1/11/23 Background & review, PDEs

Heat Eq in 1D: $u_t = u_{xx}$ — parabolic PDE

Wave Eq in 1D: $u_{tt} = u_{xx}$ — hyperbolic PDE

Steady state temp in 2D: $u_{xx} + u_{yy} = 0$ — elliptic PDE

A PDE is ill-posed if for a given input, there is no solution.

Ex: $u_t = u_x$, $x \in (0, 2\pi)$, $t > 0$

$u(x, 0) = \sin x$ on $x \in [0, 2\pi]$

$u(0, t) = 0$ on $t > 0$

$u(2\pi, t) = 0$ on $t > 0$

Note: A numerical scheme for this problem will "blow-up".

Why ill posed: No solution satisfying B.C.s

Ex: $u'' = \sin x$, $u'(0) = 0$, $u'(2\pi) = 0$.

Why ill-posed: Non-uniqueness

$u'' = \sin x \Rightarrow u' = -\cos(x) + B \overset{BC}{\Rightarrow} u' = -\cos(x) + 1$

$\Rightarrow u(x) = -\sin(x) + x + C$ (non-uniqueness)

Ex: $u_t = -u_{xx}$, $x \in (0, 2\pi)$, $t > 0$

$u(x, 0) = f(x)$, $x \in [0, 2\pi]$

$u(0, t) = 0$, $t > 0$

$u(2\pi, t) = 0$, $t > 0$

This problem is ill-behaved. Why? Tiny changes to our function f (in some norm) can make a large change to the solution u (also in norm).

What we want for num'l methods

1) Stability 2) Accuracy 3) Efficiency 4) Simple

Methods we will study

1) Finite difference methods: Easy, simple, transparent, but inflexible order of accuracy & difficult for complex geometries

2) Fourier / Spectral methods: Conceptually simple, "infinite-order" accuracy, very difficult for complex geometries, and can suffer instability

3) Finite volume methods: Solid mathematical theory, can approach non-smooth solutions, but low order accuracy (piecewise constant)

4) Finite element methods: Piecewise polynomial, solid math theory, high-order & geometric flexibility, but technical math & complicated to implement

1/18/23: Finite difference methods for 1D stationary problems

Basic idea: Approximate DEs with finite approximations.

$$u'(x) \approx \frac{u(x+h)-u(x)}{h} \quad \text{or} \quad \frac{u(x)-u(x-h)}{h} \quad \text{or} \quad \frac{u(x+h)-u(x-h)}{2h}$$

Consider $M+2$ equispaced points on $[0,1]$, spacing $h = \frac{1}{M+1}$. We'll call $x_j = jh$, and we'll denote u_j the computational approx to $u(x_j)$.

Define the distance operators D_+, D_-, D_0:

$$D_+ u_j = \frac{u_{j+1}-u_j}{h}, \quad D_- u_j = \frac{u_j - u_{j-1}}{h}, \quad D_0 u_j = \frac{u_{j+1}-u_{j-1}}{2h}.$$

$D_+ u(x) \approx D_- u(x) = u'(x) + O(h)$, $D_0 u(x) = u'(x) + O(h^2)$.

And can chain: $D_+ D_- u(x) = D_- D_+ u(x) \approx u''(x) + O(h^2)$.

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

Note: In big O, note we are hiding higher derivatives of u, so we need derivatives to be well behaved.

Ex: $-u''(x) = f(x)$ on $(0,1)$ with $u(0) = g_0$, $u(1) = g_1$.

This is a steady state heat equation.

We can construct a scheme by $D_+ D_- u_j \approx u''(x_j)$.

This yields $-u_{j-1} + 2u_j - u_{j+1} = h^2 f_j$, $u_0 = g_0$, $u_{M+1} = g_1$, $0 \le j \le M$.

Yields $\underline{\underline{A}}\,\underline{u} = \underline{f} + \frac{g_0}{h^2}\underline{e}_1 + \frac{g_1}{h^2}\underline{e}_M$

where

$$\underline{\underline{A}} = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 \end{bmatrix}, \quad \underline{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_M \end{bmatrix}, \quad \underline{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_M \end{bmatrix}$$

Observations: $u \mapsto u''$ is a symmetric operator, and A symmetric

- $\underline{\underline{A}}$ is invertible
- $\underline{\underline{A}}$ is sparse, $3M-1$ entries
- Brute force $\sim O(M^3)$, can find $O(M)$ algorithms. (Cholesky)
- Use iterative methods, approximate $\underline{\underline{A}}^{-1}\underline{f}$, not $\underline{\underline{A}}^{-1}$.

Is it true $u_j \approx u(x_j)$, and $u_j \to u(x_j)$ as $h \to 0$?

Local Truncation Error: What is the residual plugging our true solution into the numerical scheme?

Def $\quad \tau_j := -D_+ D_- u(x_j) - f(x_j)$.

We can show $\tau_j = ch^2 u^{(4)}(x_j) + O(h^4)$ where c a constant through a Taylor's analysis of $D_+ D_- u(x_j)$.

Then define $\underline{\tau} = [\tau_1 \cdots \tau_M]^T$, and consider $\|\underline{\tau}\|_2$.

We say a numerical scheme is **consistent** if $\lim_{h \to 0} \|\underline{\tau}\|_2 = 0$.

In our case, $\|\underline{\tau}\|_2 = O(h^2) \xrightarrow{h \to 0} 0$. And since the LTE is $O(h^2)$, we say the scheme is consistent to second order. But, consistency doesn't imply accuracy, just gives a suggestion.

Now we look at stability. Abstractly, our problem looks like $\underline{f}, g_0, g_1 \xrightarrow{A^{-1}} \underline{u}$. Hence we need $\underline{\underline{A}}^{-1}$ to behave nicely.

