x(t),t)x'(t) - \int_a^{x(t)} f(u)_x(x,t)dx\\
&= u(x(t),t)x'(t) - \big(f(u(x(t),t)) - f(u(a,t))\big),
\end{aligned}$$

and $dG_b(x(t),t)/dt$ analogously.

Rankine-Hugoniot Condition (2/2)

$$(d/dt)G_a(x(t),t) = u(x(t),t)x'(t) - (f(u(x(t),t)) - f(u(a,t))).$$

Discontinuity at $u(x(t),t)$: $(d/dt)G_a$ doesn't exist. One-sided limits:

$$\left[\frac{dG_a(x(t),t)}{t}\right]^- = u^-x'(t) - (f(u^-) - f(u(a,t))),$$

$$\left[\frac{dG_b(x(t),t)}{t}\right]^+ = -u^+x'(t) - (f(u(b,t)) - f(u^+)).$$

Adopted shorthand: $u^- := u(x(t)^-, t)$, $\qquad u^+ := u(x(t)^+, t)$.
Plug into integral form: $u^-x'(t) - f(u^-) - u^+x'(t) + f(u^+) = 0$.

$$x'(t) = \frac{f(u^+) - f(u^-)}{u^+ - u^-}.$$

This is the called the Rankine-Hugoniot Condition.

Rankine-Hugoniot and Weak Solutions

Theorem (Rankine-Hugoniot and Weak Solutions)

If u is piecewise C^1 and is discontinuous only along isoated curves, and if u satisfies the PDE when it is C^1, and the Rankine-Hugoniot condition holds along all discontinuous curves, then u is a weak solution of the conservation law.

Riemann Problems: Example 1

Consider the following Riemann problem:

$$u_t + \left(\frac{u^2}{2}\right)_x = 0,$$

$$u(x,0) = \begin{cases} 1 & x < 0, \\ -1 & x \geq 0. \end{cases}$$

The IC is just propagated in time (at "speed 0") to form a weak solution (a shock).

Riemann Problems: Example 2

$$u_t + \left(\frac{u^2}{2}\right)_x = 0,$$

$$u(x,0) = \begin{cases} -1 & x < 0, \\ 1 & x \geq 0. \end{cases}$$

(IC sign flip compared to previous slide)

The propagated ICs also form a weak solution. But consider

$$u(x,t) = \begin{cases} -1 & x \leq -t, \\ x/t & -t < x < t, \\ 1 & x > t. \end{cases}$$

This is also a weak solution (a rarefaction wave).
Conclusion: Our current notion of weak solution is *too* weak.

Bad Shocks and Good Shocks

In the shock version of the 'ambiguous' Riemann problem, where do the characteristics go?

- Out of the shock.
- In the first example, the shock is self-steepening.
- In the second example, it is not.

Comment on the stability of that situation.

Smearing out the initial profile or adding viscosity would wash out the solution into a rarefaction fan.

Ad-Hoc Idea: Ban Bad Shocks

Recall: what is $f'(u)$?

Characteristic speed.

Devise a way to ban unstable shocks.

A discontinuity propagating with speed s (cf. Rankine-Hugoniot) satsifes the entropy condition if

$$f'(u^-) > s > f'(u^+).$$

If f is convex, f' is monotonically non-decreasing, and the Rankine-Hugoniot speed automatically falls between $f'(u^-)$ and $f'(u^+)$. So for convex f, $f'(u^-) > f'(u^+)$ is sufficient (and implies $u^- > u^+$ by convexity).

Vanishing Viscosity Solutions

Goal: neither uniqueness nor existence poses a problem.

How?

Consider adding an artificial viscosity:

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{x,x}^\varepsilon \qquad \text{with small } \varepsilon > 0.$$

By 'washing out' the solution, the viscous term increases smoothness, and, we hope, restores uniqueness.

Then we would wish to define an vanishing viscosity weak solution as

$$\lim_{\varepsilon \to 0} u^\varepsilon(x,t) = u(x,t)$$

in some norm.

Entropy-Flux Pairs

What are features of (physical) entropy?

- Constant along particle paths in smooth flow
- Jumps to higher values across a shock

Definition (Entropy/Entropy Flux)

An entropy $\eta(u)$ and an entropy flux $\psi(u)$ are functions so that η is convex and

$$\eta(u)_t + \psi(u)_x = 0$$

for smooth solutions of the conservation law.

Finding Entropy-Flux Pairs

$\eta(u)_t + \psi(u)_x = 0$. Find conditions on η and ψ.

For smooth u, the chain rule gives $\eta'(u)u_t + \psi'(u)u_x = 0$. Similarly, we can rewrite the conservation law:

$$\begin{aligned} u_t + f'(u)u_x &= 0 \\ \Leftrightarrow \quad \eta'(u)u_t + \eta'(u)f'(u)u_x &= 0. \end{aligned}$$

This gives us $\psi'(u) = \eta'(u)f'(u)$.
Lots of solutions for scalar conservation laws. For systems and in multiple dimensions: may have no solutions.

Come up with an entropy-flux pair for Burgers.

$f(u) = u^2/2$. If we take $\eta(u) = u^2$, then $\psi'(u) = 2u \cdot u$, i.e. $\psi(u) = 2u^3/3$.

Back to Vanishing Viscosity (1/2)

$$u_t + f(u)_x = \varepsilon u_{xx}$$

What's the evolution equation for the entropy?

Note: Viscosity solutions are always smooth. Allowed to do derivative gymnastics.

$$\begin{aligned} \eta'(u)u_t + \eta'(u)f'(u)u_x &= \varepsilon\eta'(u)u_{xx} \\ \Leftrightarrow \quad \eta(u)_t + \psi(u)_x &= \varepsilon(\eta'(u)u_x)_x - \varepsilon\eta''(u)u_x^2. \end{aligned}$$

Back to Vanishing Viscosity (2/2)

$$\eta(u)_t + \psi(u)_x = \varepsilon(\eta'(u)u_x)_x - \varepsilon\eta''(u)u_x^2.$$

Integrate this over $[x_1, x_2] \times [t_1, t_2]$.

$$\begin{aligned} & \int_{t_1}^{t_2}\int_{x_1}^{x_2} \eta(u)_t + \psi(u)_x dxdt \\ = \quad & \varepsilon \int_{t_1}^{t_2} [\eta'(u(x_2,t))u_x(x_2,t) - \eta'(u(x_1,t))u_x(x_1,t)]dt \\ & -\varepsilon \int_{t_1}^{t_2}\int_{x_1}^{x_2} \underbrace{\eta''(u)u_x^2}_{\geq 0} dxdt. \end{aligned}$$

As $\varepsilon \to 0$, the first term goes to zero. The second term involves an integral over the square of the derivative of a steepening u (as $\varepsilon \to 0$), and so will not vanish. Accordingly, $\eta(u)_t + \psi(u)_x \leq 0$ weakly.

Entropy Solution

Definition (Entropy solution)

The function $u(x, t)$ is the entropy solution of the conservation law if for all convex entropy functions and corresponding entropy fluxes, the inequality

$$\eta(u)_t + \psi(u)_x \leq 0$$

is satisfied in the weak sense.

Conservation of Entropy?

What can you say about conservation of entropy in time?

$$
\begin{aligned}
0 &\geq \int_{t_1}^{t_2} \int_{x_1}^{x_2} \eta(u)_t + \psi(u)_x dx dt \\
&= \left[\int_{x_1}^{x_2} \eta(u(x,t)) dx \right]_{t_1}^{t_2} + \left[\int_{t_1}^{t_2} \psi(u(x,t)) dt \right]_{x_1}^{x_2},
\end{aligned}
$$

so that

$$
\int_{x_1}^{x_2} \eta(u(x,t_2)) dx \leq \int_{x_1}^{x_2} \eta(u(x,t_1)) dx - \underbrace{\left[\int_{t_1}^{t_2} \psi(u(x,t)) dt \right]_{x_1}^{x_2}}_{\text{Outflow/Inflow}}
$$

If u is compactly supported, then we can choose x_1 and x_2 on either side of u's support and obtain that entropy can only decrease. (Physically, entropy only increases. Could have chosen concave for that.)

Total Variation

$$\mathrm{TV}(u) = \limsup_{\varepsilon \to 0} \frac{1}{\varepsilon} \int |u(x+\varepsilon) - u(x)| \, dx.$$

Simpler form if u is differentiable?

$$\mathrm{TV}(u) = \int \left| u'(x) \right| dx$$

Hiking analog?

Elevation change

Total Variation and Conservation Laws

Theorem (Total Variation is Bounded [Dafermos 2016, Thm. 6.2.6])

Let u be a solution to a conservation law with $f''(u) \geq 0$. Then:

$$\mathrm{TV}(u(t + \Delta t, \cdot)) \leq \mathrm{TV}(u(t, \cdot)) \quad \textit{for } \Delta t \geq 0.$$

- For smooth solutions (and non-crossing characteristics), all function values live $\Rightarrow$ TV stays unchanged.
- For solutions with shocks, local minima and maxima may disappear into the shock $\Rightarrow$ TV decreases.

Theorem (L^1 contraction [Dafermos 2016, Thm. 6.3.2])

Let u, v be viscosity solutions of the conservation law. Then

$$\|u(t + \Delta, \cdot) - v(t + \Delta t, \cdot)\|_{L^1(\mathbb{R})} \leq \|u(t, \cdot) - v(t, \cdot)\|_{L^1(\mathbb{R})} \quad \textit{for } \Delta t \geq 0$$

Outline

Finite Difference for Conservation Laws? (1/2)

$$\begin{cases} u_t + \left(\frac{u}{2}\right)^2_x = 0 \\ u(x,0) = \begin{cases} 1 & x < 0, \\ 0 & x \geq 0. \end{cases} \end{cases}$$

Entropy Solution?

$$u(x,t) = \begin{cases} 1 & x \leq \frac{1}{2}t, \\ 0 & x > \frac{1}{2}t. \end{cases}$$

Rewrite the PDE to 'match' the form of advection $u_t + au_x = 0$:

$$u_t + uu_x = 0.$$

Equivalent?

Finite Difference for Conservation Laws? (2/2)

Recall the *upwind scheme* for $u_t + au_x = 0$:

$$u_{j,\ell+1} = u_{j,\ell} - a \cdot \frac{\Delta t}{\Delta x}(u_{j,\ell} - u_{j-1,\ell}).$$

Write the upwind FD scheme for $u_t + uu_x = 0$:

$$u_{j,\ell+1} = u_{j,\ell} - \frac{\Delta t}{\Delta x} u_{j,\ell}(u_{j,\ell} - u_{j-1,\ell}).$$

- For $j \neq 0$, $u_{j,0} - u_{j-1,0} = 0$
- For $j = 0$, $u_{j,0} = 0$.

Altogether,

$$u_{j,\ell+1} = u_{j,\ell}.$$

Bad.

Schemes in Conservation Form

Definition (Conservative Scheme)

A conservation law scheme is called conservative iff it can be written as

$$u_{j,\ell+1} = u_{j,\ell} - \frac{\Delta t}{\Delta x}[f^*_{j+1/2}(\boldsymbol{u}_\ell) - f^*_{j-1/2}(\boldsymbol{u}_\ell)],$$

where $f^*\dots$

- is Lipschitz continuous,
- satisfies $f^*(u, \cdots, u) = f(u)$ (consistency).

Theorem (Lax-Wendroff)

If the solution $\{u_{j,\ell}\}$ to a conservative scheme converges (as $\Delta t, \Delta x \to 0$) boundedly almost everywhere to a function $u(x,t)$, then u is a weak

Lax-Wendroff Theorem: Proof

Summation by parts: With $\Delta^+ a_k = a_{k+1} - a_k$ and $\Delta^- a_k = a_k - a_{k-1}$:

$$\sum_{k=1}^{N} a_k(\Delta^- \varphi_k) + \sum_{k=1}^{N} \varphi_k(\Delta^+ a_k) = -a_1\varphi_0 + \varphi_N a_{N+1}.$$

Let $\varphi_{j,\ell} = \varphi(x_j, t_\ell)$ for $\varphi \in C_0^1$ (compact support). Then

$$0 = \sum_{\ell=1}^{\infty} \sum_{j} \left(\frac{\Delta_2^+ u_{j,\ell}}{h_t} + \frac{\Delta^+ f_{j-1/2}^*}{h_x} \right) \varphi_{j,\ell} h_x h_t$$

$$= -\sum_{\ell=1}^{\infty} \sum_{j} \left(\frac{\Delta_2^- \varphi_{j,\ell}}{h_t} u_{j,\ell} + \frac{\Delta_1^- \varphi_{j,\ell}}{h_x} f_{j-1/2}^* \right) h_x h_t - \sum_j u_{j,1}\phi_{j,0} h_x$$

$$\underset{f^*(u,u)=u}{\overset{\text{DCT}}{\rightarrow}} -\int_0^\infty \int_{-\infty}^\infty (\varphi_t u + \varphi_x f(u))\,\mathrm{d}x\mathrm{d}t - \int_{-\infty}^\infty u(x,0)\phi(x,0)\,\mathrm{d}x = 0.$$

Finite Volume Schemes

Finite volume: Idea?

- Consider the solution constant in each cell: $\bar{u}_j$
- $\bar{u}_j$ is the cell average of cell I_j:

$$\bar{u}_j = (1/h_x) \int_{x_{j-1/2}}^{x_{j+1/2}} u(x) dx$$

- Choose h_x, h_t so that $\max |f'(u)| h_t < h_x$.

Then in a sequence of cells (A, B, C, D, E), the solution in cell C in the next timestep is not influenced at all by the solution in cells A and E.

Idea: Solve Riemann problem at each cell interface.

Developing Finite Volume

$$\int_{t_\ell}^{t^{\ell+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} (u_t + f(u)_x)\mathrm{d}x\mathrm{d}t = 0$$

$$\begin{aligned}
\frac{1}{h_x}\int_{x_{j-1/2}}^{x_{j+1/2}} u^{\ell+1}\mathrm{d}x - \frac{1}{h_x}\int_{x_{j-1/2}}^{x_{j+1/2}} u^{\ell}\mathrm{d}x & \\
+\frac{1}{h_x}\int_{t_\ell}^{t_{\ell+1}} f(u_{j+1/2})\mathrm{d}t - \frac{1}{h_x}\int_{t_\ell}^{t_{\ell+1}} f(u_{j-1/2})\mathrm{d}t &= 0 \\
\Leftrightarrow \quad \bar{u}_{j,\ell+1} - \bar{u}_{j,\ell} & \\
+\frac{1}{h_x}\int_{t_\ell}^{t_{\ell+1}} f(u_{j+1/2})\mathrm{d}t - \frac{1}{h_x}\int_{t_\ell}^{t_{\ell+1}} f(u_{j-1/2})\mathrm{d}t &= 0.
\end{aligned}$$

Flux Integrals?

$$\frac{1}{h_x}\int_{t_\ell}^{t_{\ell+1}} f(u_{j+1/2})\mathrm{d}t?$$

The substitution

$$\bar{x} = ax, \qquad \bar{t} = at.$$

leaves the conservation law and the Riemann ICs invariant.
$\Rightarrow$ The Riemann solution must be self-similar under scaling.

Thus: the Riemann solution $u(x,t)$ can be viewed as function of only one variable $\xi = x/t$.

Thus u is constant along $x = x_{j\pm 1/2}$, so that

$$\frac{1}{h_x}\int_{t_\ell}^{t_{\ell+1}} f(u_{j+1/2})\mathrm{d}t = \frac{h_t}{h_x} f(u_{j+1/2}).$$

The Godunov Scheme

Altogether:

$$\bar{u}_{j,\ell+1} = \bar{u}_{j,\ell} - \frac{h_t}{h_x}(f(u_{j+1/2,\ell}) - f(u_{j-1/2,\ell})).$$

Overall algorithm?

- Reconstruct $u^-_{j\pm1/2,\ell}$ and $u^+_{j\pm1/2,\ell}$
- Evolve the Riemann problem at $x_{j\pm1/2}$:
  Numerical flux / Riemann solver: $f^*(u^-_{j\pm1/2,\ell}, u^+_{j\pm1/2,\ell})$
- Average the Riemann solutions to obtain $\bar{u}_{j,\ell+1}$

Heuristic time step restriction?

Will run into problems if wave from one cell interface interacts with other interface: $h_t \le h_x / \max_j |f'(u_j)|$

Riemann Problem

$$\begin{cases} u_t + f(u)_x = 0, \\ u(x,0) = \begin{cases} u_l & x < 0, \\ u_r & x \geq 0 \end{cases} \end{cases}$$

Exact solution in the Burgers case?

$$u(x,t) = \begin{cases} \begin{cases} u_l & x < st, \\ u_r & x \geq st, \end{cases} & u_l \geq u_r, \\ \begin{cases} u_l & x < u_l t, \\ x/t & u_l t \leq x < u_r t, \\ u_r & x \geq u_r t, \end{cases} & u_l < u_r, \end{cases}$$

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l} = \frac{\frac{1}{2}[u_r^2 - u_l^2]}{u_r - u_l} = \frac{1}{2}(u_l + u_r).$$

Why is the rarefaction part independent of u_l and u_r?

Rieamnn Solver for a General Conservation Law

To complete the scheme: Need $f^*(u^-, u^+)$. For Burgers: already known. For a general (convex/concave-f) conservation law?

Assume $f''(u) > 0$.
Let u_s such that $f'(u_s) = 0$ (called the stagnation state: why?)

$$f^*(u^-, u^+) = \begin{cases} f(u^-) & \text{if shock with } s > 0, \\ f(u^+) & \text{if shock with } s \leq 0, \\ f(u^-) & \text{if rarefaction with } f'(u^-) \geq 0, \\ f(u^+) & \text{if rarefaction with } f'(u^+) \leq 0, \\ f(u_s) & \text{if rarefaction with } f'(u^-) \leq 0 \leq f'(u^+). \end{cases}$$

Equivalent to

$$f^*(u^-, u^+) = \begin{cases} \max_{u^+ \leq u \leq u^-} f(u) & \text{if } u^- > u^+, \\ \min_{u^- \leq u \leq u^+} f(u) & \text{if } u^- \leq u^+. \end{cases}$$

More Riemann Solvers

Downside of Godunov Riemann solver?

Not easy/efficient to implement in general. Want simpler Riemann solvers.

Back to Advection

Consider only $f(u) = au$ for now. Riemann solver inspiration from FD?