We say our scheme is **stable** if $\|\underline{\underline{A}}^{-1}\|_2 \leq C$ for h sufficiently small, C independent of h.

Recall $$\|\underline{\underline{A}}^{-1}\|_2 = \sup_{\underline{x} \neq 0} \frac{\|\underline{\underline{A}}^{-1}\underline{x}\|_2}{\|\underline{x}\|_2}.$$

We can show our $\underline{\underline{A}}$ is symmetric positive definite, eigenvalues/vecs given by $\lambda_j = \frac{4}{h^2}\sin^2(\pi h j/2)$, $\underline{v_j} = \sqrt{2}\sin(\underline{x}\pi j)$.

Linear Algebra: Recall if $\underline{\underline{A}} \in \mathbb{R}^{M \times M}$, $\exists$ SVD $\underline{\underline{A}} = \underline{\underline{U}}\,\underline{\underline{\Sigma}}\,\underline{\underline{V}}^T$ and $\|\underline{\underline{A}}\|_2 = \sigma_1$. If A symmetric, $\exists$ $\underline{\underline{A}} = \underline{\underline{U}}\,\underline{\underline{D}}\,\underline{\underline{U}}^T$. If A

symmetric & invertible, $\underline{\underline{A}}^{-1} = \underline{\underline{U}}\,\underline{\underline{D}}^{-1}\underline{\underline{U}}^T$. If A symmetric, $|\lambda_j| = \sigma_j$. We can show in our example, $\|\underline{\underline{A}}^{-1}\|_2 \sim \frac{1}{\pi^2}$ since $\|\underline{\underline{A}}^{-1}\|_2$ equals $1/(\text{minimum eigenvalue of } \underline{\underline{A}})$. Hence our scheme is stable.

1/23/23:

Original question, is scheme accurate? Let $e_j := u_j - u(x_j)$, $\underline{e} = (e_1 \cdots e_M)^T$. We say our scheme is convergent if $\lim_{h \downarrow 0} \|\underline{e}\|_2 = 0$. Note that this is a strong statement. $h \downarrow 0 \Rightarrow M \to \infty$.

Define $\underline{U}$ to be the vector of exact solutions, so we have $\underline{\underline{A}}\,\underline{U} = \underline{f} + \frac{g_0}{h^2}\underline{e}_1 + \frac{g_1}{h^2}\underline{e}_M + \underline{\tau}$. Therefore,

$$-\underline{\tau} = \underline{\underline{A}}(\underline{u} - \underline{U}) = \underline{\underline{A}}\,\underline{e}, \text{ giving us that}$$

$$\|\underline{e}\|_2 = \|\underline{\underline{A}}^{-1}\underline{\tau}\|_2 \le \|\underline{\underline{A}}^{-1}\|_2 \|\underline{\tau}\|_2 \le C\,O(h^2).$$

So in this case, stability ($\|\underline{\underline{A}}^{-1}\|_2 \le C$) and consistency ($\|\underline{\tau}\|_2 = O(h^2)$) imply accuracy. Specifically, it is second order accurate.

Overview:

- Consistency: Local truncation error small relative to h
- Stability: Scheme is well behaved for small h
- Linearity: Global Error = Consistency order (ish)

Lax Equivalence Theorem: Stability & Consistency give Convergence, but depends on your definition of stability and consistency.

Elliptic PDE: $-\nabla \cdot (\underline{\underline{K}}(x,y)\nabla u) = f(x,y)$, $(x,y) \in (0,1)^2$

$$u(0,y) = g_0(y), \quad u(1,y) = g_1(y)$$
$$u(x,0) = h_0(x), \quad u(x,1) = h_1(x).$$

For well-posedness, require $\underline{\underline{K}}$ is symmetric positive definite. Models spatially dependent heat diffusion.

We'll focus on Poisson's Equation: $-\Delta u = f$. If $f \equiv 0$, we

get Laplace's Equation. Again, let $h = \frac{1}{M+1}$, and we have $u_{ij} \approx u(x_i, y_j)$, $x_i = ih$, $y_j = jh$. We use the same approximations as before, $O(h^2)$.

$$u_{xx}(x_i, y_j) \approx D_+^x D_-^x u_{ij}$$
$$u_{yy}(x_i, y_j) \approx D_+^y D_-^y u_{ij}$$

Scheme:

$$\begin{array}{ccc} & -u_{i,j+1} & \\ -u_{i-1,j} & 4u_{ij} & -u_{i+1,j} \\ & -u_{i,j-1} & \end{array} = h^2 f_{ij}, \quad i,j = 1, \dots, M,$$

w/ $u_{0,j} = g_0(y_j)$, $u_{M+1,j} = g_1(y_j)$,
$u_{i,0} = h_0(x_i)$, $u_{i,M+1} = h_1(x_i)$.

1/25/23:

Recall $D_+^x D_-^x u(x,y) \approx u_{xxxx}(x,y) h^2 = O(h^2)$, so we wish for our scheme to have order 2 accuracy.

Again, we can order $u(x_i, y_j)$ as a vector to get $\underline{\underline{A}}\,\underline{u} = \underline{\hat{f}}$ where $\underline{\hat{f}}$ depends on $\underline{f}$ and B.C.s.

Considerations in 2D:

- $\underline{\underline{A}}$ not tridiagonal, but still sparse
- ordering of u_{ij} matters
- $\underline{\underline{A}}$ is $M^2 \times M^2$, $\underline{u}$ has M^2 entries
- No simple trick for $O(M^2)$, but can find iterative $O(M^2 \log M)$ methods.

Similar to 1D:

- 2nd order accurate/convergent in h
- Still quadratic accuracy, but halving h quadruples nodes $\Rightarrow$ linear payoff in 2D (not quadratic)

In general, degree d, $2d+1$ stencil, but $M^d \sim \left(\frac{1}{h}\right)^d$ degrees of freedom. But iterative methods can achieve $O(dM^2 \log(M))$ time. Still a stable scheme w/ second-order accuracy in h. With $d \geq 3$, sublinear cost vs accuracy payoff.

Trick: Error looks like $u_{xxxx} + u_{yyyy} = \Delta(\Delta u) - 2u_{xxyy} = \Delta f - 2u_{xxyy}$

And, consider second stencil

$$\widetilde{\Delta}_5 u_{ij} = \begin{array}{ccc} -u_{i-1,j+1} & & -u_{i+1,j+1} \\ & 4u_{ij} & \\ -u_{i-1,j-1} & & -u_{i+1,j-1} \end{array} \approx 2h^2 \Delta u(x_i, y_j).$$

And, LTE of $\widetilde{\Delta}_5$ has u_{xxyy} in it. Combining the stencils,

$$\Delta u(x_i, y_j) \approx \Delta_9 u_{ij} := \frac{1}{6h^2}\begin{pmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{pmatrix}$$

which has LTE $\frac{h^2}{12}\Delta^2 u + O(h^4) = \frac{h^2}{12}\Delta f + O(h^4)$. For Laplace's eq, $f = 0 \Rightarrow \Delta f = 0$ yielding 4^{th} order accuracy in h. If $\Delta f \neq 0$, say Poisson, we just say order is $\frac{h^2}{12}\Delta f$. And if f from data instead, $\Delta_9 u_{ij} = f_{ij} + \frac{h^2}{12}\Delta_5 f_{ij}$.

Method of deferred corrections makes an approximate solution $\tilde{u}$, then adds to RHS the estimated LTE of stencil, solves again.

Solvers for Initial Value Problems

Start with time dependent ODE $\underline{u}'(t) = \underline{f}(\underline{u}, t)$, $\underline{u}(0) = \underline{u}_0$. Wish to find solution on $t \in [0, T]$.

Note, method of lines involves discretizing PDEs along all vars but 1, say time. Ex: $u_t = u_{xx} + uu_x \rightarrow \underline{u}'(t) = \underline{D}^2\underline{u} + \underline{u} \circ (\underline{D}\underline{u}) = \underline{f}(\underline{u})$ where $\underline{D}$ is a discretization (a matrix) of $\partial/\partial x$, and $\circ$ is elementwise product. Yields an IVP.

Recall Picard-Lindelöf Theorem provides existence and uniqueness for ODEs in some ball around $\underline{u}_0$, to given $\underline{f}$ Lipschitz continuous.

Uses Picard iterates: $\underline{u}(t) = \underline{u}(0) + \int_0^t \underline{f}(\underline{u}(s), s)\,ds$.

Let's discretize time: $t_0 = 0$, $t_n = nk$, $T = Nk$, $\underline{u}_n \approx \underline{u}(t_n)$.

Note that $\underline{u}_{n+1} \approx \underline{u}_n + \int_{t_n}^{t_{n+1}} \underline{f}(t, \underline{u}(t))\,dt$.

Forward Euler: $\int_{t_n}^{t_{n+1}} \underline{f}(t, \underline{u}(t))\,dt \approx (t_{n+1} - t_n)\,\underline{f}(t_n, u_n) =: k\underline{f}_n$

yielding the explicit scheme $\underline{u}_{n+1} = \underline{u}_n + k\underline{f}_n$.

Backward Euler does Riemann sum other dir. $\underline{u}_{n+1} = \underline{u}_n + k\underline{f}_{n+1}$, is implicit.

Crank-Nicolson uses trapezoidal quadrature: $\underline{u}_{n+1} = \underline{u}_n + \frac{k}{2}(\underline{f}_n + \underline{f}_{n+1})$.

An ODE scheme is convergent to order p if given $\underline{e_n} = \underline{u}(t_n) - \underline{u_n}$, $\max_{n\in[N]} \|\underline{e_n}\| = O(k^p)$ for some choice of norm.

Again, consistency means $LTE \xrightarrow{k\to 0} 0$. Denote the forward time difference operator $D^+ \underline{u_n} = \frac{\underline{u_{n+1}} - \underline{u_n}}{k}$. FE: $\frac{\underline{u_{n+1}} - \underline{u_n}}{k} = \underline{f_n}$

For forward Euler, $LTE_n := \| D^+\underline{u}(t_n) - \underline{f}(t_n, \underline{u}(t_n))\| \approx k\|\underline{u}''(t_n)\| = Ck$

Note we worked in units of $\frac{\partial u}{\partial t}$, not u, or else we would have order 2. Stability is trickier as error builds over each iteration.

1/30/23:

We say a numerical scheme is 0-stable if there exists a constant C s.t. for all k sufficiently small,

$$\max_{n\in[N]} \|\underline{e_n}\| \le C\left(\|\underline{e_0}\| + \max_{n\in[N]} \|R_n \underline{u}(t_n)\|\right)$$

($\|\underline{e_0}\|$: typically 0; $\max_{n\in[N]} \|R_n \underline{u}(t_n)\|$: LTE$_n$)

where $R_n \underline{u}(t_n) := D^+\underline{u}(t_n) - \underline{f}(t_n, \underline{u}(t_n))$, so $\|R_n\underline{u}(t_n)\| = LTE_n$.

Note: Technically, also includes $\max_{n\in[N]} R_n \underline{u_n}$, which in our case is zero. For implicit methods, may not be zero.

So, consistency + 0-stability $\Rightarrow$ Convergence.

This can be shown explicitly for Forward Euler using Lipschitz Continuity and iterated errors. Constant C depends on Lipschitz constant & T.