For $a \geq 0$, want ETBS:

$$\begin{aligned} 0 &= \frac{u_{j,\ell+1} - u_{j,\ell}}{h_t} + a\frac{u_{j,\ell} - u_{j-1,\ell}}{h_x} \\ &= \frac{u_{j,\ell+1} - u_{j,\ell}}{h_t} + \frac{f(u_{j,\ell}) - f(u_{j-1,\ell})}{h_x} \\ &= \frac{u_{j,\ell+1} - u_{j,\ell}}{h_t} + \frac{f^*(u_{j,\ell}, u_{j+1,\ell}) - f^*(u_{j-1,\ell}, u_{j,\ell})}{h_x}. \end{aligned}$$

Clearly equivalent to a finite volume scheme! Upwind numerical flux?

$$f^*(u^-, u^+) = \begin{cases} au^- & a \geq 0 \\ au^+ & a < 0 \end{cases} = \frac{au^- + au^+}{2} - \frac{|a|}{2}(u^+ - u^-).$$

Side Note: First Order Upwind, Rewritten

$$\frac{u_{j,\ell+1} - u_{j,\ell}}{h_t} + \frac{f^*(u_{j,\ell}, u_{j+1,\ell}) - f^*(u_{j-1,\ell}, u_{j,\ell})}{h_x}$$

with

$$f^*(u^-, u^+) = \frac{au^- + au^+}{2} - \frac{|a|}{2}(u^+ - u^-).$$

$$\frac{u_{j,\ell+1} - u_{j,\ell}}{h_t} + a\frac{u_{j+1,\ell} - u_{j-1,\ell}}{2h_x} = \frac{|a|\,h_x}{2} \cdot \frac{u_{j+1,\ell} - 2u_{j,\ell} + u_{j-1,\ell}}{h_x^2},$$

i.e. it is equivalent to ETCS (unstable!) with a second-order discretization of ∂_x^2, i.e. a dissipation, with a coefficient that vanishes as $h_x \to 0$.

Lax-Friedrichs

Generalize linear upwind flux for a nonlinear conservation law:

$$f^*(u^-, u^+) = \frac{au^- + au^+}{2} - \frac{|a|}{2}(u^+ - u^-).$$

$$f^*(u^-, u^+) = \frac{f(u^-) + f(u^+)}{2} - \frac{\alpha}{2}(u^+ - u^-).$$

Choice of α (consistent with linear)? Idea: $\alpha = |f'((u^- + u^+)/2)|$
Unfortunately: may converge to a weak solution that violates the entropy condition (not shown). Better:

$$\alpha = \max\left(\left|f'(u^-)\right|, \left|f'(u^+)\right|\right).$$

Called local Lax-Friedrichs. Global variant (with global max) also OK.

Demo: Finite Volume Burgers (Part I)

Outline

Improving Accuracy

Consider our existing discrete FV formulation:

$$\bar{u}_{j,\ell+1} = \bar{u}_{j,\ell} - \frac{h_t}{h_x}(f(u_{j+1/2,\ell}) - f(u_{j-1/2,\ell})).$$

What obstacles exist to increasing the order of accuracy?

- Temporal Accuracy
- Spatial Accuracy
- Nonsmoothness (in both space and time)

What order of accuracy can we expect?

- Near shocks: no convergence in L^∞, first-order in L^2.
- Elsewhere: hopefully, as high as we would like

Improving the Order of Accuracy

Improve temporal accuracy.

Rewrite FV using the method of lines:

$$\frac{d\bar{u}_j(t)}{dt} + \frac{f^*(u^-_{j+1/2}(t), u^+_{j+1/2}(t)) - f^*(u^-_{j-1/2}(t), u^+_{j-1/2}(t))}{h_x} = 0.$$

What's the obstacle to higher spatial accuracy?

Letting $u^-_{j+1/2} = \bar{u}_j = u^+_{j-1/2}$.

How can we improve the accuracy of that approximation?

Include more cells in the reconstruction of the state $u^\pm_{j+1/2}$.

Increasing Spatial Accuracy

Temporary Assumptions:

- $f'(u) \geq 0$
- $f^*_{j+1/2} = f(\bar{u}_j)$ (e.g. Godunov in this situation)

Reconstruct $u_{j+1/2}$ using $\{\bar{u}_{j-1}, \bar{u}_j, \bar{u}_{j+1}\}$. Accuracy? Names?

$$\begin{aligned} u^{(1)}_{j+1/2} &= \frac{1}{2}(\bar{u}_j + \bar{u}_{j+1}), && \text{(2nd order central)} \\ u^{(2)}_{j+1/2} &= \frac{3}{2}\bar{u}_j - \frac{1}{2}\bar{u}_{j-1}, && \text{(2nd order upwind)} \end{aligned}$$

Compute fluxes, use increments over cell average:

$$f^{*,(1)}_{j+1/2} = f\Big(\bar{u}_j + \underbrace{\frac{1}{2}(\bar{u}_{j+1} - \bar{u}_j)}_{\tilde{u}^{(1)}_j}\Big), \qquad f^{*,(2)}_{j+1/2} = f\Big(\bar{u}_j + \underbrace{\frac{1}{2}(\bar{u}_j - \bar{u}_{j-1})}_{\tilde{u}^{(2)}_j}\Big).$$

Lax-Wendroff

For $u_t + au_x$, from finite difference:

$$f^*(u^-, u^+) = \frac{au^- + au^+}{2} - \frac{a^2}{2} \cdot \frac{\Delta t}{\Delta x}(u^+ - u^-).$$

Taylor in time: $u_{\ell+1} = u_\ell + \partial_t u_\ell \cdot h_t + \partial_t^2 u_\ell \cdot h_t/2 + O(h_t^3)$.

$$\begin{aligned} u_t &= -f(u)_x, \\ u_{tt} &= -f(u)_{xt} = -(f(u)_t)_x = -(f'(u)u_t)_x = (f'(u)f(u)_x)_x. \end{aligned}$$

$$\frac{u_{j,\ell+1} - u_{j,\ell}}{h_t} + \frac{f(u_{j+1,\ell}) - f(u_{j-1,\ell})}{2h_x}$$
$$= \frac{h_t}{2h_x}\left[f'(u_{j+1/2,\ell})\frac{f(u_{j+1,\ell}) - f(u_{j,\ell})}{h_x} - f'(u_{j-1/2,\ell})\frac{f(u_{j,\ell}) - f(u_{j-1,\ell})}{h_x}\right]$$

As a Riemann solver:

$$f^*(u^-, u^+) = \frac{f(u^-) + f(u^+)}{2} - \frac{h_t}{h_x}[f'(u^\circ)(f(u^+) - f(u^-))].$$

Monotone Schemes

Definition (Monotone Scheme)

A scheme

$$\begin{aligned} u_{j,\ell+1} &= u_{j,\ell} - \lambda(f^*(u_{j-p},\ldots,u_{j+q}) - f^*(u_{j-p-1},\ldots,u_{j+q-1})) \\ &=: G(u_{j-p-1},\ldots,u_{j+q}) \end{aligned}$$

is called a montone scheme if G is a monotonically nondecreasing function $G(\uparrow,\uparrow,\ldots,\uparrow)$ of each argument.

Monotonicity for Three-Point Schemes

Three-Point Scheme:

$$G(u_{j-1}, u_j, u_{j+1}) = u_j - \lambda[f^*(u_j, u_{j+1}) - f^*(u_{j-1}, u_j)].$$

When is this monotone?

If $f^*(\uparrow, \downarrow)$, then $G(\uparrow, ?, \uparrow)$. To clean up the second argument, consider

$$\frac{\partial G}{\partial u_j} = 1 - \lambda[\underbrace{f_1^* - f_2^*}_{\geq 0}] \geq 0.$$

(The subscripts indicate partial derivatives with respect to the first and second argument.)

If $\lambda(f_1^* - f_2^*) \leq 1$, then $G(\uparrow, \uparrow, \uparrow)$.

Note: Also obtain a time-step restriction.

Lax-Friedrichs is Monotone

$$f^*(u^-, u^+) = \frac{f(u^-) + f(u^+)}{2} - \frac{\alpha}{2}(u^+ - u^-).$$

Show: This is monotone.

Let $\alpha = \max_u |f'(u)|$.

$$\begin{aligned} f_1^* &= \frac{1}{2}[f'(u_j) + \alpha] \geq 0, \\ f_2^* &= \frac{1}{2}[f'(u_{j+1}) - \alpha] \leq 0. \end{aligned}$$

So $f^*(\uparrow, \downarrow)$. Assume h_t is chosen small enough so that $\lambda(f_1^* - f_2^*) \leq 1$ is satisfied.

Monotone Schemes: Properties

Theorem (Good properties of monotone schemes)

- *Local maximum principle:*

$$\min_{i \in \text{stencil around } j} u_i \leq G(u)_j \leq \max_{i \in \text{stencil around } j} u_i.$$

- *L^1-contraction:*

$$\|G(u) - G(v)\|_{L^1} \leq \|u - v\|_{L^1}.$$

- *TVD:*

$$TV(G(u)) \leq TV(u).$$

- *Solutions to monotone schemes satisfy all entropy conditions.*

Godunov's Theorem

Theorem (Godunov)

Monotone schemes are at most first-order accurate.

What now?

> Maybe relax this condition? Maybe only ask for TVD?

Linear Schemes

Definition (Linear Schemes)

A scheme is called a linear scheme if it is linear when applied to a linear PDE:

$$u_t + a u_x = 0,$$

where a is a constant.

Write the general case of a linear scheme for $u_t + u_x = 0$:

$$u_{j,\ell+1} = \sum_{k=-K}^{K} c_k(\lambda) u_{j-k,\ell},$$

where $c_k(\lambda)$ are constants which may depend on $\lambda = h_t/h_x$. Such a linear scheme is monotone iff $c_k(\lambda) \geq 0$ for all k.

Also called positive schemes.

Linear + TVD = ?

Theorem (TVD for linear Schemes)

For linear schemes, TVD $\Rightarrow$ monotone.

What does that mean?

> Linear TVD schemes are at most first order accurate.

Now what?

> Not all bad: Implies that *nonlinear* TVD schemes at least stand a chance.

Harten's Lemma

Theorem (Harten's Lemma)

If a scheme can be written as

$$\bar{u}_{j,\ell+1} = \bar{u}_{j,\ell} + \lambda(C_{j+1/2}\Delta_+\bar{u}_j - D_{j-1/2}\Delta_-\bar{u}_j)$$

with $C_{j+1/2} \geq 0$, $D_{j+1/2} \geq 0$, $1 - \lambda(C_{j+1/2} + D_{j+1/2}) \geq 0$ *and* $\lambda = h_t/h_x$, *then it is TVD.*

As a matter of notation, we have

$$\begin{aligned} \Delta_+ u_j &= u_{j+1} - u_j, \\ \Delta_- u_j &= u_j - u_{j-1}. \end{aligned}$$

We have omitted the time subscript for the time level ℓ.

Harten's Lemma: Proof

$$\begin{aligned}
\Delta_+\bar{u}_{j,\ell+1} &= \Delta_+\bar{u}_{j,\ell} + \lambda\Delta_+\big(C_{j+1/2}\Delta_+\bar{u}_j - D_{j-1/2}\Delta_-\bar{u}_j\big)\\
&= \Delta_+\bar{u}_{j,\ell} + \lambda\big(C_{j+3/2}\Delta_+\bar{u}_{j+1} - D_{j+1/2}\underbrace{\Delta_+\bar{u}_j}_{=\Delta_-\bar{u}_{j+1}}\\
&\quad -C_{j+1/2}\Delta_+\bar{u}_j + D_{j-1/2}\Delta_-\bar{u}_j\big)\\
&= [1-\lambda(C_{j+1/2}+D_{j+1/2})]\Delta_+\bar{u}_j\\
&\quad +\lambda C_{j+3/2}\Delta_+\bar{u}_{j+1} + \lambda D_{j-1/2}\Delta_-\bar{u}_j.\\
|\Delta_+\bar{u}_{j,\ell+1}| &\le [1-\lambda(C_{j+1/2}+D_{j+1/2})]|\Delta_+\bar{u}_j|\\
&\quad +\lambda\underbrace{C_{j+3/2}|\Delta+\bar{u}_{j+1}|}_{C_{j'+1/2}|\Delta_+\bar{u}_{j'}|}+\lambda\underbrace{D_{j-1/2}|\Delta_-\bar{u}_j|}_{D_{j''+1/2}|\Delta_+\bar{u}_{j''}|}.\\
\mathrm{TV}(\bar{u}_{\ell+1}) &= \sum_j|\Delta_+\bar{u}_{j,\ell+1}| \le \sum_j\big[1-\lambda(C_{j+1/2}+D_{j+1/2})\\
&\quad +\lambda C_{j+1/2}+\lambda D_{j+1/2}\big]|\Delta_+\bar{u}_j| \le \mathrm{TV}(u_\ell).
\end{aligned}$$

Minmod Scheme

Still assume $f'(u) \geq 0$.

$$f^{*,(1)}_{j+1/2} = f\Big(\bar{u}_j + \underbrace{\frac{1}{2}(\bar{u}_{j+1} - \bar{u}_j)}_{\tilde{u}_j^{(1)}}\Big), \qquad f^{*,(2)}_{j+1/2} = f\Big(\bar{u}_j + \underbrace{\frac{1}{2}(\bar{u}_j - \bar{u}_{j-1})}_{\tilde{u}_j^{(2)}}\Big).$$

Design a 'safe' thing to use for $\tilde{u}$:

$$\operatorname{minmod}(a,b) := \begin{cases} a & |a| < |b|, ab > 0, \\ b & |b| < |a|, ab > 0, \\ 0 & ab \leq 0, \end{cases} \qquad \tilde{u}_j := \operatorname{minmod}(\tilde{u}_j^{(1)}, \tilde{u}_j^{(2)}).$$

Intuition: TV growth driven by local extrema
$\rightarrow$ if slopes have different signs, revert to first order.

Then consider $f^{*,(3)}_{j+1/2} = f(\bar{u}_j + \tilde{u}_j)$. Called a slope limiter.

Minmod is TVD

Show that Minmod is TVD:

Rewrite

$$\bar{u}_{j,\ell+1} = \bar{u}_j - \lambda[f(\bar{u}_j + \tilde{u}_j) - f(\bar{u}_{j-1} + \tilde{u}_{j-1})] = \bar{u}_j - \lambda[-D_{j-1/2}\Delta_- \bar{u}_j],$$

with

$$\begin{aligned} D_{j-1/2} &= \frac{f(\bar{u}_j + \tilde{u}_j) - f(\bar{u}_{j-1} + \tilde{u}_{j-1})}{\bar{u}_j - \bar{u}_{j-1}} = f'(\xi)\frac{\bar{u}_j - \bar{u}_{j-1} + \tilde{u}_j - \tilde{u}_{j-1}}{\bar{u}_j - \bar{u}_{j-1}} \\ &= \underbrace{f'(\xi)}_{\geq 0}\left[1 + \underbrace{\frac{\tilde{u}_j}{\bar{u}_j - \bar{u}_{j-1}}}_{0 \leq \cdot \leq \frac{1}{2}} - \underbrace{\frac{\tilde{u}_{j-1}}{\bar{u}_j - \bar{u}_{j-1}}}_{0 \leq \cdot \leq \frac{1}{2}}\right] \geq 0. \end{aligned}$$

Minmod: CFL restriction?

Derive a time step restriction for Minmod.

$$D_{j-1/2} \leq 3/2 f'(\xi) \leq \frac{3}{2} \max_u |f'(u)|.$$

Plugging this into the Harten CFL bound gives:

$$1 - \lambda D_{j-1/2} \geq 1 - \frac{3}{2}\lambda \max_u |f'(u)| \geq 0 \Leftarrow \boxed{\lambda \max |f'(\xi)| \leq \tfrac{2}{3}.}$$

What about Time Integration?

$$u^{(1)} = u_\ell + h_t L(u_\ell), \qquad u_{\ell+1} = \frac{u_\ell}{2} + \frac{1}{2}(u^{(1)} + h_t L(u^{(1)})).$$

Above: A version of RK2 with L the ODE RHS. Will this cause wrinkles?

Use: TV is convex. $TV(\alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v}) \le \alpha TV(\boldsymbol{u}) + (1-\alpha) TV(\boldsymbol{v})$.

$$\begin{aligned}
\text{TV}(u_{\ell+1}) &= \text{TV}\left(\frac{u_\ell}{2} + \frac{1}{2}(u^{(1)} + h_t L(u^{(1)}))\right) \\
&\le \frac{1}{2}\text{TV}(u_\ell) + \frac{1}{2}\text{TV}(u^{(1)} + h_t L(u^{(1)})) \\
&\overset{\text{TVD}}{\le} \frac{1}{2}\text{TV}(u_\ell) + \frac{1}{2}\text{TV}(u^{(1)}) \\
&\overset{\text{TVD}}{\le} \frac{1}{2}\text{TV}(u_\ell) + \frac{1}{2}\text{TV}(u_\ell) = \text{TV}(u_\ell).
\end{aligned}$$

General idea: time steppers out of convex comb. of Fw Euler.
(SSP / Strong-Stability Preserving Schemes) Above: SSPRK(2,2)

Total Variation is Convex

Show: $\mathrm{TV}(\cdot)$ is a convex functional.

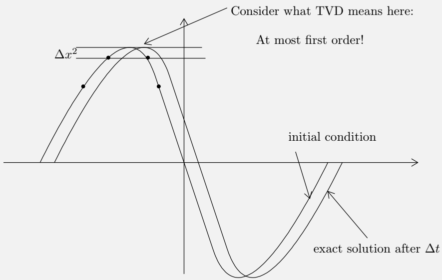
With $0 \le \alpha \le 1$:

$$\begin{aligned}
& \mathrm{TV}(\alpha u + (1-\alpha) v) \\
\le \quad & \sum_j |\alpha(u_j - u_{j-1}) + (1-\alpha)(v_j - v_{j-1})| \\
\le \quad & \sum_j \alpha \, |u_j - u_{j-1}| + (1-\alpha) \, |v_j - v_{j-1}| \\
= \quad & \alpha \, \mathrm{TV}(u) + (1-\alpha) \, \mathrm{TV}(v).
\end{aligned}$$

TVD and High Order

Can TVD schemes be high order everywhere? (aside from near shocks)

Consider $u_t + u_x = 0$.



The solution has an error of h_x^2, which means the approximation to the derivative has error h_x: first order. [Osher/Chakravarthy '84]

High Order at Smooth Extrema

- TVB Schemes [Shu '87]
- ENO [Harten/Engquist/Osher/Chakravarthy '87]
  - Define $W_j = w(x_{j+1/2}) = \int_{x_{1/2}}^{x_{j+1/2}} u(\xi, t) d\xi = h_x \sum_{i=1}^{j} \bar{u}_i$
    - Observe $u_{j+1/2} = w'(x_{j+1/2})$.
    - Approximate by interpolation/numerical differentiation.
  - Start with the linear function $p^{(1)}$ through W_{j-1} and W_j
  - Compute divided differences on (W_{j-2}, W_{j-1}, W_j)
  - Compute divided differences on (W_{j-1}, W_j, W_{j+1})
  - Use the one with the smaller magnitude (of the divided differences) to extend $p^{(1)}$ to quadratic
  - (and so on, adding points on the side with the lowest magnitude of the divided differences)
- WENO [Liu/Osher/Chan '94]

Outline

Systems of Conservation Laws

Linear system of hyperbolic conservation laws, $A \in \mathbb{R}^{m \times m}$:

$$\begin{aligned} \boldsymbol{u}_t + A\boldsymbol{u}_x &= 0, \\ \boldsymbol{u}(x,0) &= \boldsymbol{u}_0(x). \end{aligned}$$

Assumptions on A?