Hence, Forward Euler is 1st order convergent, but constant $\frac{e^{LT}}{L}$.

Note for $f(t,u) = -10^6 u$, $L = 10^6$, e^{10^6} very large.

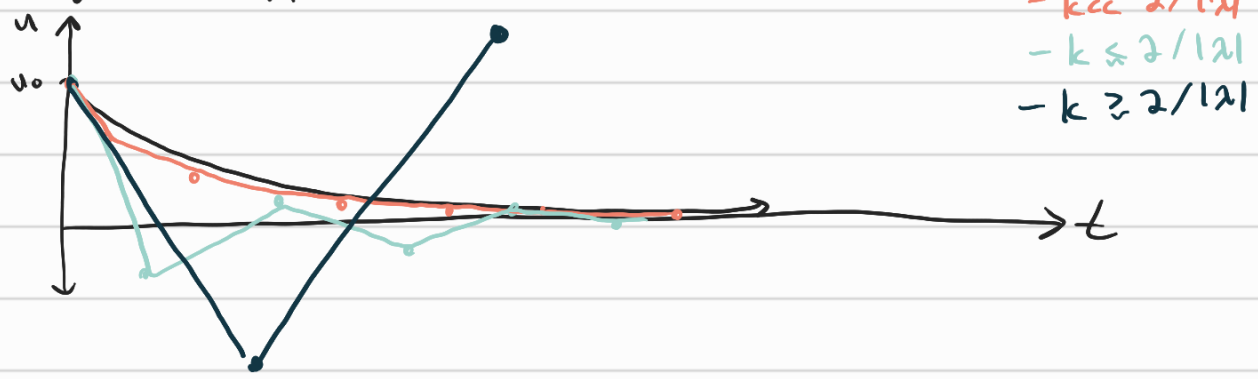
We want to derive a new notion of stability, consider the simpler problem $u'(t) = \lambda u(t)$, $u(0) = u_0$, $\lambda \in \mathbb{C}$. Exact sol $u(t) = u_0 e^{\lambda t} = u_0 e^{\mu t}(\cos\omega t + i\sin\omega t)$, $\lambda = \mu + i\omega$.

If $\mu \le 0$, we expect $|u_{n+1}| \le |u_n|$, as $\mu \le 0$ implies oscillatory behavior or $u \to 0$ for stability.

2/1/23:

For Forward Euler, we have $u_{n+1} = u_n + k\lambda u_n$. Define $z = \lambda k$, so instead of modifying λ and k, just modify z.

Note that $|u_{n+1}| \leq |u_n| \Leftrightarrow |1+z| \leq 1$. We call $1+z$ the amplification factor. Performing algebra, $k \leq -\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}$. And if λ real, require $k \leq \frac{2}{|\lambda|}$.



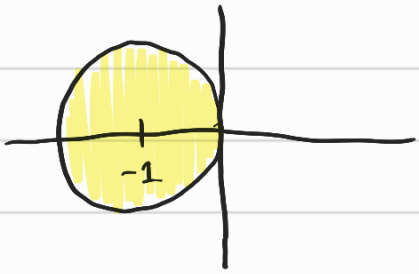
So, now consider the problem w/ forward Euler.

$$u' = \lambda u, \quad u(0) = 1, \quad 0 \leq t \leq 1, \quad \mathrm{Re}(\lambda) \ll -1$$

We showed this is consistent, but we need $k \ll 1$ for stability ($k < 2/|\lambda|$). Such problems where the stability criterion is much stricter than the accuracy (consistency) requirement are called <u>stiff problems</u>.

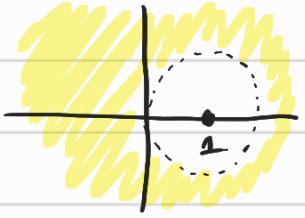
The notion of <u>absolute stability</u> is the requirement $|u_{n+1}| \leq |u_n|$ applied to the ODE problem $u' = \lambda u$ for $\mathrm{Re}(\lambda) \leq 0$, and a time step k. The set of values $z = \lambda k \in \mathbb{C}$ satisfying $|u_{n+1}| \leq |u_n|$ is called the <u>Region of Stability (ROS)</u>. A numerical method is called <u>A-stable</u> if it satisfies $|u_{n+1}| \leq |u_n|$ for some $k > 0$.

Forward Euler A-stablity $|z+1| \leq 1$. Not A-stable for all $k \geq 0$!



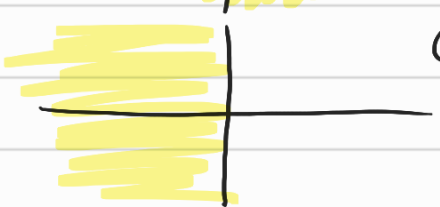
For Backward Euler, $u_{n+1} = u_n + k\lambda u_{n+1}$, so $\frac{|u_{n+1}|}{|u_n|} = \frac{1}{|1-z|}$, giving us the region of stability pictured. Since $\mathrm{Re}(\lambda) \leq 0$, $k > 0$, $\mathrm{Re}(z) < 0$, so in fact Backward Euler is A-stable

B.E.



Similarly, Trapezoidal/ Crank Nicholson methods are A-stable with ROS of $\mathrm{Re}(z) \leq 0$.

C.N.



Forward Euler note, choosing $\lambda = 0 \pm i$, we see that $z = k\lambda = \pm ik$ which only lies in the ROS of $k = 0$. So in fact, no k value (positive) will result in non-increasing solutions. For Backward Euler, for $\pm ki$ to be in ROS, we just require $k \neq 0$.

Takeaway: Explicit methods easy to implement, but suffer instability. Implicit methods are more difficult to implement, but much more stable.

Now, consider a linear ODE: $\underline{u}'(t) = \underline{\underline{A}}\,\underline{u}$, $\underline{u}(0) = \underline{u}_0 \in \mathbb{R}^M$ where $\underline{\underline{A}}$ is diagonalizable, $\underline{\underline{A}} = \underline{\underline{V}}\,\underline{\underline{\Lambda}}\,\underline{\underline{V}}^{-1}$, so performing the change of variables $\underline{w} = \underline{\underline{V}}^{-1}\underline{u}$, we get $\underline{w}'(t) = \underline{\underline{\Lambda}}\,\underline{w}$. So a new notion of stability is $|(w_{n+1})_m| \leq |(w_n)_m| \Rightarrow k\lambda_m \in \text{ROS}$. Or, we require

$$k\lambda(\underline{\underline{A}}) \subseteq \text{ROS}$$

Or for nonlinear systems, $\underline{u}'(t) = \underline{f}(t, \underline{u})$, $\underline{u}(0) = \underline{u}_0$, we can linearize our problem around some t_n to get $\underline{u}'(t) = \frac{\partial \underline{f}}{\partial \underline{u}}(t_n, \underline{u}(t_n))\,\underline{u}$, and we can extend our notion of stability to

$$k\lambda\left(\frac{\partial \underline{f}}{\partial \underline{u}}(t_n, \underline{u}(t_n))\right) \subseteq \text{ROS}$$

but in practice we can only compute $\underline{u}_n$ not $\underline{u}(t_n)$.

Note: "Nearly all" matrices are diagonalizable.

2/6/23:

Moving forward, we will assume a general quadrature

$$\int_{t_n}^{t_{n+1}} \underline{f}(t, \underline{u}(t))\,dt \approx \sum_{j=1}^{s} k b_j \underline{f}(t_{n,j}, \underline{u}(t_{n,j})), \quad t_{n,j} = t_n + kc_j$$

for some constants b_j, c_j, and $s \geq 1$. What do we choose to approximate $\underline{u}(t_{n,j})$?

First, we will choose $c_2 = \frac{1}{2}$, $s = 1$, so we determine $b_2 = 1$ for consistency. How to compute $\underline{u}(t_{n,2})$? Straight-forward method is to use forward Euler. So $\underline{u}(t_n + \frac{k}{2}) \approx \underline{U}_2 := \underline{u}_n + \frac{k}{2}\underline{f}(t_n, \underline{u}_n)$

$$\Rightarrow \underline{u}_{n+1} = \underline{u}_n + k\underline{f}(t_n + \tfrac{k}{2}, \underline{U}_2).$$

Def $D^+ \underline{u}(t_n) := \underline{f}(t_n + \frac{k}{2}, u(t_n + \frac{k}{2})) + O(k^2)$, yielding an order 2 scheme. BUT, $u(t_n + \frac{k}{2}) \neq \underline{U}_1$

In fact, through Taylor analysis, we find that this procedure is second order accurate, we call this the Explicit Midpoint Method

$$\underline{u}(t_n + \tfrac{k}{2}) \approx \underline{U}_1 := \underline{u}_n + \tfrac{k}{2} \underline{f}(t_n, \underline{u}_n),$$

$$\underline{u}_{n+1} = \underline{u}_n + k \underline{f}(t_n + \tfrac{k}{2}, \underline{U}_1).$$

This can be generalized to multi-stage methods.

$$\underline{u}(t_{n,j}) \approx \underline{U}_j := \underline{u}_n + k \sum_{l=1}^{s} a_{j,l} \underline{f}(t_{n,l}, \underline{U}_l), \quad t_{n,j} = t_n + k c_j$$

$$\underline{u}_{n+1} = \underline{u}_n + k \sum_{j=1}^{s} b_j \underline{f}(t_{n,j}, \underline{U}_j)$$

where we must determine $a_{j,l}$, b_j, c_j coefficients. Also known as Runga-Kutta methods.

- If $a_{j,l} \neq 0$ for any $l \geq j$, the scheme is implicit
- Quite cumbersome for $s \geq 3$, feasible for $s \leq 2$.

Theorem: There is no Runga Kutta method with $s = p$ if $p \geq 5$.

Stages	1	2	3	4	5	6	7	8	9	10
Achievable R-K Order	1	2	3	4	4	5	6	6	7	7

diminishing returns $\longrightarrow$

This is why RK-4 is so popular.

We display coefficients w/ a Butcher Tableau

c_1	a_{11}	a_{12}	$\cdots$	a_{1s}
c_2	a_{21}	a_{22}	$\cdots$	a_{2s}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
c_s	a_{s1}	a_{s2}		a_{ss}
	b_1	b_2	$\cdots$	b_s

What we've seen so far

0	0
	1

Forward Euler

1	1
	1

Backward Euler

0	0	0
0	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Crank-Nicholson

RK-2

0	0	0
c	c	0
	$1-\frac{1}{2c}$	$\frac{1}{2c}$

for $c \in (0, 1]$

$c = 1$ - explicit trapezoid

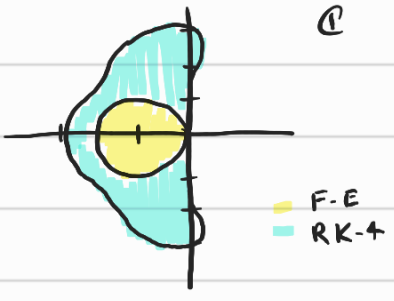
$c = \frac{1}{2}$ - explicit midpoint

Classical RK-4

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

We can show RK methods are 0-stable, hence convergence of the LTE implies convergence. And can also plot ROSs for A-stability. RK4 extends the ROS of Forward Euler. The update for RK is simply a polynomial in $z = \lambda k$, cannot be A-Stable for explicit RK.