System is hyperbolic [cf. Loret '08] if A is diagonalizable with real eigenvalues.

Let $A\boldsymbol{r}_p = \lambda_p \boldsymbol{r}_p$ $(p = 1, \ldots, m)$. Called strictly hyperbolic if the eigenvalues are distinct. $AR = R\Lambda$.

Substitution $\boldsymbol{v} = R^{-1}\boldsymbol{u}$ attains $\boldsymbol{v}_t + \Lambda \boldsymbol{v}_x = 0$, called characteristic variables.

Recall: Rewrote wave equation in this form early on.

Linear System Solution

$$\boldsymbol{v} = R^{-1}\boldsymbol{u}, \qquad \boldsymbol{v}_t + \Lambda \boldsymbol{v}_x = 0.$$

Write down the solution.

$$u(x,t) = \sum_p \boldsymbol{r}_p v_p(x - \lambda_p t, 0),$$

where

$$\boldsymbol{v}(x,0) = R^{-1}\boldsymbol{u}(x,0).$$

What is the impact on boundary conditions? E.g. $(\lambda_p) = (-c, 0, c)$ for a BC at $x = 0$ for $[0,1]$?

Can only impose BCs on incoming waves! E.g. only one BC (on v_3) at $x = 0$.

Characteristics for Systems (1/2)

Consider system $\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0$. Write in quasilinear form:

$$\boldsymbol{u}_t + A(\boldsymbol{u})\boldsymbol{u}_x = 0 \qquad \text{with} \quad A(\boldsymbol{u}) = J_{\boldsymbol{f}}(\boldsymbol{u}).$$

When hyperbolic?

A diagonalizable w/real eigenvalues. “Strictly” hyperbolic for distinct eigenvalues. Both now local properties.

Characteristics for Systems (2/2)

What about characteristics/shock speeds?

- By considering eigenstates: can still define characteristics.
  m characteristics through each point.
- Characteristic locations no longer obey an ODE.

Are values of $\boldsymbol{u}$ still constant along characteristics?

No, only the coefficients of the eigenstates are constant along characteristics, and only locally.

Shocks and Riemann Problems for Systems

$$\begin{aligned} \boldsymbol{u}_t + A\boldsymbol{u}_x &= 0, \\ \boldsymbol{u}(x,0) &= \begin{cases} \boldsymbol{u}_l & x < 0, \\ \boldsymbol{u}_r & x > 0. \end{cases} \end{aligned}$$

Solution? (Assume strict hyperbolicity with $\lambda_1 < \lambda_2 < \cdots < \lambda_m$.)

$$\boldsymbol{u}_l = \sum_{p=1}^{m} \alpha_p \boldsymbol{r}_p, \quad \boldsymbol{u}_r = \sum_{p=1}^{m} \beta_p \boldsymbol{r}_p. \qquad \text{Then} \quad v_p(x,0) = \begin{cases} \alpha_p & x < 0, \\ \beta_p & x > 0. \end{cases}$$

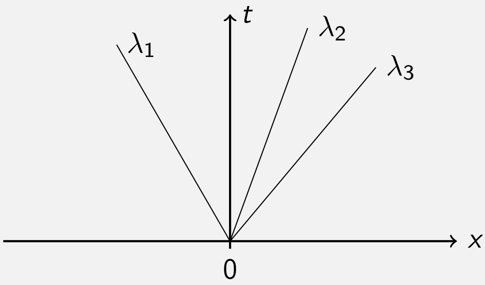
Let $P(x,t)$ be the maximum value of p for which $x - \lambda_p t > 0$, then

$$\boldsymbol{u}(x,t) = \sum_{p=1}^{P(x,t)} \beta_p \boldsymbol{r}_p + \sum_{p=P(x,t)+1}^{m} \alpha_p \boldsymbol{r}_p.$$

Shock Fans (1/2)

What does the solution look like?

Fan of values constant between each characteristic.



Jump across the characteristic associated with λ_p?

$$[\boldsymbol{u}] = (\beta_p - \alpha_p)\boldsymbol{r}_p.$$

Shock Fans (2/2)

Do those jumps satisfy Rankine-Hugoniot?

$$[\boldsymbol{f}] = A[\boldsymbol{u}] = (\beta_p - \alpha_p)A\boldsymbol{r}_p = \lambda_p[\boldsymbol{u}],$$

where λ_p is the propagation speed of the jump.

How can we find intermediate values of $\boldsymbol{u}$?

"Split up" the jump into a sum of jumps:

$$\boldsymbol{u}_r - \boldsymbol{u}_l = (\beta_1 - \alpha_1)\boldsymbol{r}_1 + \cdots + (\beta_m - \alpha_m)\boldsymbol{r}_m.$$

Use Rankine-Hugoniot as a constraint.
This works much the same way in the nonlinear case.

Two Dimensions

$u_t + f(u)_x + g(u)_y = 0$. Finite volume methods generalize in principle:

$$\frac{d\bar{u}_{ij}(t)}{\mathrm{dt}} + \frac{1}{h^2}\int_{y_{j-1/2}}^{y_{j+1/2}} f(u(x_{i+1/2}, y, t)) - f(u(x_{i-1/2}, y, t))dy$$
$$+ \frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}} g(u(x, y_{j+1/2}, t)) - g(u(x, y_{j-1/2}, t))dx$$

Downside: Stencil full ($n \times n$), not star-shaped (cf. FD)

However:

- If a method is TVD in two dimensions, it is at most first order accurate except in trivial cases. [Goodman/Leveque '85].
- The 'reconstruction' idea in complex geometry can become computationally expensive at high order.

Later: discontinuous Galerkin (DG) for high order with c. laws

Outline

Outline

Function Spaces

Consider

$$f_n(x) = \begin{cases} -1 & x \le -\frac{1}{n}, \\ \frac{3n}{2}x - \frac{n^3}{2}x^3 & -\frac{1}{n} < x < \frac{1}{n}, \\ 1 & x \ge 1/n. \end{cases}$$

Converges to the step function. Problem?

f_n continuous, step function not. Want: limits that preserve smoothness properties. Limits defined by norms.

Norms

Definition (Norm)

A norm $\|\cdot\|$ maps an element of a *vector space* into $[0, \infty)$. It satisfies:

- $\|x\| = 0 \Leftrightarrow x = 0$
- $\|\lambda x\| = |\lambda| \|x\|$
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

Convergence

Definition (Convergent Sequence)

$x_n \to x :\Leftrightarrow \|x_n - x\| \to 0$ (convergence in norm)

Definition (Cauchy Sequence)

For all $\epsilon > 0$ there exists an n for which $\|x_\nu - x_\mu\| \leq \epsilon$ for $\mu, \nu \geq n$.

Banach Spaces

Definition (Complete/"Banach" space)

Cauchy $\Rightarrow$ Convergent

What's special about Cauchy sequences?

Limits appear out of thin air. Can be used to construct things.

Counterexamples?

- $\mathbb{Q}$ with absolute value
- C^0 with L^2 norm

More on C^0

Let $\Omega \subseteq \mathbb{R}^n$ be open. Is $C^0(\Omega)$ with $\|f\|_\infty := \sup_{x\in\Omega} |f(x)|$ Banach?

> $f(x) = 1/x$ clearly satisfies $f \in C^0(\Omega)$, but its norm is unbounded, so $\|\cdot\|_\infty$ is not a norm on this space.

Is $C^0(\bar{\Omega})$ with $\|f\|_\infty := \sup_{x\in\Omega} |f(x)|$ Banach?

> Assume $(f_i)_i$ is Cauchy.
>
> - For each x, $(f_i(x))_i$ is Cauchy, so a pointwise limit exists. Call that f.
> - Let $\varepsilon > 0$. There exists N so that $|f_n(x) - f_m(x)| < \varepsilon$ for all $n, m \geq N$ and $x \in \bar{\Omega}$. Taking the limit $m \to \infty$ yields $|f_n(x) - f(x)| < \varepsilon$, i.e. uniform convergence, forcing f to be continuous.

C^m Spaces

Let $\Omega \subseteq \mathbb{R}^n$.

Consider a multi-index $\boldsymbol{k} = (k_1, \dots, k_n)$ and define the symbols

$$D^{\boldsymbol{k}} f = \frac{\partial^{|\boldsymbol{k}|}}{\partial x_1^{k_1} \cdots \partial x_n^{k_n}}, \qquad |\boldsymbol{k}| = k_1 + \cdots + k_n.$$

Definition (C^m Spaces)

$$C^m(\Omega) = \left\{ f \in C^0(\Omega) : D^{\boldsymbol{k}} f \in C^0 \text{ for all } \boldsymbol{k} \text{ with } |\boldsymbol{k}| \leq m \right\},$$

$$C^\infty(\Omega) = \left\{ f \in C^0(\Omega) : D^{\boldsymbol{k}} f \in C^0(\Omega) \text{ for all } \boldsymbol{k} \right\},$$

$$C_0^m(\Omega) = \{ f \in C^m(\Omega) : f \text{ has compact support} \},$$

where compact support means that there is a compact (closed and bounded) set $S \subset \Omega$ for which $f(x) = 0$ if $x \notin S$.

L^p Spaces

Let $1 \leq p < \infty$.

Definition (L^p Spaces)

$$L^p(\Omega) := \left\{ u : (u : \mathbb{R} \to \mathbb{R}) \text{ measurable}, \int_\Omega |u|^p \, dx < \infty \right\},$$

$$\|u\|_p := \left(\int_\Omega |u|^p \, dx \right)^{1/p}.$$

Definition (L^∞ Space)

$$L^\infty(\Omega) := \{ u : (u : \mathbb{R} \to \mathbb{R}), |u(x)| < \infty \text{ almost everywhere} \},$$

$$\|u\|_\infty = \inf \{ C : |u(x)| \leq C \text{ almost everywhere} \}.$$

L^p Spaces: Properties

Theorem (Hölder's Inequality)

For $1 \leq p, q \leq \infty$ *with* $1/p + 1/q = 1$ *and measurable* u *and* v,

$$\|uv\|_1 \leq \|u\|_p \|v\|_q .$$

Theorem (Minkowski's Inequality (Triangle inequality in L^p))

For $1 \leq p \leq \infty$ *and* $u, v \in L^p(\Omega)$,

$$\|u + v\|_p \leq \|u\|_p + \|v\|_p .$$

Inner Product Spaces

Let V be a vector space.

Definition (Inner Product)

An inner product is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ such that for any $f, g, h \in V$ and $\alpha \in \mathbb{R}$

$$
\begin{aligned}
\langle f, f \rangle &\geq 0, \\
\langle f, f \rangle &= 0 \Leftrightarrow f = 0, \\
\langle f, g \rangle &= \langle f, g \rangle, \\
\langle \alpha f + g, h \rangle &= \alpha \langle f, h \rangle + \langle g, h \rangle .
\end{aligned}
$$

Definition (Induced Norm)

$$\|f\| = \sqrt{\langle f, f \rangle}.$$

Hilbert Spaces

Definition (Hilbert Space)

An inner product space that is complete under the induced norm.

Let Ω be open.

Theorem (L^2)

$L^2(\Omega)$ equals the closure of (set of all imits of Cauchy sequences in) $C_0^\infty(\Omega)$ under the induced norm $\|\cdot\|_2$.

Theorem (Hilbert Projection)

Let $M \subseteq V$ be a closed subspace of a Hilbert space V. For any $u \in V$ there exists a unique $v \in M$ such that $u = v + w$ with $w \in M^\perp$.

Weak Derivatives

Define the space L^1_{loc} of locally integrable functions.

$$L^1_{\text{loc}}(\Omega) = \Big\{ u : (u : \mathbb{R} \to \mathbb{R}) \text{ measurable}, \int_\Omega |u(x)\varphi(x)|\, dx < \infty \text{ for every } \varphi \in C_0^\infty(\Omega) \Big\}$$

Definition (Weak Derivative)

$v \in L^1_{\text{loc}}(\Omega)$ is the weak partial derivative of $u \in L^1_{\text{loc}}(\Omega)$ of multi-index order $\boldsymbol{k}$ if

$$\int_\Omega v\varphi dx = (-1)^{|\boldsymbol{k}|} \int_\Omega u D^{\boldsymbol{k}} \varphi dx \qquad \text{for all } \varphi \in C_0^\infty(\Omega).$$

In this case, $D^{\boldsymbol{k}} u := v$.

Weak Derivatives: Examples (1/2)

Consider all these on the interval $[-1, 1]$.

$$f_1(x) = 4(1-x)x$$

$D_w f_1(x) = 4 - 8x$. For ("strongly") differentiable functions, weak and strong derivatives coincide.

$$f_2(x) = \begin{cases} 2x & x \leq 1/2, \\ 2-2x & x > 1/2. \end{cases}$$

"Kinks" in the function are allowed (but jumps are not):

$$D_w f_2(x) = \begin{cases} 2 & x \leq 1/2, \\ -2 & x > 1/2. \end{cases}$$

Weak Derivatives: Examples (2/2)

$$f_3(x) = \sqrt{\frac{1}{2}} - \sqrt{|x - 1/2|}$$

Even cusps are allowed:

$$D_w f_3(x) = \begin{cases} \frac{1}{2\sqrt{1/2 - x}} & x < 1/2, \\ -\frac{1}{2\sqrt{x - 1/2}} & x > 1/2. \end{cases}$$

Sobolev Spaces

Let $\Omega \subset \mathbb{R}^n$, $k \in \mathbb{N}_0$ and $1 \leq p < \infty$.

Definition ((k,p)-Sobolev Norm/Space)

$$\|u\|_{k,p} := \sqrt[p]{\sum_{|\boldsymbol{\alpha}| \leq k} \|D_w^{\boldsymbol{\alpha}} u\|_p^p},$$

$$|u|_{k,p} := \sqrt[p]{\sum_{|\boldsymbol{\alpha}| = k} \|D_w^{\boldsymbol{\alpha}} u\|_p^p}.$$

$$W^{k,p}(\Omega) := \left\{ u : (u : \Omega \to \mathbb{R}), \|u\|_{k,p} < \infty \right\}.$$

More Sobolev Spaces

$W^{0,2}$?

Equal to L^2.

$W^{s,2}$?

Also called H^s, a Hilbert space, with an induced norm. From what scalar product?

$H_0^1(\Omega)$?

Closure of the space $C_0^\infty(\Omega)$ under $\|u\|_{k,p}$.
The Sobolev way of saying zero on the boundary.

Outline

An Elliptic Model Problem

Let $\Omega \subset \mathbb{R}^n$ open, bounded, $f \in H^1(\Omega)$.

$$\begin{aligned} -\nabla \cdot \nabla u + u &= f(x) \quad (x \in \Omega), \\ u(x) &= 0 \qquad (x \in \partial\Omega). \end{aligned}$$

Let $V := H_0^1(\Omega)$. Integration by parts? (Gauss's theorem applied to $a\boldsymbol{b}$):

$$\int_\Omega \nabla a \cdot \boldsymbol{b} + \int_\Omega a \nabla \cdot \boldsymbol{b} = \int_\Omega \nabla \cdot (a\boldsymbol{b}) = \int_{\partial\Omega} \hat{\boldsymbol{n}} \cdot (a\boldsymbol{b}).$$

Weak form?

Multiply by test function $v \in V$, integrate by parts:

$$\int_\Omega \nabla u \cdot \nabla v - \underbrace{\int_{\partial\Omega} \hat{n} \cdot (v \nabla u)}_{=0\ (v \in H_0^1)} + \int_\Omega uv = \int_\Omega fv.$$

Motivation: Bilinear Forms and Functionals

$$\int_\Omega \nabla u \cdot \nabla v + \int_\Omega uv = \int fv.$$

This is the weak form of the strong-form problem. The task is to find a $u \in V$ that satisfies this for all test functions $v \in V$.

Recast this in terms of bilinear forms and functionals:

$$\begin{aligned} a(u,v) &= \langle \nabla u, \nabla v \rangle + \langle u, v \rangle, \\ g(v) &= \langle f, v \rangle, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the L^2 inner product. Then the weak form is equivalent to

$$a(u,v) = f(v) \qquad \text{for all } v \in V.$$

This motivates further study of Hilbert spaces and objects in them.

Dual Spaces and Functionals

Bounded Linear Functional

Let $(V, \|\cdot\|)$ be a Banach space. A linear functional is a linear function $g : V \to \mathbb{R}$. It is bounded ($\Leftrightarrow$ continuous) if there exists a constant C so that $|g(v)| \leq C \|v\|$ for all $v \in V$.

Dual Space

Let $(V, \|\cdot\|)$ be a Banach space. Then the dual space V' is the space of bounded linear functionals on V.

Dual Space is Banach (cf. e.g. Trèves 1967)

V' is a Banach space with the dual norm

$$\|g\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{|g(v)|}{\|v\|_V}.$$

Functionals in the Model Problem

Is g from the model problem a bounded functional? (In what space?)

Must use same space as rest of problem: $H^1(\Omega)$.

$$\|g\|_{V'} = \sup_{v \in H^1 \setminus \{0\}} \frac{|\langle f, v\rangle_{L^2}|}{\|v\|_{H^1}} \le \frac{\|f\|_{L^2}\|v\|_{L^2}}{\|v\|_{L^2} + \|D_w v\|_{L^2}} \le \frac{\|f\|_{L^2}\|v\|_{L^2}}{\|v\|_{L^2}} = \|f\|_{L^2}$$

using Cauchy-Schwarz. Find: $f \in L^2$ leads to bounded g in H^1.

That bound felt loose and wasteful. Can we do better?

Define negative-index Sobolev norms:

$$\|f\|_{H^{-1}} = \sup_{v \in H^1(\Omega) \setminus \{0\}} \frac{|\langle f, v\rangle_{L^2}|}{\|v\|_{H^1}}.$$

Bound (by definition) $|g(v)| \le \|f\|_{H^{-1}} \|v\|_{H^1}$. Allows $f \in H^{-1}$.

Riesz Representation Theorem (1/3)

Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$.

Theorem (Riesz)

Let g be a bounded linear functional on V, i.e. $g \in V'$. Then there exists a unique $u \in V$ so that $g(v) = \langle u, v \rangle$ for all $v \in V$.

Let $g \in V'$. $N(\cdot)$ below represents the nullspace.
Case 1. $N(g) = V$. $u = 0$ works, unique by scalar product axioms.
Case 2. $N(g) \neq V$. Let $w \in N(g)^{\perp} \setminus \{0\}$. Let $\alpha = g(w) \neq 0$.

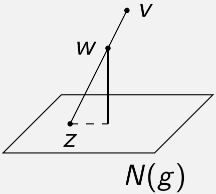
$$g\left(\frac{g(v)}{\alpha} w\right) = \frac{g(v)}{\alpha} g(w) = g(v) \qquad \text{for all } v \in V.$$

Let $v \in V$ be arbitrary, and let $z := v - (g(v)/\alpha)w$. (Feel reminded of Gram-Schmidt?) Then $g(z) = g(v) - g(v) = 0$, i.e. $z \in N(g)$, i.e. $\langle z, w \rangle_V = 0$ since $w \in N(g)^{\perp}$.

Riesz Representation Theorem: Proof (2/3)

Have $w \in N(g)^{\perp} \setminus \{0\}$, $\alpha = g(w) \neq 0$, and $z := v - (g(v)/\alpha)w \perp w$.

$$0 = \left\langle v - \frac{g(v)}{\alpha} w, w \right\rangle \quad \Leftrightarrow \quad \left\langle \frac{g(v)}{\alpha} w, w \right\rangle = \langle v, w \rangle \qquad \text{for all } v \in V.$$



Multiplying by $\alpha / \langle w, w \rangle$ yields

$$g(v) = \left\langle v, \underbrace{\frac{\overbrace{g(w)}^{=\alpha}}{\langle w, w \rangle_V} w}_{u:=} \right\rangle .$$

Riesz Representation Theorem: Proof (3/3)

Uniqueness of u?

Suppose we have two: u and $\hat{u}$ so that

$$g(v) = \langle u, v \rangle = \langle \hat{u}, v \rangle \quad \Rightarrow \quad \langle u - \hat{u}, v \rangle = 0 \quad \text{for all } v \in V,$$

Plugging in $v = u - \hat{u}$ yields $u - \hat{u} = 0$ by the properties of the inner product.

Back to the Model Problem

$$\begin{aligned} a(u,v) &= \langle \nabla u, \nabla v \rangle_{L^2} + \langle u, v \rangle_{L^2} \\ g(v) &= \langle f, v \rangle_{L^2} \\ a(u,v) &= g(v) \end{aligned}$$

Have we learned anything about the solvability of this problem?

In this particular case, observe that $a(u,v) = \langle u, v \rangle_{H^1}$. By the Riesz Representation theorem and knowing that g is a bounded linear functional in H^1, we know that there exists a unique u so that

$$a(u,v) = \langle u, v \rangle_{H^1} = g(v).$$

Poisson

Let $\Omega \subset \mathbb{R}^n$ open, bounded, $f \in H^{-1}(\Omega)$.

$$\begin{aligned} -\nabla \cdot \nabla u &= f(x) \quad (x \in \Omega), \\ u(x) &= 0 \qquad (x \in \partial\Omega). \end{aligned}$$

This is called the Poisson problem (with Dirichlet BCs).

Weak form?

$$\underbrace{\int_\Omega \nabla u \cdot \nabla v dx}_{a(u,v)} = \underbrace{\int_\Omega f(x)v(x)dx}_{g(v)} \quad \text{for all } v \in V.$$

We know that g is a bounded linear functional in H_0^1, but $a(u,v)$ is no longer identical to our inner product. Maybe we can come up with some conditions that make a 'sufficiently similar' to an inner product?

Ellipticity

Let V be Hilbert space.

V-Ellipticity

A bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is called coercive if there exists a constant $c_0 > 0$ so that

$$c_0 \|u\|_V^2 \le a(u,u) \quad \text{for all } u \in V,$$

and a is called continuous if there exists a constant $c_1 > 0$ so that

$$|a(u,v)| \le c_1 \|u\|_V \|v\|_V \quad \text{for all } u, v \in V.$$

If a is both coercive and continuous on V, then a is said to be V-elliptic.

Lax-Milgram Theorem

Let V be Hilbert space with inner product $\langle \cdot, \cdot \rangle$.

Lax-Milgram, Symmetric Case

Let a be a V-elliptic bilinear form that is also symmetric, and let g be a bounded linear functional on V.
Then there exists a unique $u \in V$ so that $a(u, v) = g(v)$ for all $v \in V$.

a defines an inner product $\langle u, v \rangle_a = a(u, v)$ on V, with linearity and symmetry trivial, and:

- Show $a(u, u) \geq 0$.
  $a(u, u) \geq c_0 \|u\|_V^2 \geq 0$ by coercivity,
- Show $a(u, u) = 0 \Rightarrow u = 0$.
  $0 = a(u, u) \geq c_0 \|u\|_V^2 \geq 0$, i.e. $\|u\|_V = 0$, i.e. $u = 0$.

From the Riesz representation theorem, there exists a unique $u \in V$ so that $a(u, v) = \langle u, v \rangle_a = g(v)$.

Back to Poisson

Can we declare victory for Poisson?

Continuity of a holds:

$$\left|\int_\Omega \nabla u \cdot \nabla v dx\right| = |\langle \nabla u, \nabla v \rangle_{L^2}| \leq \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} \leq \|u\|_{H^1} \|v\|_{H^1}.$$

However coercivity is less clear:

$$\int_\Omega \nabla u \cdot \nabla u dx \overset{?}{\geq} c_1 \left(\int_\Omega \nabla u \cdot \nabla u dx + \int_\Omega u^2 dx \right).$$

Can this inequality hold in general, without further assumptions?

No: a constant would violate it.

Poincaré-Friedrichs Inequality (1/3)

Theorem (Poincaré-Friedrichs Inequality)

Suppose $\Omega \subset \mathbb{R}^n$ is bounded and $u \in H_0^1(\Omega)$. Then there exists a constant $C > 0$ such that

$$\|u\|_{L^2} \leq C \|\nabla u\|_{L^2}.$$

Outline: Helpful identity, result in $C_0^\infty(\Omega)$, result in $H_0^1(\Omega)$.

A helpful identity. For $u \in C_0^\infty(\Omega)$,

$$\begin{aligned}
\nabla \cdot (u^2 \boldsymbol{x}) &= \partial_{x_1}(u^2 x_1) + \cdots + \partial_{x_n}(u^2 x_n) \\
&= u^2 + 2(u\partial_{x_1} u)x_1 + \cdots + u^2 + 2(u\partial_{x_n} u)x_n \\
&= nu^2 + 2u(\nabla u \cdot \boldsymbol{x}). \\
\Rightarrow \quad u^2 &= \frac{1}{n}\nabla \cdot (u^2 \boldsymbol{x}) - \frac{2}{n} u(\nabla u \cdot \boldsymbol{x}).
\end{aligned}$$

Poincaré-Friedrichs Inequality (2/3)

Prove the result in $C_0^\infty(\Omega)$.

$$
\begin{aligned}
\|u\|_{L^2}^2 &= \int_\Omega u^2 d\boldsymbol{x} = \int_\Omega \frac{1}{n}\nabla\cdot(u^2\boldsymbol{x}) - \frac{2}{n}u(\nabla u\cdot\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{1}{n}\int_{\partial\Omega} \underbrace{\hat{\boldsymbol{n}}\cdot(u^2\boldsymbol{x})}_{0} ds_{\boldsymbol{x}} - \frac{2}{n}\int_\Omega u(\nabla u\cdot\boldsymbol{x})d\boldsymbol{x} \\
&\leq \frac{2}{n}\max_{\boldsymbol{x}\in\Omega}|\boldsymbol{x}|\int_\Omega |u\nabla u|\,d\boldsymbol{x} \leq \frac{2}{n}\max_{\boldsymbol{x}\in\Omega}|\boldsymbol{x}|\,\|u\|_{L^2}\,\|\nabla u\|_{L^2} \\
\Rightarrow \|u\|_{L^2} &\leq \underbrace{\frac{2}{n}\max_{\boldsymbol{x}\in\Omega}|\boldsymbol{x}|}_{C}\,\|\nabla u\|_{L^2}.
\end{aligned}
$$

Poincaré-Friedrichs Inequality (3/3)

Prove the result in $H_0^1(\Omega)$.

Let $u \in H_0^1(\Omega)$. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, let $(u_k) \subset C_0^\infty$. Then the inequality holds for each u_k, and $\|u_k\|_{L^2} \to \|u\|_{L^2}$ and $\|\nabla u_k\|_{L^2} \to \|\nabla u\|_{L^2}$.

Back to Poisson, Again

Show that the Poisson bilinear form is coercive.

$$\frac{1}{C^2+1}\|u\|^2_{H^1(\Omega)} = \frac{1}{C^2+1}\left(\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2}\right) \leq \|\nabla u\|^2_{L^2} = a(u,u).$$

Draw a conclusion on Poisson:

Because of coercivity and continuity of a, the Poisson weak form admits a unique solution in $H^1_0(\Omega)$.

Outline

Ritz-Galerkin

Some key goals for this section:

- How do we use the weak form to compute an approximate solution?
- What can we know about the accuracy of the approximate solution?

Can we pick one underlying principle for the construction of the approximation?

Considered: Weak form $a(u, v) = g(v)$ for all $v \in V \subseteq H$, where H is a Hilbert space. (Think of V as H_0^1 for example.)
Idea: Choose a finite-dimensional subspace $V_h \subset V$, find a solution $u_h \in V_h$ to the weak-form problem

$$a(u_h, v_h) = g(v_h) \qquad \text{for all } v_h \in V_h.$$

This is called Ritz-Galerkin approximation.

Galerkin Orthogonality

$$a(u, v) = g(v) \qquad \text{for all } v \in V, a(u_h, v_h) \;\; = g(v_h) \qquad \text{for all } v_h \in V_h.$$

Observations?

Observe that the 'continuous' weak form also allows v_h to be plugged in:

$$a(u, v_h) = g(v_h) \qquad \text{for all } v_h \in V_h.$$

Subtracting the two leads to Galerkin Orthgonality:

$$a(u_h - u, v_h) = 0 \qquad \text{for all } v_h \in V_h,$$

i.e. using $a(\cdot, \cdot)$ as a (sort of) inner product, the error $u - u_h$ is orthogonal to the space of test functions.

Céa's Lemma

Let $V \subset H$ be a closed subspace of a Hilbert space H.

Céa's Lemma

Let $a(\cdot,\cdot)$ be a coercive and continuous bilinear form on V. In addition, for a bounded linear functional g on V, let $u \in V$ satisfy

$$a(u,v) = g(v) \qquad \text{for all } v \in V.$$

Consider the finite-dimensional subspace $V_h \subset V$ and $u_h \in V_h$ that satisfies

$$a(u_h,v_h) = g(v_h) \qquad \text{for all } v_h \in V_h.$$

Then

$$\|u - u_h\|_V \leq \frac{c_1}{c_0} \inf_{v_h \in V_h} \|u - v_h\|_V .$$

Céa's Lemma: Proof

Recall Galerkin orthgonality: $a(u_h - u, v_h) = 0$ for all $v_h \in V_h$. Show the result.

For any $v_h \in V_h$,

$$
\begin{aligned}
c_0 \|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \qquad \text{(coercivity)} \\
&= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\
&= a(u - u_h, u - v_h) \qquad \text{(Galerkin orth.)} \\
&\leq c_1 \|u - u_h\|_V \|u - v_h\|_V .
\end{aligned}
$$

Dividing by $\|u - u_h\|_V$ completes the proof.

Elliptic Regularity

Definition (H^s Regularity)

Let $m \geq 1$, $H_0^m(\Omega) \subseteq V \subseteq H^m(\Omega)$ and $a(\cdot,\cdot)$ a V-elliptic bilinear form. The bilinear form $a(u,v) = \langle f, v\rangle$ for all $v \in V$ is called H^s regular, if for every $f \in H^{s-2m}$ there exists a solution $u \in H^s(\Omega)$ and we have with a constant $C(\Omega, a, s)$,

$$\|u\|_{H^s} \leq C(\Omega, a, s)\, \|f\|_{H^{s-2m}}\,.$$

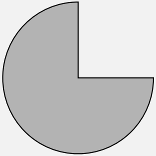
Theorem (Elliptic Regularity (cf. Braess Thm. 7.2))

Let a be a H_0^1-elliptic bilinear form with sufficiently smooth coefficient functions.

- *If Ω is convex, then then Dirichlet problem is H^2 regular.*
- *Let $s \geq 2$. If $\partial\Omega$ is C^s, the Dirichlet problem is H^s regular.*

Elliptic Regularity: Counterexamples

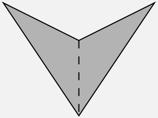
Are the conditions on the boundary essential for elliptic regularity?



Consider $\triangle u = 0$, $u(e^{i\phi}) = \sin(2/3\phi)$, $u = 0$ elsewhere.

- $u(z) = \mathrm{Im}(z^{2/3})$ with $z = x + iy \in \mathbb{C}$.
- Derivative: $(2/3)z^{-1/3}$: unbounded $\Rightarrow$ $u \notin H^2$!

Are there any particular concerns for mixed boundary conditions?



Homogeneous Neumann on dashed line with (e.g.) left half, Dirichlet elsewhere.

- Solution could be found by solving on whole domain using reflected Dirichlet BCs.
- Reentrant corner $\Rightarrow$ $u \notin H^2$ (in gen.)

Estimating the Error in the Energy Norm

Come up with an idea of a bound on $\|u - u_h\|_{H^1}$.

$$\|u - u_h\|_{H^1} \underset{\text{Céa}}{\leq} C \inf_{v_h \in V_h} \|u - v_h\|_{H^1} \leq C \|u - I_h u\|_{H^1}$$

$$\underset{\text{TBD}}{\leq} C_1 h \|u\|_{H^2} \underset{H^2 \text{ reg.}}{\leq} c_2 h \|f\|_{L^2}.$$

What's still to do?

- we still need to figure out what V_h will be,
- I_h is some interpolation operator that we will define more precisely later, and
- we need to worry about the interpolation error bound ("TBD")
- Finally, H^1 is kind of a weird norm. Can we get an error estimate in L^2?

L^2 Estimates

Let H be a Hilbert space with the norm $\|\cdot\|_H$ and the inner product $\langle \cdot, \cdot \rangle$. (Think: $H = L^2$, $V = H^1$.)

Theorem (Aubin-Nitsche)

Let $V \subseteq H$ be a subspace that becomes a Hilbert space under the norm $\|\cdot\|_V$. Let the embedding $V \to H$ be continuous. Then we have for the finite element solution $u \in V_h \subset V$:

$$\|u - u_h\|_H \leq c_1 \|u - u_h\|_V \sup_{g \in H} \left[\frac{1}{\|g\|_H} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_V \right],$$

if with every $g \in H$ we associate the unique (weak) solution φ_g of the equation (also called the dual problem)

$$a(w, \varphi_g) = \langle g, w \rangle \quad \textit{for all } w \in V,$$

Aubin-Nitsche: Proof

The norm of an element in a Hilbert space can be determined via the scalar product: $\|w\|_H = \sup_{g \in H} \langle g, w \rangle / \|g\|_H$.

$$\begin{aligned} \langle g, u - u_h \rangle \underset{\text{Def. } \varphi_g}{=} & \; a(u - u_h, \varphi_g) \underset{\text{Galerkin orth.}}{=} a(u - u_h, \varphi_g - v_h) \\ \underset{\text{cont. } a}{\leq} & \; c_1 \|u - u_h\|_V \|\varphi_g - v_h\|_V \, . \end{aligned}$$

Since this argument is valid for any $v_h \in V_h$, we obtain

$$\langle g, u - u_h \rangle \leq c_1 \|u - u_h\|_V \inf_{v_h \in V_h} \|\varphi_g - v_h\|_V \, .$$

Plugging into the norm relationship yields

$$\|u - u_h\|_H = \sup_{g \in H} \frac{\langle g, w \rangle}{\|g\|_H} \leq c_1 \|u - u_h\|_V \sup_{g \in H} \left[\frac{1}{\|g\|_H} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_V \right] .$$

L^2 Estimates using Aubin-Nitsche

$$\|u - u_h\|_H \le c_1 \|u - u_h\|_V \sup_{g \in H} \left[\frac{1}{\|g\|_H} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_V \right],$$

If $u \in H_0^1(\Omega)$, what do we get from Aubin-Nitsche?

As before (e.g. Poisson: symmetry of a: primal prob. = dual prob.):

$$\inf_{v_h \in V_h} \|\varphi_g - v_h\|_{H^1} \le C \|\varphi_g - I_h \varphi_g\|_{H^1} \le C_1 h \|\varphi_g\|_{H^2} \le c_2 h \|g\|_{L^2}.$$

So $\|u - u_h\|_{L^2} \le Ch \|u - u_h\|_{H^1}$.

So does Aubin-Nitsche give us an L^2 estimate?

Had (aside from missing pieces): $\|u - u_h\|_{H^1} \le c_2 h \|f\|_{L^2}$.
If we have $f \in L^2(\Omega)$ and hence $u \in H^2(\Omega)$ (H^2 regularity), then

$$\|u - u_h\| \le Ch^2 \|f\|$$

Outline

Finite Elements in 1D: Discrete Form

$\Omega := [\alpha, \beta]$. Look for $u \in H_0^1(\Omega)$, so that $a(u, \varphi) = \langle f, \varphi \rangle$ for all $\varphi \in H_0^1(\Omega)$. Choose $V_h = \text{span}\{\varphi_1, \ldots, \varphi_n\}$ and expand $u_h = \sum_{i=1}^n u_h^i \varphi_i \in V_h$. Find the discrete system.

$$a\left(\sum_{i=1}^n u_h^i \varphi_i, \varphi\right) = \langle f, \varphi \rangle \quad \text{for all } \varphi \in V_h,$$

We may as well choose the basis (φ_i) to represent $\varphi \in V_h$:

$$a\left(\sum_{i=1}^n u_h^i \varphi_i, \varphi_j\right) = \langle f, \varphi_j \rangle \quad \text{for all } j \in \{1, \ldots, n\}.$$

This *could* lead to a linear system $Au = b$, where $A = \{a_{i,j}\} \in \mathbb{R}^{n \times n}$ with $a_{i,j} = a(\varphi_i, \varphi_j)$, $u = \{u_h^i\}$, $b_j = \langle f, \varphi_j \rangle$, but we choose not to go this route.

Grids and Hats

Let $I_i := [\alpha_i, \beta_i]$, so that $\bar{\Omega} = \bigcup_{i=0}^{N} I_i$ and $I_i^\circ \cap I_j = \emptyset$ for $i \neq j$. Consider a grid

$$\alpha = x_0 < \cdots < x_N < x_{N+1} = \beta,$$

i.e. $\alpha_i = x_i$, $\beta_i = x_{i+1}$ for $i \in \{0, \ldots, N\}$. The $\{x_i\}$ are called nodes of the grid. $h_i := x_{i+1} - x_i$ for $i \in \{0, \ldots, N\}$ and $h := \max_i h_i$. V_h? Basis?

$$P_h^1 := \{v_h \in C^0(\bar{\Omega}) : \text{for all } i \in \{0, \ldots, N\}, v_h|_{I_i} \in \mathbb{P}_1\}.$$

For $i \in \{0, \ldots, N+1\}$, let

$$\varphi_i(x) := \begin{cases} \frac{1}{h_{i-1}}(x - x_{i-1}) & x \in I_{i-1}, \\ \frac{1}{h_i}(x_{i+1} - x) & x \in I_i, \\ 0 & \text{otherwise} \end{cases} \in P_h^1.$$

Observe: The set $\{\varphi_i\}_i$ forms a basis of P_h^1.

Degrees of Freedom and Matrices

Define something more general than basis coefficients to solve for.