Fact: No explicit method can be A-stable



2/8/23 Butcher Tableaus:

c_1	a_{11}	a_{12}	$\cdots$	a_{1s}
c_2	a_{21}	a_{22}	$\cdots$	a_{2s}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
c_s	a_{s1}	a_{s2}		a_{ss}
	b_1	b_2	$\cdots$	b_s

$$\begin{cases} t_{n,j} = t_n + k c_j \\ \underline{u}(t_{n,j}) \approx \underline{U}_j := \underline{u}_n + k \sum_{i=1}^{s} a_{j,i} \underline{f}(t_{n,i}, \underline{U}_i) \\ \underline{u}_{n+1} = \underline{u}_n + k \sum_{j=1}^{s} b_j \underline{f}(t_{n,j}, \underline{U}_j) \end{cases}$$

All RK methods are 0-stable by defn, and "A-stable" inside ROS

In practice, we use methods in combination, one lower order and one higher order, to approximate error.

> $\underline{u}_n$ - Lower order (ex. RK3) > $\underline{\tilde{u}}_n$ - Higher order (ex RK4)

And then $\|\underline{e}_n\| = \|\underline{u}(t_n) - \underline{u}_n\| \approx \|\underline{\tilde{u}}_n - \underline{u}_n\|$. But, this requires over twice the amount of work for getting $\underline{u}_n$.

Solution: Embedded Multi-Stage methods in which we solely

change the b_j coefficients to hopefully increase order of accuracy.
So $\underline{u}_{n+1} = \underline{u}_n + k\sum_{j=1}^{s} b_j f(t_{n,j}, \underline{U}_j)$, and also we have
$\tilde{\underline{u}}_{n+1} = \underline{u}_n + k\sum_{j=1}^{s} \tilde{b}_j f(t_{n,j}, \underline{U}_j)$, no need to recalculate U's. One example is the Dormand-Prince 4(5) method, which has 7 stages.
b_j's yield order 4 accuracy, $\tilde{b}_j$'s yield order 5 accuracy. Very ugly tableau.
Used in Julia as DP5(), MATLAB as ode45(), python: scipy.integrate.ode().
We then adaptively modify k, given some ε_{tol}, $\hat{k}$ chosen s.t.

$$\left(\frac{\hat{k}}{k}\right)^p \|\underline{u}_n - \tilde{\underline{u}}_n\| \lesssim \varepsilon_{tol}$$

given that the embedded method (worse one) is of order p.

Theorem: Polynomial Interpolation: There exists a unique polynomial $p(x)$ of degree at most n that interpolates $n+1$ points.
Constructed thru Newton divided differences

$$f[x_j] = f(x_j), \quad f[x_j, \cdots, x_{j+s}] = \frac{f[x_{j+1}, \cdots, x_{j+s}] - f[x_j, \cdots, x_{j+s-1}]}{x_{j+s} - x_j}$$

and $p(x) = \sum_{s=0}^{n} f[x_0, \cdots, x_s] \prod_{j=0}^{s-1} (x - x_j)$.

Difference Equations: $u_n + \sum_{j=1}^{s} \alpha_j u_{n-j} = 0$, $u_{n-j} = u_{n-j,0}$, $j = 1, 2, \cdots, s$.
Similar to ODEs, assume $u_n = z^n \Rightarrow p(z) := \sum_{j=0}^{s} \alpha_j z^{s-j} = 0$, $(\alpha_0 = 1)$.
So solns of form $u_n \sim z_j^n$ where $z_1, \cdots, z_s$ are the roots of $p(z)$.
This is like taking $u(t) = e^{zt}$ for ODEs. So we can show solns u_n are stable if $|z_j| \le 1$, asymptotically stable if $|z_j| < 1$.

For the IVP $\underline{u}'(t) = \underline{f}(t, \underline{u})$, $\underline{u}(0) = \underline{u}_0$, we have $\underline{u}_n \approx \underline{u}(t_n)$.
a general s-step mult:-step scheme w/ step k has the form

$$\sum_{j=0}^{s} \alpha_j \underline{u}_{n+s-j} = k \sum_{j=0}^{s} \beta_j \underline{f}(t_{n+s-j}, \underline{u}_{n+s-j}), \quad \alpha_j, \beta_j \in \mathbb{R}.$$

2/13/23: Comments on multi-step

- $s = 1$ is general single step (and stage)
- $s > 1$ uses time history
- $\alpha_0 \neq 0$ to solve for $\underline{u}_{n+1}$
- Rescale eq s.t. $\alpha_0 = 1$
- $\beta_0 = 0$ is explicit, $\beta_0 \neq 0$ is explicit
- Assume either $\alpha_j \neq 0$ or $\beta_j \neq 0$ $\forall j$

For simplicity, assume ODE is autonomous, so $\sum_{j=0}^{s} \alpha_j \underline{u}_{n+1-j} = k \sum_{j=0}^{s} \beta_j \underline{f}_{n+1-j}$.

In general, α_j approximates $\frac{d}{dt}\underline{u}(t)$, β_j approximates $\frac{1}{k}\int_{t_n}^{t_{n+1}} \underline{f}(\underline{u}(r))\,dr$.

Note: For $s \geq 2$, must calculate $\underline{u}_1, \ldots, \underline{u}_{s-1}$ from $\underline{u}_0$ to begin. Typically derived from multi-stage method.

$s=1$: $\underline{u}_{n+1} + \alpha_1 \underline{u}_n = k(\beta_0 \underline{f}_{n+1} + \beta_1 \underline{f}_n)$. For any notion of consistency, should have $\alpha_1 = -1$. And RHS should approximate $\int_{t_n}^{t_{n+1}} \underline{f}(\underline{u}(r))\,dr$, so require $\beta_0 + \beta_1 = 1$ for consistency. So, $\underline{u}_{n+1} - \underline{u}_n = k(\beta \underline{f}_{n+1} + (1-\beta)\underline{f}_n)$.