- For $i \in \{0, \ldots, N+1\}$, let $\gamma_i : C(\bar{\Omega}) \to \mathbb{R}$.
  Here: $v \mapsto \gamma_i(v) := v(x_i) \in \mathbb{R}$.
  Generally: could be derivatives etc. (cf. splines).
- $\{\gamma_i\}_{i=0}^{N+1}$ are global degrees of freedom in P_h^1.
- $\{\gamma_i\}_{i=0}^{N+1}$ forms a basis of the dual space $(P_h^1)'$.
  (i.e. uniquely determine $\varphi \in V_h$, global unisolvence)

Define shape functions and assemble the stiffness matrix:

Shape functions $\hat{\varphi} \in V_h$ satisfy $\gamma_j(\hat{\varphi}_i) = \delta_{i,j}$ for $i, j \in \{0, \ldots, N+1\}$.

$$a(u_h, \hat{\varphi}_i) = \langle f, \varphi_i \rangle \Leftrightarrow \sum_{j=1}^{N} \underbrace{\gamma_j(u_h)}_{=u_h^i} \underbrace{a(\hat{\varphi}_j, \hat{\varphi}_i)}_{(A_h)_{i,j}} = \underbrace{\langle f, \varphi_i \rangle}_{(\boldsymbol{b}_h)_i} (j = 1, \ldots, N)$$

A Matrix Property for Efficiency

$$(A_h)_{i,j} = a(\hat{\varphi}_j, \hat{\varphi}_i).$$

Anything special about the matrix?

Only $a_{i,i}, a_{i,i+1}, a_{i,i-1} \neq 0$ in the ith row of A is nonzero. Sparse.

Error Estimation

According to Céa, what's our main missing piece in error estimation now?

An interpolation operator

$$\begin{aligned} I_h^1 : C^0(\bar{\Omega}) &\to P_h^1, \\ v &\mapsto \sum_{i=0}^{N+1} \gamma_i(v)\hat{\varphi}_i \in P_h^1. \end{aligned}$$

Next: need to estimate its accuracy.

Interpolation Error (1D-only)

For $v \in H^2(\Omega)$,

$$\begin{aligned} \left\| v - I_h^1 v \right\|_{L^2} &\leq h^2 \left| v \right|_{H^2} \quad \text{for all } h > 0, \\ \left| (v - I_h^1 v) \right|_{H^1} &\leq h \left| v \right|_{H^2} \quad \text{for all } h > 0. \end{aligned}$$

If $v \in H^1(\Omega) \setminus H^2(\Omega)$,

$$\begin{aligned} \left\| v - I_h^1 v \right\|_{L^2} &\leq h \left| v \right|_{H^1} \quad \text{for all } h > 0, \\ \lim_{h \to 0} \left| (v - I_h^1 v) \right|_{H^1} &= 0. \end{aligned}$$

Is I_h^1 defined for $v \in H^2$? for $v \in H^1 \setminus H^2$?

Depends on the dimension n and the domain Ω. Need to consider the Sobolev Embedding Theorem.

Interpolation Error: Towards an Estimate

Provide an a-priori estimate.

$$\|u - u_h\|_{H^1} \leq \frac{c_1}{c_0} \inf_{v_h \in P_h^1} \|u - v_h\|_{H^1} \leq \frac{c_1}{c_0} \left\|u - I_h^1 u\right\|_{H^1} \leq \frac{c_1}{c_0} h \, |u|_{H^2} \, .$$

What's the relationship between $I_h^1 u$ and u_h?

None!

Local-to-Global

Is there a simple way of constructing the polynomial basis?

The basis functions $\{\varphi_i\}_{i=1}^N$ can be viewed as a composition of

- grid-independent reference basis functions on a reference element, and
- geometric transformations from the reference element to the grid.

Local-to-Global: Math

Construct a polynomial basis using this approach.

Let $\hat{\kappa} = [0, 1]$ be the reference interval and consider the affine transformations $T_I : \hat{x} \in \hat{\kappa} \mapsto x = x_i + \hat{x} h_i$ for $i \in \{0, \ldots, N\}$. Define the shape functions

$$\begin{aligned} \hat{\varphi}_0(\hat{x}) &:= 1 - \hat{x} \quad \text{for all } \hat{x} \in \hat{\kappa}, \\ \hat{\varphi}_1(\hat{x}) &:= \hat{x} \quad \text{for all } \hat{x} \in \kappa. \end{aligned}$$

These functions form a basis of $P_1(\hat{\kappa})$. Then

$$\varphi_i(x) = \begin{cases} (\hat{\varphi}_1 \circ T_{i-1}^{-1})(x) & x \in [x_{i-1}, x_i], \\ (\hat{\varphi}_0 \circ T_i^{-1})(x) & x \in [x_i, x_{i+1}]. \end{cases}$$

Demo

Demo: Developing FEM in 1D

Going Higher Order

$\Omega \subset \mathbb{R}$ with a grid as above.

Possible extension:

$$P_h^k := \{v_h \in C^0(\bar{\Omega}) : \text{for all } i \in \{1, \ldots, N\}, v_h|_{I_i} \in \mathbb{P}_k\}.$$

Higher Order Approximation

Let $0 \leq \ell \leq k$. Then for $v \in H^{\ell+1}(\Omega)$,

$$\left\| v - I_h^k v \right\|_{L^2} + h \left| (v - I_h^k v) \right|_{H^1} \leq Ch^{\ell+1} \left| v \right|_{H^{\ell+1}}.$$

High-Order: Degrees of Freedom

Define some degrees of freedom (or DoFs) for high-order 1D FEM.

Let $\{\gamma_j\}_{j=0}^{N+1} \in (V_h^1)'$ be the linear functionals so that

$$\gamma_j(v_h) = v_h(x_j) \quad \text{for all } v_h \in V_h^1.$$

Using terminology from classical mechanics, these functions are called (global) degrees of freedom. The functions $\{\varphi_i\}_{i=0}^{N+1}$ that are defined so that

$$\gamma_j(\varphi_i) = \delta_{i,j} \quad (i, j \in \{0, \ldots, N+1\}, \varphi_i \in V_h^1)$$

holds are called (global) shape functions. One can also define local shape functions on the reference element.

High-Order: Local Basis

Define local form functions for high-order 1D FEM.

The local form functions are typically chosen to be Lagrange polynomials:

$$\hat{\varphi}_i^k(\hat{x}) = \frac{\prod_{j=0, j\neq i}^{k}(\hat{x} - \hat{x}_j)}{\prod_{j=0, j\neq i}^{k}(\hat{x}_i - \hat{x}_j)},$$

where $\hat{x}_j = j/k$ for $i = 0, \ldots, k$.
$x_{i,j} := x_i + (j/k)h_i$ for $i = 0, \ldots, N$ and $j = 0, \ldots, k-1$, further $x_{N+1,0} = 0$. Then

$$\dim(V_h^k) = k(N+1) + 1.$$

High-Order: Global Basis

Obtain the global shape functions for high-order 1D FEM.

Define

$$\varphi_{i,0}(x) := \begin{cases} \hat{\varphi}_k^k \circ T_{i-1}^{-1}(x) & x \in [x_{i-1}, x_i], \\ \hat{\varphi}_0^k \circ T_i^{-1}(x) & x \in [x_i, x_{i+1}], \\ 0 & \text{otherwise}, \end{cases}$$

and

$$\varphi_{i,j}(x) := \begin{cases} \hat{\varphi}_j^k \circ T_i^{-1}(x) & x \in [x_i, x_{i+1}], \\ 0 & \text{otherwise}. \end{cases}$$

for $j = 0, \ldots, k-1$ und $i = 0, \ldots, N$.

Outline

A Boundary Value Problem

Consider the following elliptic PDE

$$-\nabla \cdot (\kappa(\boldsymbol{x}) \nabla u) = f(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Omega \subset \mathbb{R}^2,$$
$$u(\boldsymbol{x}) = 0 \quad \text{when} \quad \boldsymbol{x} \in \partial\Omega.$$

Weak form?

Multiply by a test function $v \in H_0^1(\Omega)$ and integrate by parts:

$$\int_\Omega [-\nabla \cdot (\kappa(\boldsymbol{x}) \nabla u) - f(\boldsymbol{x})]\, v \,\mathrm{d}\boldsymbol{x} = 0$$

$$\Leftrightarrow -\int_{\partial\Omega} v\,[\kappa \widehat{\boldsymbol{n}} \cdot \nabla u]\,\mathrm{d}\Gamma + \int_\Omega [\kappa(\boldsymbol{x}) \nabla u \cdot \nabla v - f(\boldsymbol{x})\, v]\,\mathrm{d}\boldsymbol{x} = 0.$$

The boundary integral vanishes since $v \in H_0^1$ and we find

$$\int_\Omega \kappa(\boldsymbol{x}) \nabla u \cdot \nabla v \,\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\, v \,\mathrm{d}\boldsymbol{x}.$$

Weak Form: Bilinear Form and RHS Functional

Hence the problem is to find $u \in V$, such that

$$a(u, v) = g(v), \quad \text{for all } v \in V = H_0^1(\Omega)$$

where...

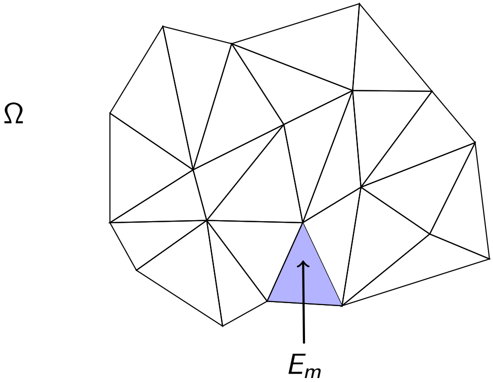
$$a(u, v) := \int_\Omega \kappa(\boldsymbol{x}) \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{x},$$
$$g(v) := \int_\Omega f(\boldsymbol{x}) v \, \mathrm{d}\boldsymbol{x},$$

Is this symmetric, coercive, and continuous?

- Symmetric: yes.
- Coercive: When there exists c so that $0 < c \leq \kappa(\boldsymbol{x})$ for all $\boldsymbol{x}$.
- Continuous: When there exists C so that $\kappa(\boldsymbol{x}) \leq C < \infty$ for all $\boldsymbol{x}$.

Triangulation: 2D

Suppose the domain is a union of triangles E_m, with vertices x_i.



$$\bar{\Omega} = \bigcup_{i=1}^{M} E_m.$$

Elements and the Bilinear Form

If the domain, Ω, can be written as a disjoint union of elements, E_k,

$$\Omega = \cup_{m=1}^{M} E_m \quad \text{with} \quad E_i^\circ \cap E_j^\circ = \emptyset \text{ for } i \neq j,$$

what happens to a and g?

$$a(u, v) = \sum_{m=1}^{M} \int_{E_m} \kappa(\boldsymbol{x}) \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{x},$$

$$g(v) = \sum_{m=1}^{M} \int_{E_m} q(\boldsymbol{x}) \, v \, \mathrm{d}\boldsymbol{x}.$$

Basis Functions

Expand

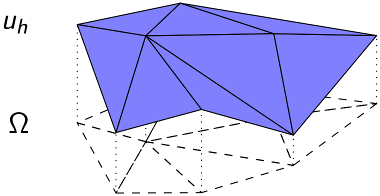
$$u_N(\boldsymbol{x}) = \sum_{i=1}^{N_p} u_i \varphi_i,$$

and plug into the weak form.

$$\sum_{j=1}^{N_p} u_j a(\varphi_j, \varphi_i) = g(\varphi_i), \quad \text{for } i = 1 \ldots N_p.$$

Global Lagrange Basis

Approximate solution u_h: Piecewise linear on Ω



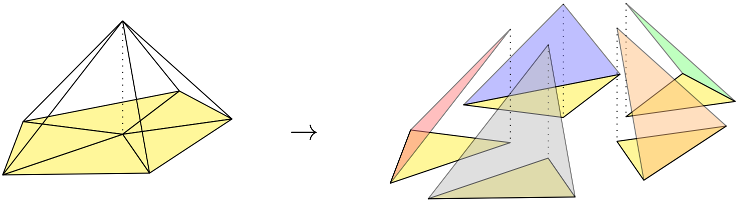
The Lagrange basis for V_h consists of piecewise linear φ_i, with...

$$\varphi_i(\boldsymbol{x}_i) = 1 \quad \text{and} \quad \varphi_i(\boldsymbol{x}_j) = 0, \quad \text{for} \quad i \neq j.$$

Basis Functions Features

Features of the basis?

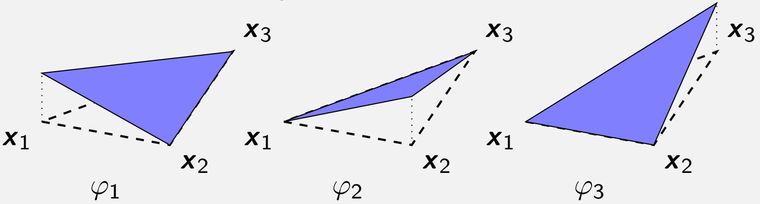
- For the piecewise linear Lagrange basis, each φ_i is continuous on Ω.
- Restricted to E_m, each φ_i is linear.



Local Basis

What basis functions exist on each triangle?

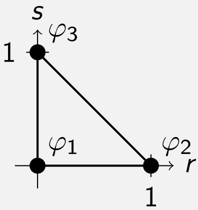
On each triangle, E_m, we have three non-zero basis functions, one for each vertex of the triangle:



In the Figure, $\varphi_1(\boldsymbol{x}_1) = 1$, $\varphi_1(\boldsymbol{x}_2) = 0$, and $\varphi_1(\boldsymbol{x}_3) = 0$.

Local Basis Expressions

Write expressions for the nodal linear basis in 2D.



- $\varphi_1(r, s) = 1 - r - s$
- $\varphi_2(r, s) = r$
- $\varphi_3(r, s) = s$

Higher-Order, Higher-Dimensional Simplex Bases

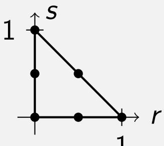
What's an n-simplex?

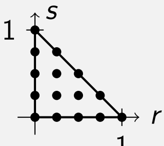
$r_i \geq 0$, $\sum r_i \leq 1$. ($\to$ barycentric) Interval, $\triangle$, tetrahedron, ...

Give a higher-order polynomial space on the n-simplex:

$$P^N := \operatorname{span}\left\{\prod_{i=1}^{d} x_i^{n_i} : \sum n_i \leq N\right\}$$

Give nodal sets (on the $\triangle$) for P^N and $\dim P^N$ in general.





$$\dim P^N = N_p = \frac{(N+1)(N+2)}{2}$$

Avoiding Runge: e.g. Warburton '06

Finding a Nodal/Lagrange Basis in General

Given a nodal set $(\xi_i)_{i=1}^{N_p} \subset \hat{E}$ (where $\hat{E}$ is the reference element) and a basis $(\varphi_j)_{j=1}^{N_p} : \hat{E} \to \mathbb{R}$, find a Lagrange basis.

Set up a Vandermonde matrix:

$$V := \begin{bmatrix} \varphi_1(\xi_1) & \cdots & \varphi_{N_p}(\xi_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\xi_{N_p}) & \cdots & \varphi_{N_p}(\xi_{N_p}) \end{bmatrix}.$$

Then $\ell_i := \sum_{j=1}^{N^p} (V^{-T})_{i,j} \varphi_j$ is a Lagrange basis.

Higher-Order, Higher-Dimensional Tensor Product Bases

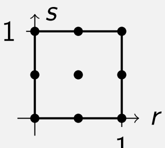
What's a tensor product element?

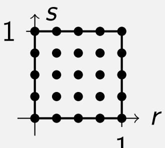
$[0,1]^n \subset \mathbb{R}^n$. Interval, quad, hexahedron.

Give a higher-order polynomial space on the n-simplex:

$$Q^N := \operatorname{span}\left\{\prod_{i=1}^{d} x_i^{n_i} : \max n_i \le N\right\}$$

Give the nodal sets (on the quad) for Q^N.



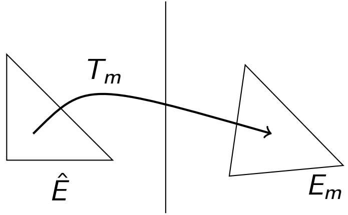


Tensor Product Elements: Lagrange Basis

Lagrange Basis for Tensor Product Elements?

> Can use tensor product of one-dimensional basis $\Rightarrow$ Lower complexity for this and many other operations.

Element Mappings



Construct a mapping $T_m : \hat{E} \to E_m$. Reference element $\hat{E}$, global $\triangle$ E_m.

$$T_m(r, s) = (\boldsymbol{x}_2 - \boldsymbol{x}_1)r + (\boldsymbol{x}_3 - \boldsymbol{x}_1)s + \boldsymbol{x}_1.$$

What is the Jacobian of T_m?

$$J_T = \begin{bmatrix} \partial x/\partial r & \partial x/\partial s \\ \partial y/\partial r & \partial y/\partial s \end{bmatrix} = \begin{bmatrix} \frac{\partial T}{\partial r} & \frac{\partial T}{\partial s} \end{bmatrix}$$
$$= \begin{bmatrix} (\boldsymbol{x}_2 - \boldsymbol{x}_1) & (\boldsymbol{x}_3 - \boldsymbol{x}_1) \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

More on Mappings

Is an affine mapping sufficient for a tensor product element?

No, because affine mappings preserve parallel lines: Global elements could only be parallelograms.
Idea: Consider a mapping $T_m \in (Q^1)^n$.

How might we accomplish curvilinear elements using the same idea?

- Use isoparametric mappings $T_m \in (P^N)^n$ (if FEM basis is P^N)
- Use subparametric mappings $T_m \in (P^M)^n$ ($M < N$ if FEM basis is P^N)
- Use superparametric mappings $T_m \in (P^M)^n$ ($M > N$ if FEM basis is P^N)

Constructing the Global Basis

Construct a basis on the element E_m from the reference basis $(\hat{\varphi}_j)_{j=1}^{N_p} : E_m \to \mathbb{R}$.

$$\varphi_{i,j}(\boldsymbol{x}) = \hat{\varphi}_j(T_m^{-1}(\boldsymbol{x})).$$

What's the gradient of this basis?

$$\begin{aligned}
\nabla_{\boldsymbol{x}}\varphi_j(T^{-1}(\boldsymbol{x})) &= \left[\frac{d}{d\boldsymbol{x}}\varphi_j(T^{-1}(\boldsymbol{x}))\right]^T \\
&= \left[\left(\frac{d\varphi_j}{d\boldsymbol{r}}\right)_{T^{-1}(\boldsymbol{x})} J_T^{-1}(\boldsymbol{x})\right]^T \\
&= J_T^{-T}(\boldsymbol{x})\nabla_{\boldsymbol{r}}\varphi_j(T^{-1}(\boldsymbol{x})).
\end{aligned}$$

Assembling a Linear System

Express the matrix and vector elements in

$$\sum_{j=1}^{N_p} u_j a(\varphi_j, \varphi_i) = g(\varphi_i) \quad \text{for } i = 1, \ldots, N_p.$$

$$a(\varphi_i, \varphi_j) = \sum_{m=1}^{M} \int_{E_m} \kappa(\boldsymbol{x}) \nabla \varphi_i \cdot \nabla \varphi_j \, \mathrm{d}\boldsymbol{x},$$

$$g(\varphi_i) = \sum_{m=1}^{M} \int_{E_m} f(\boldsymbol{x}) \varphi_i \, \mathrm{d}\boldsymbol{x}.$$

Integrals on the Reference Element

Evaluate

$$\int_E \kappa(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \varphi_i(\boldsymbol{x})^T \nabla_{\boldsymbol{x}} \varphi_j(\boldsymbol{x}) d\boldsymbol{x}.$$

$$\begin{aligned} &\int_E \kappa(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \varphi_i(\boldsymbol{x})^T \nabla_{\boldsymbol{x}} \varphi_j(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_E \kappa(\boldsymbol{x}) (J_T^{-T} \nabla_{\boldsymbol{r}} \varphi_i)^T (J_T^{-T} \nabla_{\boldsymbol{r}} \varphi_j) d\boldsymbol{x} \\ &\overset{P^1}{=} (J_T^{-T} \nabla_{\boldsymbol{r}} \varphi_i)^T (J_T^{-T} \nabla_{\boldsymbol{r}} \varphi_j) |J_T| \int_{\hat{E}} \kappa(T(\boldsymbol{r})) d\boldsymbol{r} \end{aligned}$$

And now the RHS functional.

$$\int_E f(\boldsymbol{x}) \varphi_i(\boldsymbol{x}) d\boldsymbol{x} = |J_T| \int_{\hat{E}} f(T(\boldsymbol{r})) \varphi_i(\boldsymbol{r}) d\boldsymbol{r}.$$

Inhomogeneous Dirichlet BCs

Handle an inhomogeneous boundary condition $u(\boldsymbol{x}) = \eta(\boldsymbol{x})$ on $\partial\Omega$.

- Find a function $u^0 \in H^1(\Omega)$ with boundary values $u^0(\boldsymbol{x}) = \eta(\boldsymbol{x})$ on $\partial\Omega$. ("lifted" from boundary to volume)
- Define $\hat{u} := u - u^0 \in H_0^1(\Omega)$.
- Insert $u = \hat{u} + u^0$ into the weak form:

$$a(\hat{u} + u^0, v) = a(\hat{u}, v) + a(u^0, v) = g(v),$$
$$a(\hat{u}, v) = \underbrace{g(v) - a(u^0, v)}_{\hat{g}(v):=},$$

  where still $\hat{u} \in H_0^1$.

Altogether:

- Inhomogeneous BC just leads to extra term on RHS.
- No change in function spaces.

Demo

- **Demo:** Developing FEM in 2D
- **Demo:** 2D FEM Using Firedrake
- **Demo:** Rates of Convergence

Outline

Conditions on the Mesh

Let Ω be a polygonal domain.

Admissibility (Braess, Def. II.5.1)

A partition (mesh) $\mathcal{T} = \{E_1, \ldots, E_M\}$ of Ω into triangular or quadrilateral elements is called admissible if

- $\bar{\Omega} = \bigcup_{i=1}^{M} E_i$.
- If $E_i \cap E_j$ consists of exactly one point, then it is a common vertex of E_i and E_j.
- If $E_i \cap E_j$ consists of more than one point for $i \neq j$, then $E_i \cap E_j$ is a common edge of E_i and E_j.

Give an example of a non-admissible partition.

One with a hanging node.

Mesh Resolution, Shape Regularity

Definition (Diameter)

A bounded set Ω has diameter $d(\Omega) = \sup\{|x-y| : x, y \in \Omega\}$.

Mesh Resolution

When every element of a partition has diameter at most $2h$, we write $\mathcal{T}_h$ instead of $\mathcal{T}$.

Definition (Shape Regularity (Braess, Def. II.5.1))

A family of partitions $\{\mathcal{T}_h\}$ is called shape regular if

there exists a number $\kappa > 0$ so that every $E \in \mathcal{T}_h$ contains a circle

Cone Conditions

Definition (Lipschitz Domain)

A bounded domain $\Omega \subset \mathbb{R}^n$ is called a Lipschitz domain provided that...

> for every $x \in \partial\Omega$ there exists a neighborhood of x within which $\partial\Omega$ can be represented as the graph of a Lipschitz function.

Lipschitz domains satisfy a cone condition:

> The interior angles at vertices are positive, so that a cone can be placed in Ω with its tip at the vertex.

Theorem (Rellich Selection Theorem (Braess, Thm. II.1.9))

Let $m \geq 0$, let Ω be Lipschitz. Then the imbedding $H^{m+1}(\Omega) \to H^m(\Omega)$ is compact, i.e. any bounded sequence in the range of the imbedding has a

The Interpolation Operator

Theorem (Interpolation Operator (Braess, Lemma II.6.2))

Let $\Omega \subset \mathbb{R}^2$ be Lipschitz. Let $t \geq 2$, and $z_1, z_2, \ldots, z_s$ are $s := t(t+1)/2$ prescribed points in $\bar{\Omega}$ such that the interpolation operator $I : H^t \to \mathbb{P}^{t-1}$ is well-defined. Then there exists a constant c so that for $u \in H^t(\Omega)$

$$\|u - Iu\|_{H^t} \leq c(\Omega, (z_i)) \, |u|_{H^t} \, .$$

Theorem (Approx. for Congruent $\triangle$ (Braess, Remark II.6.5))

Let $E_h := h\hat{E}$, i.e. a scaled version of a reference triangle, with $h \leq 1$. Then, for $0 \leq m \leq t$, there exists a C so that

$$\|u - Iu\|_{H^m(E_h)} \leq C h^{t-m} \, |u|_{H^t(E_h)} \, .$$

Approximation for Congruent Triangles: Proof (1/2)

Set up a function on E_h and $\hat{E}$. Work out the scaling for the derivative.

Let $u \in H^t(E_h)$. Define $v \in H^t(\hat{E})$ by $v(y) := u(hy)$.
Then $D_w^\alpha v = h^{|\alpha|} D_w^\alpha u$ for $|\alpha| \leq t$.

Work out the scaling for the Sobolev seminorm.

$$|v|^2_{H^\ell(\hat{E})} = \sum_{|\alpha|=\ell} \int_{\hat{E}} (D_w^\alpha v)^2 = \sum_{|\alpha|=\ell} \int_{E_h} h^{2\ell} (D_w^\alpha u)^2 h^{-2} = h^{2\ell-2} |u|^2_{H^\ell(E_h)}.$$

Work out the scaling for the Sobolev norm. Recall $h \leq 1$.

$$\|u\|^2_{H^m(E_h)} = \sum_{\ell \leq m} |u|^2_{H^\ell(E_h)} = \sum_{\ell \leq m} h^{-2\ell+2} |v|^2_{H^\ell(E_h)} \leq C' h^{-2m+2} \|v\|^2_{H^m(\hat{E})}.$$

Approximation for Congruent Triangles: Proof (1/2)

$$\|u - Iu\|_{H^m(E_h)} \leq Ch^{t-m} |u|_{H^t(E_h)} \quad (0 \leq m \leq t)$$

- $|v|^2_{H^\ell(\hat{E})} = |u|^2_{H^\ell(E_h)}$
- $\|u\|^2_{H^m(E_h)} \leq C' h^{-2m+2} \|v\|^2_{H^m(\hat{E})}$

Prove the estimate.

Inserting $u - Iu$ into this estimate in place of u:

$$\begin{aligned} \|u - Iu\|_{H^m(E_h)} &\leq C' h^{-m+1} \|v - Iv\|_{H^m(\hat{E})} \leq C' h^{-m+1} \|v - Iv\|_{H^t(\hat{E})} \\ &\leq C' c h^{-m+1} |v|_{H^t(\hat{E})} \leq C' c h^{t-m} |u|_{H^t(E_h)} . \end{aligned}$$

H^m Polynomial Approximation on Meshes

Definition (Broken Norm)

Given a partition $\mathcal{T}_h = \{E_i\}_{i=1}^M$ and a function u such that $u \in H^m(E_i)$,

$$\|u\|_{H^m,h} := \sqrt{\sum_{i=1}^{M} \|u\|^2_{H^m(E_i)}}.$$

Approximation Theorem (Braess, Theorem II.6.4)

Let $t \geq 2$, suppose $\mathcal{T}_h$ is a shape-regular triangulation of Ω. Then there exists a constant c such that, for $0 \leq m \leq t$ and $u \in H^t(\Omega)$,

$$\|u - I_h u\|_{H^m,h} \leq c(\Omega, \kappa, t) h^{t-m} |u|_{H^t(\Omega)},$$

Outline

Weak Forms as Minimization Problems

Let V be a linear space, and $a : V \times V \to \mathbb{R}$ a bilinear form, and $g \in V'$.

Theorem (Solutions of Weak Forms are Quadratic Form Minimizers)

If a is SPD, then

$$J(v) := \tfrac{1}{2}a(v,v) - g(v)$$

attains its minimum over V at u iff $a(u,v) = g(v)$ for all $v \in V$.

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - g(u+tv) \\ &= J(u) + t[a(u,v) - g(v)] + \frac{t^2}{2}a(v,v). \end{aligned}$$

for $u, v \in V$ and $t \in \mathbb{R}$.
If u satisfies $a(u,v) = g(v)$, $J(u+v) > J(u)$.
If J has a min at u, derivative of $t \mapsto J(u+tv)$ must vanish at $t = 0$

Example: Lagrange Multipliers in $\mathbb{R}^2$

$$
\begin{aligned}
f(x,y) = x^2 + y^2 &\rightarrow \text{min!} \\
g(x,y) = x + y &= 2
\end{aligned}
$$

Write down the Lagrangian.

$$\mathcal{L}(x,y,\lambda) = f(x,y) + \lambda g(x,y) = x^2 + y^2 + \lambda(x + y - 2).$$

Write down a necessary condition for a constrained minimum.

$$0 = \nabla \mathcal{L} = \begin{bmatrix} \nabla f + \lambda \nabla g \\ g \end{bmatrix}.$$

Saddle Point Problems

X, M Hilbert spaces. $a : X \times X \to \mathbb{R}$ and $b : X \times M \to \mathbb{R}$ continuous bilinear forms, $f \in X'$, $g \in M'$. Minimize

$$J(u) = \frac{1}{2}a(u,u) - \langle f, u \rangle \qquad \text{subject to} \qquad b(u,\mu) = \langle g, \mu \rangle \quad (\mu \in M).$$

Apply the method of the Lagrange multipliers.

$$\mathcal{L}(u,\lambda) = J(u) + [b(u,\lambda) - \langle g, \lambda \rangle] \quad (\lambda \in M).$$

- J and $\mathcal{L}(\cdot,\lambda)$ agree when constraint is satisfied.
- Idea: Select $\lambda \in M$ to 'tweak' $\mathcal{L}$ so that minimizer of $\mathcal{L}(\cdot,\lambda)$ satsifies the constraints. (Finite-dim: $-\nabla f = J_g^T \boldsymbol{\lambda}$)

Yields saddle point problem: find $(u,\lambda) \in X \times M$ so that

$$\begin{aligned} a(u,v) + b(v,\lambda) &= \langle f, v \rangle \quad (v \in X), \\ b(u,\mu) &= \langle g, \mu \rangle \quad (\mu \in M). \end{aligned}$$

Example: Saddle Point Problem in $\mathbb{R}^2$

$$\begin{aligned} f(x,y) = x^2 + y^2 \quad &\to \quad \text{min!} \\ g(x,y) = x + y \quad &= \quad 2 \end{aligned}$$

Lagrangian: $\mathcal{L}(x,y,\lambda) = f(x,y) + \lambda g(x,y) = x^2 + y^2 + \lambda(x + y - 2)$.

Show that $x = y = 1$, $\lambda = -2$ is a saddle point.

The Hessian has the form

$$\mathcal{H}_{\mathcal{L}} = \begin{bmatrix} H_f & \nabla g \\ \nabla g^T & 0 \end{bmatrix}.$$

$$\mathcal{H}_{\mathcal{L}} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} = M \begin{bmatrix} A & \\ & -BA^{-1}B^T \end{bmatrix} M^T,$$

demonstrating indefiniteness using Sylvester's Law of Inertia.
(cf. Benzi et al. '05, Section 3.4)

Stokes Equation

$$\begin{aligned} \Delta \boldsymbol{u} + \nabla p &= -\boldsymbol{f} \quad (x \in \Omega), \\ \nabla \cdot \boldsymbol{u} &= 0 \quad (x \in \Omega), \\ \boldsymbol{u} &= \boldsymbol{u}_0 \quad (x \in \partial\Omega). \end{aligned}$$

What are the pieces?

- $\boldsymbol{u}$ is the velocity field,
- p is the pressure,
- $\boldsymbol{f}$ is an externally applied force field,
- Pressure gradient gives rise to an additional force that prevents a density change.
- $\nabla \cdot \boldsymbol{u} = 0$ is the incompressibility constraint:
  Pressure falls/rises where a source/sink would be created.

Stokes: Properties

$$\begin{aligned} \Delta \boldsymbol{u} + \nabla p &= -\boldsymbol{f} \quad (x \in \Omega), \\ \nabla \cdot \boldsymbol{u} &= 0 \quad (x \in \Omega), \\ \boldsymbol{u} &= \boldsymbol{u}_0 \quad (x \in \partial\Omega). \end{aligned}$$

Can we choose any $\boldsymbol{u}_0$?

$$\int_{\partial\Omega} \boldsymbol{u}_0 \cdot \hat{\boldsymbol{n}} dS_{\boldsymbol{x}} = \int_{\partial\Omega} \boldsymbol{u} \cdot \hat{\boldsymbol{n}} dS_{\boldsymbol{x}} = \int_{\Omega} \nabla \cdot \boldsymbol{u} d\boldsymbol{x} = 0$$

is a compatibility condition. Satisfied e.g. for $\boldsymbol{u}_0 \equiv 0$.

Does Stokes fully determine the pressure?

Only up to an additive constant. Additionally demand $\int_\Omega p dx = 0$.

Stokes: Variational Formulation

$$\Delta \boldsymbol{u} + \nabla p = -\boldsymbol{f}, \qquad \nabla \cdot \boldsymbol{u} = 0 \quad (x \in \partial\Omega).$$

Choose some function spaces (for homogeneous $\boldsymbol{u}_0 = 0$).

$$X = H_0^1(\Omega)^n, \qquad M = L_0^2(\Omega) := \left\{ q \in L^2(\Omega) : \int_\Omega q dx = 0 \right\}$$

Derive a weak form.

$$a(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega J_{\boldsymbol{u}} : J_{\boldsymbol{v}}, \qquad b(\boldsymbol{v}, q) = \int_\Omega \nabla \cdot \boldsymbol{v} q,$$

$A : B = \mathrm{tr}(AB^T) = \sum_{i,j} A_{i,j} B_{i,j}$. Find $(\boldsymbol{u}, p) \in X \times M$ so that

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) &= \langle \boldsymbol{f}, \boldsymbol{v} \rangle_{L^2} \quad (\boldsymbol{v} \in X), \\ b(\boldsymbol{u}, q) &= 0 \quad (q \in M), \end{aligned}$$

where in reusing b, we used that $(-\operatorname{div})^* = \operatorname{grad}$ are adjoint.

Solvability of Saddle Point Problems

The Stokes weak form is clearly in saddle-point form.
Do all saddle point problems have unique solutions?

$$
\begin{aligned}
f(x,y) = x^2 + y^2 &\rightarrow \text{min!},\\
x + y &= 2,\\
3x + 3y &= 6.
\end{aligned}
$$

$\mathcal{L}(x,y,\lambda) = x^2 + y^2 + \lambda(x + y - 2) + \mu(3x + 3y - 6)$. (λ,μ) no longer uniquely determined.

$\rightarrow$ Need a criterion.

The inf-sup Condition

$$\begin{aligned} a(u,v) + b(v,\lambda) &= \langle f, v\rangle \quad (v \in X), \\ b(u,\mu) &= \langle g, \mu\rangle \quad (\mu \in M). \end{aligned}$$

Theorem (Brezzi's splitting theorem (Braess, III.4.3))

The saddle point problem has a unique solution if and only if

- *The bilinear form $a(\cdot,\cdot)$ is V-elliptic, where $V = \{u : b(u,\mu) = 0 \text{for all } \mu \in M\}$, i.e. there exists $c_0 > 0$ so that*

$$a(v,v) \geq c_0 \|v\|_X^2 \qquad (v \in V).$$

- *There exists a constant $c_2 > 0$ so that (inf-sup or LBB condition):*

$$\inf_{\mu \in M} \sup_{v \in X} \frac{b(v,\mu)}{\|v\|_X \|\mu\|_M} \geq c_2.$$

Interpreting the inf-sup Condition

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} = M \begin{bmatrix} A & \\ & -BA^{-1}B^T \end{bmatrix} M^T$$

$$a(v, v) \geq c_0 \|v\|_X^2, \qquad \inf_{\mu \in M} \sup_{v \in X} \frac{b(v, \mu)}{\|v\|_X \|\mu\|_M} \geq c_2.$$

For any given v, can we expect $b(v, \mu)$ to be nonzero for all μ?

No! E.g. for Stokes, the B block is short-and-fat $\Rightarrow \exists$ nullspace.

What is the inf-sup condition saying?

"b has no μ-nullspace."

Why does it suffice for a to be V-elliptic?

True in the linear algebra, too! (Think Schur complements.) (Benzi et al. '05, Thm. 3.2)

inf-sup and Stokes

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) &= \int_\Omega J_{\boldsymbol{u}} : J_{\boldsymbol{v}}, \qquad \text{where } A : B = \operatorname{tr}(AB^T), \\ b(\boldsymbol{v}, q) &= \int_\Omega \nabla \cdot \boldsymbol{v} q. \end{aligned}$$

Find $(\boldsymbol{u}, p) \in X \times M$ so that

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) &= \langle \boldsymbol{f}, \boldsymbol{v} \rangle_{L^2} \quad (\boldsymbol{v} \in X), \\ b(\boldsymbol{u}, q) &= 0 \quad (q \in M). \end{aligned}$$

Theorem (Existence and Uniqueness for Stokes (Braess, III.6.5))

There exists a unique solution of this system when $\boldsymbol{f} \in H^{-1}(\Omega)^n$.

(based on results due to Ladyšenskaya, Nečas)

Discretizations for Stokes

Demo: 2D Stokes Using Firedrake (P^1-P^1)

Give a heuristic reason why P^1-P^1 might not be great.

The differential operators being applied to $\boldsymbol{u}$ and p in the Stokes system are of different order.

Demo: Bad Discretizations for 2D Stokes

Establishing a Discrete inf-sup Condition

Suppose $b : X \times M \to \mathbb{R}$ satisfies inf-sup. Subspaces $X_h \subseteq X$, $M_h \subseteq M$.

Fortin's Criterion ([Fortin 1977])

Suppose there exists a bounded projector $\Pi_h : X \to X_h$ so that

$$b(v - \Pi_h v, \mu_h) = 0 \quad (\mu_h \in M_h).$$

If $\|\Pi_h\| \leq c$ for some constant c independent of h, then b satisfies the inf-sup-condition on $X_h \times M_h$.