* $\beta = 0$ - Forward Euler, $\beta = 1$ - Backward Euler, $\beta = \frac{1}{2}$ - Crank Nicholsen *

Adams Methods wish to approximate $\underline{u}(t_{n+1}) = \underline{u}(t_n) + \int_{t_n}^{t_{n+1}} \underline{f}(\underline{u}(r))\,dr$, so sets $\alpha_0 = 1$, $\alpha_1 = -1$, then determine other coefficients. Choose β_j s.t. $\int_{t_n}^{t_{n+1}} \underline{f}(\underline{u}(r))\,dr \approx k \sum_{j=0}^{s} \underline{f}_{n+1-j}$, note this uses points outside the region of integration $(t_{n+1}, t_n, \ldots, t_{n+1-s})$.

Explicit Adams Methods, $\beta_0 = 0$, are Adams Bashforth methods. Can find β_j's thru Taylor expansion or polynomial interpolation. There exist tables of coefficients for these methods. $(p = s)$

Implicit Adams Methods, $\beta_0 \neq 0$, are Adams Moulton methods where β_j chosen for highest order LTE. One extra degree of freedom, so we achieve one higher degree of LTE. $(p = s+1)$

Note: Neither of above are good at stiff problems.

Backwards Differentiation Methods set $\beta_j = 0$ for $j > 0$, but then need nonzero α_j's. $\sum_{j=0}^{s} \alpha_j \underline{u}_{n+1-j} = k\beta_0 \underline{f}_{n+1}$. The coefficients are computable, all implicit methods. The idea is $\underline{u}'(t_n) \approx \sum_{j=0}^{s} \alpha_j \underline{u}_{n+1-j}$. Called BDF methods. $(p = s)$

How to compute LTE for $\sum_{j=0}^{s} \alpha_j \underline{u}_{n+1-j} = k \sum_{j=0}^{s} \beta_j \underline{f}_{n+1-j}$? We compute

$$\frac{1}{k}\sum_{j=0}^{s} \alpha_j \underline{u}(t_{n+1-j}) - \sum_{j=0}^{s} \beta_j \underbrace{\underline{u}'(t_{n+1-j})}_{\underline{f}(\underline{u}(t_{n+1-j}))}.$$

Looking at the order $\frac{1}{k}$ terms, we want $\sum_{j=0}^{s} \alpha_j = 0$.

Looking at the order 1 terms, $\sum_{j=0}^{s} (s-j)\alpha_j - \sum_{j=0}^{s} \beta_j = 0$.

We just need $O(\frac{1}{k})$, $O(1)$ satisfied for consistency.

Define $\rho(w) = \sum_{j=0}^{s} \alpha_j w^{s-j}$

$\sigma(w) = \sum_{j=0}^{s} \beta_j w^{s-j}$.

Theorem: A multi-step method is consistent if and only if $\rho(1) = 0$ and $\rho'(1) = \sigma(1)$.

Reminder: Consistency says $\lim_{k \to 0} \| LTE \| = 0$.

Ex: Derive an explicit multistep method: Let $s = 2$, so

$$u_{n+2} + \alpha_1 u_n + \alpha_2 u_{n-1} = k\beta_1 f_n + k\beta_2 f_{n-1}$$

Taylor: $u_{n+1} \approx u_{n-1} + 2k u_{n-1}' + \frac{4k^2}{2} u_{n-1}'' + \frac{8k^3}{6} u_{n-1}''' + \cdots$

$u_n \approx u_{n-1} + k u_{n-1}' + \frac{k^2}{2} u_{n-1}'' + \frac{k^3}{6} u_{n-1}''' + \cdots$

$f_n = u_n' \approx u_{n-1}' + k u_{n-1}'' + \frac{k^2}{2} u_{n-1}''' + \cdots$

$O(u_{n-1})$: $1 + \alpha_1 + \alpha_2 = 0$ $\quad (\rho(1) = 0)$

$O(u_{n-1}')$: $2k + \alpha_1 k = \beta_1 k + \beta_2 k$ $\quad (\rho'(1) = \sigma(1))$

$O(u_{n-1}'')$: $\frac{4k^2}{2} + \alpha_1 \frac{k^2}{2} = \beta_1 k \Rightarrow \beta_1 = 2 + \frac{\alpha_1}{2}$

$O(u_{n-1}''')$: $\frac{4}{3}k^3 + \alpha_1 \frac{k^3}{6} = \frac{k^3}{2}\beta_1 \Rightarrow \beta_1 = \frac{8}{3} + \frac{\alpha_1}{3}$

$\alpha_1 = 4, \ \alpha_2 = -5$

$\beta_1 = 4, \ \beta_2 = 2$.

But, we see this method still "blows-up"

Theorem: A multi-step method is 0-stable if and only if the roots of $\rho(w)$ all satisfy $|w_i| \leq 1$ and any roots satisfying $|w_i| = 1$ are simple. $(w-1)$ $\quad (w^s - w^{s-1} = w^{s-2}(w-1))$

> Any $s = 1$ & Adams methods are 0-stable, and all BDF methods w/ $s \leq 6$ are 0-stable.

> For last ex: $\rho(w) = w^2 + 4w - 5 = (w+5)(w-1)$.

2/15/23:

For absolute stability, require that for $u' = \lambda u$, $Re(\lambda) \leq 0$, solns do not grow exponentially in time. Plugging in this equation, we get the difference eqn $\sum_{j=0}^{s} \alpha_j u_{n+2-j} = k\lambda \sum_{j=0}^{s} \beta u_{n+2-j}$, has characteristic eqn $\rho(w) = k\lambda\sigma(w) = z\sigma(w)$. So we require that the roots of $\rho(w) - z\sigma(w)$ satisfy $|w_i| \leq 1$.