Let $\mu_h \in M_h$. By assumption, $b(v, \mu_h) = b(\Pi_h v, \mu_h)$ for $v \in X$.

$$\sup_{v_h \in X_h} \frac{b(v_h, \mu_h)}{\|v_h\|} \geq \sup_{v_h \in \Pi_h X} \frac{b(v_h, \mu_h)}{\|v_h\|} = \sup_{v \in X} \frac{b(\Pi_h v, \mu_h)}{\|\Pi_h v\|}$$

$$\geq \frac{1}{c} \sup_{v \in X} \frac{b(v, \mu_h)}{\|v\|} \geq c_2 \|\mu_h\|.$$

H^1-Boundedness of the L^2-Projector

Assume H^2-regularity and a uniform triangulations $\mathcal{T}_h$. (Not in general!)

H^1-Boundedness of the L^2-Projector (Braess Corollary II.7.8)

Let π_h^0 be the L_2-projector onto a finite element space $V_h \subset H^1(\Omega)$. Then, for an h-independent constant c,

$$\left\|\pi_h^0 v\right\|_{H^1} \leq c \left\|v\right\|_{H^1} .$$

Ingredients?

- Regularity
- Aubin-Nitsche
- Inverse estimates (For affine, pw. polynomial family V_h: $\left\|v_h\right\|_{H^t,h} \leq Ch^{m-t} \left\|v_h\right\|_{H^m,h}$ with $0 \leq m \leq t$, e.g. $\left\|v_h\right\|_{H^1,h} \leq Ch^{-1} \left\|v_h\right\|_{L^2,h}$)

H^1-Boundedness of the L^2-Projector

Does H^1 boundedness of the H^1 projector hold?

Yes, any Hilbert space projection is bounded. (Pythagoras)

How would this break down without the uniformity assumption?

On a graded mesh, where L^2 projection introduces $O(1/h)$ growth in the H^1 seminorm (which measures oscillation, in a way).

Bubbles and the MINI Element

What is a bubble function?

$\varphi_b(r,s) = rs(1-r-s)$. (see figure on next slide)

Let B^3 be the span of the bubble function and $\mathcal{T}_h$ the triangulation.

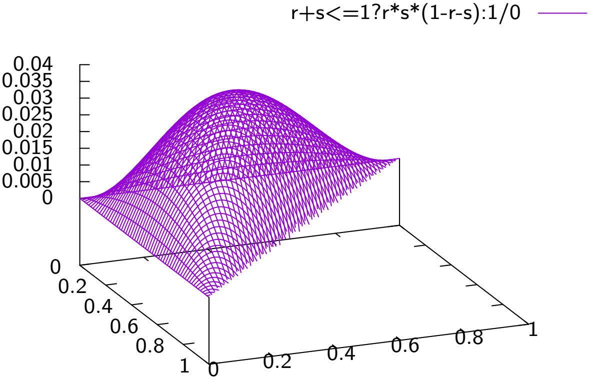
Define the MINI variational space $X_h \times M_h$.

$$
\begin{aligned}
X_h &:= \left\{ v_h \in C(\bar{\Omega})^2 \cap H_0^1(\Omega)^2 : v_h|_E \in (P^1 \oplus B^3)^2 \text{ for } E \in \mathcal{T}_h \right\} \\
M_h &:= \left\{ q_h \in C(\bar{\Omega}) \cap L_0^2(\Omega) : v_h|_E \in P^1 \text{ for } E \in \mathcal{T}_h \right\}
\end{aligned}
$$

Computational impact of the bubble DOF?

Not coupled to DOFs outside the element; can use static condensation to eliminate.

The Bubble in Pictures



MINI Satisifies an inf-sup Condition (1/4)

MINI satisifes inf-sup (Braess Theorem III.7.2)

Assume Ω is convex or has a smooth boundary. Then the MINI variational space satisfies an inf-sup condition for every variational form that itself satisfies one.

Assume uniform meshes (can generalize). Let

$$\mathcal{M}_h := \left\{ v_h \in C(\bar{\Omega}) \cap H_0^1(\Omega) : v_h|_E \in P^1 \text{ for } E \in \mathcal{T}_h \right\}.$$

Let $\pi_h^0 : H_0^1 \to \mathcal{M}_h$ be the L^2 projector.
Then $\left\|\pi_h^0 v\right\|_{H^1} \leq c_1 \left\|v\right\|_{H^1}$ from its H^1-boundedness and, from the interpolation estimate,

$$\begin{aligned}\left\|v - \pi_h^0 v\right\|_{L^2} &\leq \left\|v - \mathcal{I}v\right\|_{L^2} + \left\|\mathcal{I}v - \pi_h^0 v\right\|_{L^2} \\ &= \left\|v - \mathcal{I}v\right\|_{L^2} + \left\|\pi_h^0(\mathcal{I}v - v)\right\|_{L^2} \leq c_2 h \left|v\right|_{H^1}.\end{aligned}$$

MINI Satisifies an inf-sup Condition (2/4)

Create a projector onto the bubble space B^3.

Let $\pi_h^1 : L^2 \to B^3$ be linear so that

$$\int_E (\pi_h^1 v - v) dx = 0 \quad \text{for } E \in \mathcal{T}_h.$$

What does this bubble projector do?

- Project onto piecewise constant functions.
- Replace the constant by a bubble with the same integral.

Do we have an estimate for the bubble projector?

$$\left\| \pi_h^1 v \right\|_{L^2} \leq c_3 \left\| v \right\|_{L^2}.$$

MINI Satisifies an inf-sup Condition (3/4)

Make an overall projector Π_h onto X_h.

Define $\Pi_h v := \pi_h^0 v + \pi_h^1(v - \pi_h^0 v)$. By construction, Π_h preserves the constant mode, i.e. $\int(\Pi_h v - v)dx = 0$.

Show Fortin's criterion for Π_h.

Extend Π_h to vector-valued component-by-component.
$q_h \in M_h$ is continuous, so we may apply Gauss's theorem.

$$\begin{aligned} & b(\boldsymbol{v} - \Pi_h \boldsymbol{v}, q_h) \\ = & \int \nabla \cdot (\boldsymbol{v} - \Pi_h \boldsymbol{v}) q_h dx \\ = & \int_{\partial\Omega} (\boldsymbol{v} - \Pi_h \boldsymbol{v}) \cdot \hat{\boldsymbol{n}} \underbrace{q_h}_{0} dS_x - \int_{\Omega} (v - \Pi_h v) \cdot \underbrace{\nabla q_h}_{\text{const}} dx = 0. \end{aligned}$$

MINI Satisifies an inf-sup Condition (4/4)

- $\left\|\pi_h^0 v\right\|_{H^1} \leq c_1 \|v\|_{H^1}$ for L^2 projector $\pi_h^0 : H_0^1 \to \mathcal{M}_h$.
- $\left\|v - \pi_h^0 v\right\|_{L^2} \leq c_2 h \, |v|_{H^1}$.
- $\left\|\pi_h^1 v\right\|_{L^2} \leq c_3 \|v\|_{L^2}$.

Show H^1-boundedness of Π_h.

$$
\begin{aligned}
\|\Pi_h v\|_{H^1} &\leq \left\|\pi_h^0 v\right\|_{H^1} + \left\|\pi_h^1 (v - \pi_h^0 v)\right\|_{H^1} \\
&\underset{\text{inv.est.}}{\leq} c_1 \|v\|_{H^1} + c_4 h^{-1} \left\|\pi_h^1 (v - \pi_h^0 v)\right\|_{L^2} \\
&\leq c_1 \|v\|_{H^1} + c_4 h^{-1} c_3 \left\|v - \pi_h^0 v\right\|_{L^2} \\
&\leq c_1 \|v\|_{H^1} + c_4 c_3 c_2 \|v\|_{H^1} .
\end{aligned}
$$

Demo

Demo: 2D Stokes Using Firedrake (MINI and Taylor-Hood)

Outline

Lax-Milgram, General Case

Let V be Hilbert space with inner product $\langle \cdot, \cdot \rangle$.

Theorem (Lax-Milgram, General Case)

Let a be a V-elliptic bilinear form, and let g be a bounded linear functional on V.
Then there exists a unique $u \in V$ so that $a(u, v) = g(v)$ for all $v \in V$.

Let $u \in V$ and observe $a_u(v) := a(u, v)$ is a bounded linear functional (due to continuity of a). Let $t_u \in V$ be the Riesz representer of a_u with $a_u(v) = \langle v, t_u \rangle$ for all $v \in V$. Consider the mapping defined by that:

$$T : V \to V, \qquad u \mapsto Tu := t_u.$$

We show that T is linear, bounded, has closed range, and is onto V.

Lax-Milgram Proof (2/5)

$a(u, v) = \langle v, Tu \rangle$. Show linearity of T.

For $u, v, w \in V$ and $\alpha \in \mathbb{R}$:

$$\langle v, T(\alpha u + w) \rangle = a(\alpha u + w, v) = \alpha \langle v, Tu \rangle + \langle v, Tw \rangle .$$

Show boundedness $\Leftrightarrow$ continuity of T.

$$\|Tu\|^2 = \langle Tu, Tu \rangle = a_u(Tu) = a(u, Tu) \leq c_1 \|Tu\| \|u\| \quad \text{(continuity)}.$$

Lax-Milgram Proof (3/5)

$a(u, v) = \langle v, Tu\rangle$. Show that T has closed range. (Needed for Hilbert projection, which is needed for onto.)

Let $z_n = Tu_n$ be a sequence in range(T). By definition, $a(u_n, v) = \langle v, Tu_n\rangle = \langle v, z_n\rangle$ for all $v \in V$, so that

$$\begin{aligned} a(u_n - u_m, v) &= \langle v, z_n - z_m\rangle \\ \Rightarrow \quad a(u_n - u_m, u_n - u_m) &= \langle u_n - u_m, z_n - z_m\rangle \\ \Rightarrow \quad c_0 \|u_n - u_m\|^2 &\le \|u_n - u_m\| \, \|z_n - z_m\| \quad \text{(coercivity)} \\ \Rightarrow \quad c_0 \|u_n - u_m\| &\le \|z_n - z_m\| . \end{aligned}$$

If $z_n \to z$, (u_n) must be Cauchy, so has a limit (because V is Hilbert). Let u be the limit. Next: Show $z = Tu$.

Let $v \in V$ be arbitrary. $a(u_n, v) \to a(u, v)$ by continuity. Also: $|\langle Tu_n - z, v\rangle| \to 0$, so that $\langle v, Tu_n\rangle \to \langle v, z\rangle$, so $a(u, v) = \langle v, z\rangle$, and by definition of T, $z = Tu$.

Lax-Milgram Proof (4/5)

$a(u, v) = \langle v, Tu \rangle$. Show that T is onto V.

Suppose not. By the Hilbert projection theorem, there exists $w \in \text{range}(T)^{\perp} \setminus \{0\}$. Therefore $\langle w, Tu \rangle = 0$ for all $u \in V$. Choosing $u = w$ gives $0 = \langle w, Tw \rangle = a(w, w)$, a contradiction.

Lax-Milgram Proof (5/5)

Show existence of the solution u.

Let z be the Riesz representer of g: $g(v) = \langle v, z \rangle$ for all $v \in V$. Since $T : V \to V$ is onto, there exists a $u \in V$ so that $z = Tu$, i.e. $g(v) = \langle v, Tu \rangle = a(u, v)$ for all $v \in V$.

Show uniqueness of the solution u.

Suppose we have a second $\hat{u}$ with $z = T\hat{u}$. Then $a(u - \hat{u}, v) = 0$ for all $v \in V$, particularly $a(u - \hat{u}, u - \hat{u}) = 0$, i.e. $u = \hat{u}$.

Outline

Outline

Conservation laws

Goal: Solve *conservation laws* on bounded domain $\Omega \subset \mathbb{R}^n$:

$$\boldsymbol{q}_t + \nabla \cdot \boldsymbol{F}(\boldsymbol{q}) = 0$$

Example: Maxwell's Equations

$$\partial_t \boldsymbol{D} - \nabla \times \boldsymbol{H} = -\boldsymbol{j}, \qquad \partial_t \boldsymbol{B} + \nabla \times \boldsymbol{E} = 0,$$

$$\nabla \cdot \boldsymbol{D} = \rho, \qquad \nabla \cdot \boldsymbol{B} = 0.$$

What do we do with the divergence constraints?

Ignore them. If satisfied at initial condition, they continue to be satisfied.

Rewriting Maxwell's

Let $\boldsymbol{q} = (D_x, D_y, D_z, B_x, B_y, B_z)^T$. Consider $\boldsymbol{D} = \epsilon\boldsymbol{E}$ and $\boldsymbol{B} = \mu\boldsymbol{H}$.

$$\partial_t \boldsymbol{D} - \nabla \times \boldsymbol{H} = -0, \qquad \partial_t \boldsymbol{B} + \nabla \times \boldsymbol{E} = 0.$$

Rewrite in conservation law form: $\boldsymbol{q}_t + \nabla \cdot F(\boldsymbol{q}) = 0$

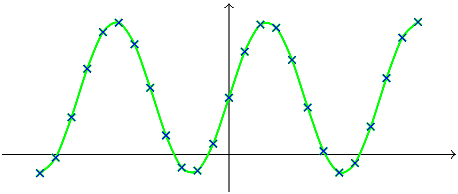
$$\boldsymbol{q}_t + \nabla \cdot \begin{pmatrix} 0 & -\frac{B_z}{\varepsilon} & \frac{B_y}{\varepsilon} \\ \frac{B_z}{\varepsilon} & 0 & -\frac{B_x}{\varepsilon} \\ -\frac{B_y}{\varepsilon} & \frac{B_x}{\varepsilon} & 0 \\ 0 & \frac{D_z}{\mu} & -\frac{D_y}{\mu} \\ -\frac{D_z}{\mu} & 0 & \frac{D_x}{\mu} \\ \frac{D_y}{\mu} & -\frac{D_x}{\mu} & 0 \end{pmatrix} = 0$$

Could we also define $\boldsymbol{q} = (E_x, E_y, E_z, H_x, H_y, H_z)^T$?

No: coeff. on the wrong side of the $\nabla\cdot$. Only OK for constant-coeff.

Outline

Solving $q_t + aq_x = 0$: Finite Differences

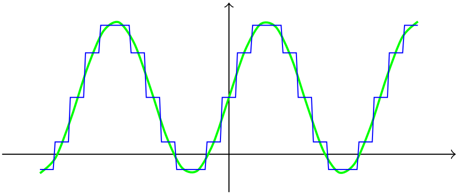


- ⊕ Simple to implement
- ⊕ High-order
- ⊕ Local and explicit in time
- ⊕ Theory available
- ⊖ High-order/geometry: pick one.
  - Upwind/downwind differencing?
    - How about in a system?
  - Boundaries?
  - Discontinuities?

$$D_t^- + aD_x^- = 0$$

$$D_t^+ f := \frac{f(t+\Delta t) - f(t)}{\Delta t}$$

Solving $q_t + aq_x = 0$: Finite Volume



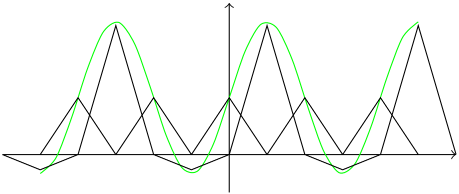
- ⊕ Robust, fast, good for c.laws
- ⊕ Local and explicit in time
- ⊕ Solid theory
- ⊖ High-order/geometry: pick one.

$$\bar{q}_k := \int_{(k-1/2)\Delta x}^{(k+1/2)\Delta x} q(x)dx$$

$$\Delta x \partial_t \bar{q}_k + f^{k+1/2} - f^{k-1/2} = 0$$

$f^{k\pm 1/2}$: flux "reconstructions"

Solving $q_t + aq_x = 0$: Finite Elements

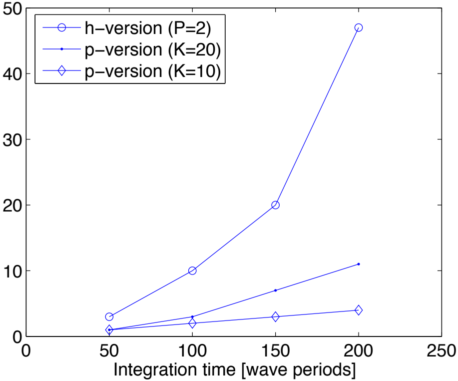


- ⊕ High-order
- ⊕ geom. flexible
- ⊕ Non-local and implicit in time
- ⊕ Solid theory
- ⊖ Not nonlinearly robust
- ⊖ Not fast: Mass matrix solve

$$\int_\Omega q_t^N \phi + a q_x^N \phi dx = 0$$

for ϕ in a test space.

Do we really want high order?



Time to compute solution at 5% error

Big assumption?

> Spectral expansion of solution decays quickly
> (i.e. solution smooth)

Figure from talk by Jan Hesthaven

Summarizing

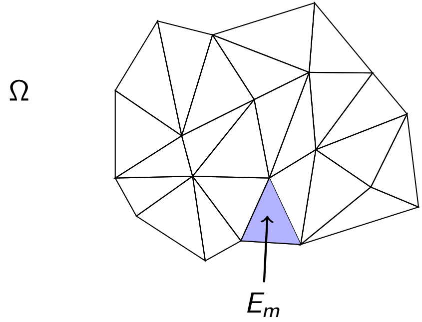
Want flexibility of finite elements *without* the drawbacks.

Let's redevelop finite elements, with a bit more care.
Strategy:

- Use n-dimensional POV for a while to expose geometric issues more clearly.
- Reduce to 1D when necessary.
- Mop up remaining issues later.

Outline

Developing the Scheme



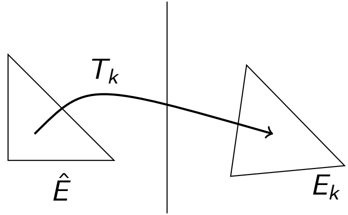
What do do about unbounded domains?

Need to truncate domain, e.g.:

- Special boundary conditions (e.g. Engquist/Majda '77, Hagstrom/Warburton '04)
- Perfectly Matched Layers (PMLs, Berenger '94)

Dealing with the Mesh, Part I

For each cell E_k, find a ref-to-global map T_k:



$$T_k : \hat{E} \to E_k$$
$$\boldsymbol{x} = (x, y, z) = T_k(r, s, t) = T_k(\boldsymbol{r})$$

- T_k affine for straight-sided simplices: $T_k(\boldsymbol{r}) = A\boldsymbol{r} + \boldsymbol{b}$
- Curved elements also possible: iso/sub/super-parametric

Dealing with the Mesh, Part II

Based on knowledge of how to do this on $\hat{E}$:

Can now *integrate* on Ω:

$$\int_\Omega f dx = \sum_{E_k} \int_{E_k} f dx = \sum_{E_k} \int_{\hat{E}} f \left| \frac{d\boldsymbol{x}}{d\boldsymbol{r}} \right| dr$$

and *differentiate* on Ω:

$$\frac{\partial f}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{x}} \frac{\partial f}{\partial \boldsymbol{r}}$$

Jacobian of T_k^{-1}?

$$\frac{d\boldsymbol{x}}{d\boldsymbol{r}} \frac{d\boldsymbol{r}}{d\boldsymbol{x}} = \mathrm{Id} \quad \Leftrightarrow \quad \left(\frac{d\boldsymbol{x}}{d\boldsymbol{r}} \right)^{-1} = \frac{d\boldsymbol{r}}{d\boldsymbol{x}}$$

Dealing with the Mesh, Part III

Approximation basis set on E_k?

Use the one we have on $\hat{E}$:

$$\phi_i^k(x) := \phi_i(T_k^{-1}(x))$$

What function space do we get if Ψ is non-affine?

- A basis of rational functions.
- Approximation results nontrivial.

Going Galerkin

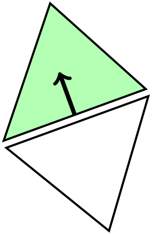
$$\int_{E_k} q_t^k \phi + (\nabla \cdot F^k)\phi dx = 0$$

Integrate by parts:

$$0 = \int_{E_k} q_t^k \phi dx - \int_{E_k} F^k \cdot \nabla \phi dx + \int_{\partial E_k} (F^k \cdot \hat{\boldsymbol{n}})\phi dx$$

Problem?

- Problem: Two values to choose from on boundary.
- Don't choose (for now).
- Call chosen answer numerical flux $(F^k \cdot n)^*$
- Feel vaguely reminded of finite volume



Strong-Form DG

Weak form:

$$0 = \int_{E_k} q_t^k \phi dx - \int_{E_k} F^k \cdot \nabla \phi dx + \int_{\partial E_k} (F^k \cdot \hat{\boldsymbol{n}})^* \phi dx$$

Integrate by parts *again*:

$$0 = \int_{E_k} q_t^k \phi + (\nabla \cdot F^k) \phi dx + \int_{\partial E_k} (F^k \cdot \hat{\boldsymbol{n}})^* - (F^k \cdot \hat{\boldsymbol{n}})^- \phi dx$$

- Strong-form DG
- Same solution as weak for linear, constant-coefficient problems.

Outline

Accuracy and Stabillity

In DG: what provides accuracy? what provides stability?

- Local approximation space provides *accuracy*
- Fluxes provide *stability*

Lax equivalence: Accuracy + Stability = Convergence

$\rightarrow$ Let flux choice be guided by stability.

Following slides based on material by Tim Warburton

Stability: Basic Setup (1/2)

$$0 = \int_{E_k} q_t^k \phi dx - \int_{E_k} F^k \cdot \nabla \phi dx + \int_{\partial E_k} (F^k \cdot \hat{\boldsymbol{n}}) \phi dS_x$$

Trick: Set $\phi = q$. Specialize $F(u) := (au, 0, 0)^T = a e_x u$.

$$\begin{aligned} 0 &= \int_{E_k} q_t^k q_k dx - \int_{E_k} a q_k e_x \cdot \nabla q_k dx + \int_{\partial E_k} (a q_k e_x \cdot \hat{\boldsymbol{n}})^* q_k dS_x \\ &= \int_{E_k} q_t^k q_k dx - \int_{E_k} a q_k \partial_x q_k dx + \int_{\partial E_k} (a q_k n_x)^* q_k dS_x \\ &= \frac{\partial_t}{2} \int_{E_k} q_k q_k dx - \int_{E_k} a q_k \partial_x q_k dx + \int_{\partial E_k} (a q_k n_x)^* q_k dS_x \end{aligned}$$

$$\Rightarrow \frac{\partial_t \|q_k\|_{2,E_k}^2}{2} = \int_{E_k} a q_k \partial_x q_k dx - \int_{\partial E_k} (a q_k n_x)^* q_k dS_x \overset{!}{\leq} 0$$

Stability: Basic Setup (2/2)

$$\frac{\partial_t \|q_k\|^2_{2,E_k}}{2} = \int_{E_k} a q_k \partial_x q_k dx - \int_{\partial E_k} (a q_k n_x)^* q_k dS_x$$

Integrate by parts:

$$\int f \partial_x f = -\int f \partial_x f + \int_{\partial} f^2 n_x$$

to see:

$$\frac{\partial_t \|q_k\|^2_{2,E_k}}{2} = \int_{\partial E_k} \frac{a(q_k)^2 n_x}{2} - (a q_k n_x)^* q_k dS_x$$

This depends on neighbors–end of element-local analysis!

Stability: Going Global

$$\frac{\partial_t \|q_k\|^2_{2,E_k}}{2} = \int_{\partial E_k} \frac{a(q_k)^2 n_x}{2} - (a q_k n_x)^* q_k \, dS_x$$

$$\begin{aligned}\frac{\partial_t \|q_k\|^2_{2,\Omega}}{2} &= \sum_k \int_{\partial E_k} \frac{a(q_k)^2 n_x}{2} - (a q_k n_x)^* q_k \, dS_x \\ &= \sum_{f \in \text{faces}} \left(\int_f \frac{a(q_k^+)^2 n_x^+}{2} - (a q_k n_x)^*_+ q_k^+ \, dS_x \right. \\ &\qquad \left. + \int_f \frac{a(q_k^-)^2 n_x^-}{2} - (a q_k n_x)^*_- q_k^- \, dS_x \right)\end{aligned}$$

- Assumption: $(a q_k n_x)^*_+ + (a q_k n_x)^*_- = 0$ ("no accumulation on interface")
- a is constant

Gather up

$$\frac{\partial_t \|q_k\|^2_{2,\Omega}}{2} = \sum_{f\in\text{faces}} \left(\int_f \frac{a(q_k^+)^2 n_x^+}{2} - (a q_k n_x)^*_+ q_k^+ dS_x \right.$$
$$\left. + \int_f \frac{a(q_k^-)^2 n_x^-}{2} - (a q_k n_x)^*_- q_k^- dS_x \right)$$

$$\frac{\partial_t \|q_k\|^2_{2,\Omega}}{2} = \sum_{f\in\text{faces}} \int_f a n_x^- \frac{(q_k^-)^2 - (q_k^+)^2}{2} - (a q_k n_x)^*_- (q_k^- - q_k^+) dS_x$$
$$= \sum_{f\in\text{faces}} \int_f \left(a n_x^- \frac{q_k^- + q_k^+}{2} - (a q_k n_x)^*_- \right) (q_k^- - q_k^+) dS_x$$

Want all that non-positive. So demand:

$$\left(a n_x^- \frac{q_k^- + q_k^+}{2} - (a q_k n_x)^*_- \right) (q_k^- - q_k^+) \overset{!}{\leq} 0$$

Picking a Flux

Want:

$$(*) = \left(a n_x^- \frac{q_k^- + q_k^+}{2} - (a q_k n_x)_-^* \right) (q_k^- - q_k^+) \overset{!}{\leq} 0$$

Ideas?

One possible choice:

$$(a q_k n_x)_-^* := a n_x^- \frac{q_k^- + q_k^+}{2}$$

- Called the *central flux*.
- Observe: $(*) = 0 \Rightarrow L^2$-norm exactly conserved!
- The lazy man's flux.
- Works.
- Problematic! Why?

Picking a flux, attempt two

Want:

$$(*) = \left(a n_x^- \frac{q_k^- + q_k^+}{2} - (a q_k n_x)_-^* \right) (q_k^- - q_k^+) \overset{!}{\leq} 0$$

More ideas?

$$(a q_k n_x)_-^* := a n_x^- \frac{q_k^- + q_k^+}{2} + \alpha \frac{q_k^- - q_k^+}{2}$$

(with $\alpha \geq 0$)

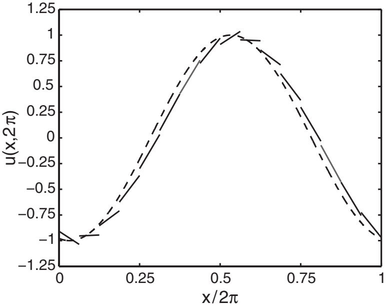
Unit considerations suggest: $\alpha = \pm a n_x^- \overset{!}{\geq} 0$.

Called the upwind flux (aka local L-F)

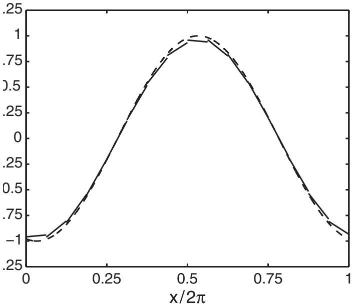
- Observe: $(*) < 0 \Rightarrow$ dissipative!
- Quite good in practice.

Comparing Fluxes (1/3)

Central



Upwind



Upwind penalizes jumps!

Figure from talk by Jan Hesthaven

Comparing Fluxes (2/3)

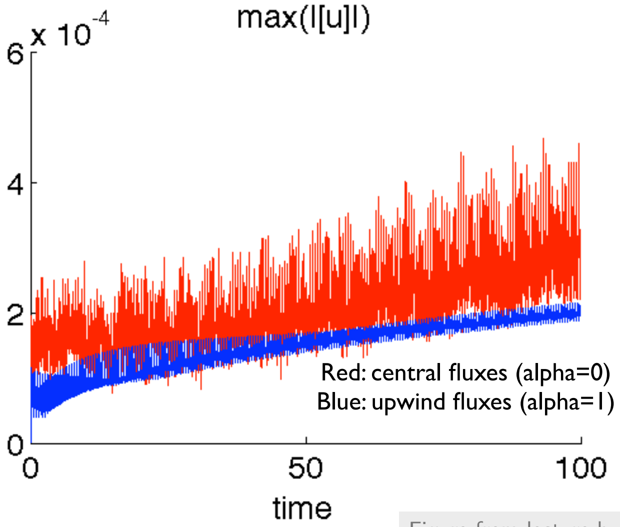


Figure from lecture by Tim Warburton

Comparing Fluxes (3/3)

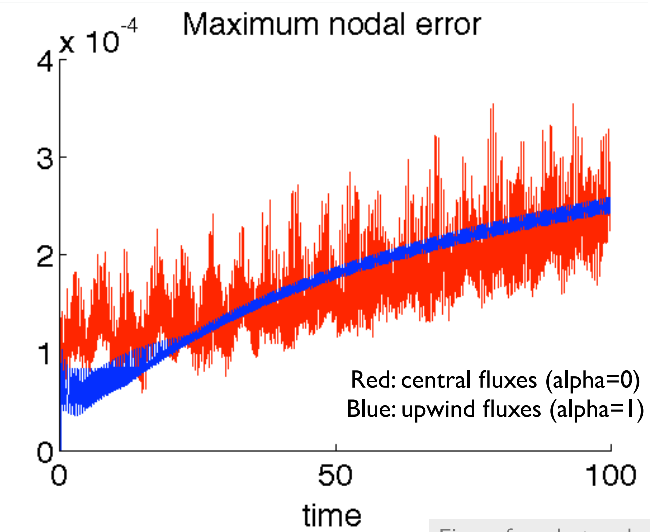


Figure from lecture by Tim Warburton

Stability Analysis

Clif notes on flux choice?

> ‘Pick the average’ or ‘pick the upwind value’

Swept under the rug: Boundary conditions

> Also important for stability!

Element coupling (and BCs) done *weakly*

> - Numerical solution really is discontinuous
> - Hence “discontinuous Galerkin”

Accuracy

Stability: (preliminary version) done!
Accuracy: Depends on approximation properties!

Need approximation space: polynomials of (total) degree at most N on the reference element.

So, expect h^{N+1} residual.

Practically often true. Theoretically:

- Lesaint, Raviart '74:
  - h^N in the general case
  - h^{N+1} for special grids
- Johnson '86: $h^{N+1/2}$

Systems of Conservation Laws

What to do about systems?

- Consider Riemann (jump) problem
  - Obtain 'fan' of different wave speeds
- *Rankine-Hugoniot condition*:

$$[\![F(q)]\!] = (\text{wave speed})\,[\![q]\!]$$

- Number states across fan $q_0, q_{-1}, q_1, \ldots$
- Set up Rankine-Hugoniot at each state boundary
- Solve for rest-state flux $F(q_0)$
- Just like Finite Volume

What about multiple dimensions?

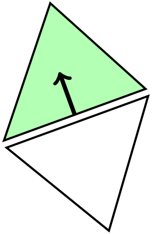
We've dealt with 1D systems.

How about the move to multiple dimensions?

In principle there is (almost) nothing to see.

Recipe:

- Reduce nD c.law to 1D c.law across boundary
- Diagonalize
- Play Rankine-Hugoniot game as before
- Transform back



Simultaneous Diagonalization

2D second-order wave equation across a boundary with normal n:

$$q_t + \begin{pmatrix} 0 & -c\, n_x & -c\, n_y \\ -c\, n_x & 0 & 0 \\ -c\, n_y & 0 & 0 \end{pmatrix} \partial_n q = 0$$

Must simultaneously diagonalize for all $(n_x, n_y)^T$ to obtain generic expression!
More symbolically:

$$q_t + (An_x)\partial_x q + (Bn_y)\partial_y q$$

Need to find matrix S that simultaneously diagonalizes An_x and Bn_y!

Demo: Finding Numerical Fluxes for DG (Part 1)

Jumps and Averages

Jump and average of a scalar quantity:

$$\{q\} := \frac{q^- + q^+}{2}$$
$$[\![q]\!] := q^+ \boldsymbol{n}^+ + q^- \boldsymbol{n}^-$$

Jump and average of a vector quantity:

$$\{\boldsymbol{q}\} := \frac{\boldsymbol{q}^- + \boldsymbol{q}^+}{2}$$
$$[\![\boldsymbol{q}]\!] := \boldsymbol{q}^+ \cdot \boldsymbol{n}^+ + \boldsymbol{q}^- \cdot \boldsymbol{n}^-$$

A Flux for Maxwell's

Wanted to solve Maxwell's equation in the time domain. Numerical flux?

Either look in the literature:

$$\hat{\boldsymbol{n}} \cdot (\boldsymbol{F}_N - \boldsymbol{F}_N^*) := \frac{1}{2} \begin{pmatrix} \{Z\}^{-1} \hat{\boldsymbol{n}} \times (Z^+ [\![\boldsymbol{H}]\!] - \alpha \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}]\!]) \\ \{Y\}^{-1} \hat{\boldsymbol{n}} \times (-Y^+ [\![\boldsymbol{E}]\!] - \alpha \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}]\!]) \end{pmatrix}.$$

or derive yourself: **Demo:** Finding Numerical Fluxes for DG (Part 2)

Good news: Scheme mathematically complete.

Outline

Implementing DG

Weak form:

$$0 = \int_{E_k} q_t^k \phi dx - \int_{E_k} F^k \cdot \nabla \phi dx + \int_{\partial E_k} (F^k \cdot n)^* \phi dx$$

What do the DoFs mean?

Two main choices:

- *Modal* DG (expansion coefficients)
- *Nodal* DG (point values at nodal locations)

We choose to use nodal DG here.

Need: set of basis functions, set of nodes

Modes

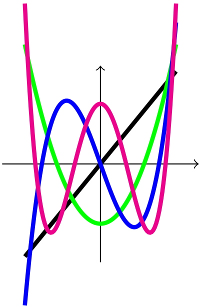
Function spaces same as for FEM: P^N, Q^N.

Numerically: better to use orthogonal polynomials with

$$\int_{\hat{E}} \phi_i \phi_j = \delta_{i,j}$$

- 1D: Legendre polys
- *n*D: Proriol ‘57/Koornwinder ‘75/Dubiner ‘93

Notation: $(\phi_i)_{i=1}^{N_p}$.



Nodes

Define set of interpolation nodes $(\xi_i)_{i=1}^{N_p}$ and ℓ_i their Lagrange basis.

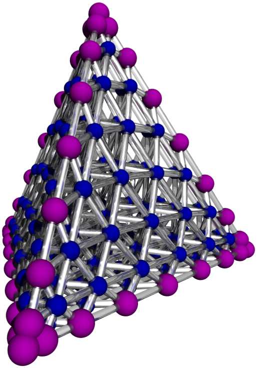
Define *generalized Vandermonde matrix*

$$V_{ij} := \phi_j(\xi_i)$$

$$V(\text{modal coeff.}) = (\text{nodal coeff.})$$

ξ_i determine cond(V)!

- Equispaced nodes: cond. exponential in N
- 1D: Gauß-Lobatto or Chebyshev
- nD: cottage industry (e.g. [Warburton '06])



In Matrix Form

$$0 = \int_{E_k} q_t^k \phi dx - \int_{E_k} F^k \cdot \nabla \phi dx + \int_{\partial E_k} (F^k \cdot n)^* \phi dx$$

Write in matrix form:

$$\mathcal{M}_{ij}^k := \int_{E_k} \ell_i \ell_j dx = |A_k| \mathcal{M} := |A_k| \int_{\hat{E}} \ell_i \ell_j dx = |A_k| V^{-T} V^{-1}$$

$$\mathcal{S}_{ij}^{k,\partial\nu} := \int_{E_k} \ell_i \partial_{x_\nu} \ell_j \mathrm{d}x,$$

$$\mathcal{M}_{ij}^{k,A} := \int_{A \subset \partial E_k} \ell_i \ell_j \mathrm{d}S_x.$$

$$0 = \mathcal{M}^k \partial_t u^k - \sum_\nu \mathcal{S}^{k,\partial_\nu} [F(u^k)] + \sum_{A \subset \partial E_k} \mathcal{M}^{k,A} (\hat{n} \cdot F)^*$$

Explicit Time Integration

$$0 = \mathcal{M}^k \partial_t u^k - \sum_{\nu} \mathcal{S}^{k,\partial_\nu}[F(u^k)] + \sum_{A \subset \partial E_k} \mathcal{M}^{k,A}(\hat{n} \cdot F)^*$$

How can we do time integration on this weak form?

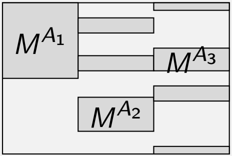
Goal: Dig out $\partial_t u$! Must invert $\mathcal{M}$.

- In 'normal' finite elements: large, unstructured, sparse matrix
- In DG: Block-diagonal
- In simplicial DG: Templated block-diagonal
- In curvilinear DG: *Still* templated block-diagonal
  e.g.: [Warburton '08], [Chan, Hewett, Warburton '17]

Trick: Multiple face mass matrices

Applying multiple face mass matrices at once:

$$\int_{\partial E_k} \hat{n} \cdot (F^*)\phi \mathrm{d}S =$$



$$\left(J_1 \hat{n} \cdot (F^*)|_{A_1} \middle| \cdots \middle| J_3 \hat{n} \cdot (F^*)|_{A_3} \right).$$

DG and Modern Computers: Possible Advantages

DG on modern processor architectures: Why?

- On-chip parallelism
  - DG inherently parallel.
- Deepening Memory Hierarchy
  - The majority of DG is local.
- Compute Bandwidth $\gg$ Memory Bandwidth
  - DG is arithmetically intense.
- Processors favor dense data.
  - Local parts of the DG operator are dense.
- Penalty on scattered access.
  - DG's cell connectivity is sparser than CG's
  - and more regular.