Ex: Forward Euler: $\rho(w) = w - 1$, $\sigma(w) = 1$, so $\rho(w) - z\sigma(w) = 0$, root $w = z + 1$,

$\Rightarrow |z+1| \leq 1$, same ROS as before.

Note: Multi-step methods easy to implement, but small ROS. The ROS, actually shrink as s grows for Adams- methods.

Note: Starting multi-step methods usually require Runga-Kutta of similar order.

General Linear Methods are the superset of multi step & stage methods.

Finite Difference Methods for PDEs

We wish to study parabolic problems, ex heat eqn. Consider a scalar PDE:

$$u_t = \underbrace{p\left(\frac{\partial}{\partial x}\right)}_{\text{(polynom in } \partial/\partial x)} u, \quad u(0,t) = u(2\pi, t).$$

where p is an operator. For prototypical eqn, define $p\left(\frac{\partial}{\partial x}\right) = \frac{\partial^2}{\partial x^2}$.

Define the Fourier Transform

$$F(\omega) = \mathcal{F}[f] = \int_0^{2\pi} f(x)\,\overline{\phi(x,\omega)}\,dx, \quad \phi(x,\omega) = \frac{1}{\sqrt{2\pi}} e^{i\omega x}$$

where $\omega \in \mathbb{Z}$. Note $\int_0^{2\pi} |f(x)|^2 dx = \sum_{\omega\in\mathbb{Z}} |F(\omega)|^2$, $\mathcal{F}$, $\mathcal{F}^{-1}$ are well-defined.

And important property, $\mathcal{F}\left(\frac{\partial}{\partial x} f\right) = i\omega F(\omega)$.

So,

$$\mathcal{F}\left[p\left(\frac{\partial}{\partial x}\right) u(x,t)\right] = P(\omega)\,U(\omega,t).$$

And

$$u_t = p\left(\frac{\partial}{\partial x}\right) u \overset{\mathcal{F}}{\Rightarrow} \frac{d}{dt} U(\omega,t) = P(\omega) U(\omega,t), \quad U(\omega,0) = U_0(\omega).$$

We've reduced this to an infinite decoupled system of ODEs. We can solve these as $U(\omega,t) = U_0(\omega) e^{P(\omega)t} \Rightarrow u(x,t) = \sum_{\omega\in\mathbb{Z}} \frac{1}{\sqrt{2\pi}} U_0(\omega) e^{P(\omega)t} e^{i\omega t}$.

The PDE above is stable if $\exists\, K, \alpha \in \mathbb{R}$ s.t. $|e^{P(\omega)t}| \le K e^{\alpha t}$, $t \ge 0$, $\omega \in \mathbb{Z}$.

Theorem: If PDE stable, Fourier-based solution is the unique solution and is "smooth".

Ex: Heat equation $u_t = u_{xx}$, $u(x,0) = u_0(x)$.

Exact soln: $\frac{1}{\sqrt{2\pi}} \sum_{\omega\in\mathbb{Z}} U_0(\omega) e^{-\omega^2 t} e^{i\omega t}$, soln unique & smooth.

($P(\omega) = -\omega^2$, $U_0(\omega) = \mathcal{F}(u_0(x))$).

Ex: $u_t = -u_{xx}$ not stable, $P(\omega) = \omega^2$, so no α, K s.t. $|e^{\omega^2 t}| < K e^{\alpha t}$.

Discretize Heat Eq: $u_t = u_{xx}$, $u(x,0) = u_0(x)$, $u(0,t) = u(2\pi,t)$.

- Equidistant mesh, $h = \Delta x$, $k = \Delta t$, $u_j^n \approx u(x_j, t_n)$, $\underline{u}^n = (u_0^n, \dots, u_{M-1}^n)^T$.
- Scheme: $D^+ u_j^n = D_- D_+ u_j^n$, $j \in [M]$, $n \geq 0$.
- Explicitly: $u_j^{n+1} = u_j^n + \frac{k}{h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n)$.
- Can also do this with Crank-Nicholson, RK4 (just not $D^+ u_j^n$ anymore)
- FE, RK4 seem to blow up, Crank-Nicholson converges for $N = M = 100$.
- Instead, we should've discretized in space, then gotten something of form $\underline{u}_t = A\underline{u}$, and can then study convergence/stab.

2/22/23:

Today, discuss stability/accuracy of above finite difference method.
We didn't see expected $O(k)$, $O(h^2)$ accuracy for FE.
We fully discretized space and time before. Now, only discretize space: $\frac{d}{dt}\underline{u}(t) = \underline{\underline{A}}\,\underline{u}(t)$, $h^2\underline{\underline{A}} = \begin{bmatrix} -2 & 1 & & 1 \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ 1 & & 1 & -2 \end{bmatrix}$, $\underline{u}(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_M(t) \end{bmatrix}$

This is called the method of lines (semi-discrete approx.). It's useful in decoupling space & time. We can now study

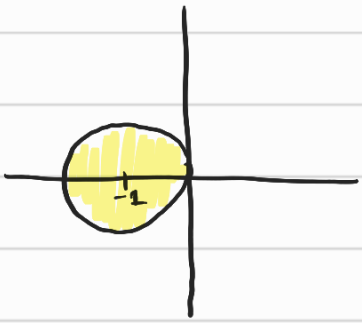
- Stability (0- or A-stability) in time
- Accuracy (time discretization)
- Convergence (on a fixed spatial discretization to ODE soln, not necessarily PDE)

Consider Forward Euler: $\underline{u}^{n+1} = \underline{u}^n + k\underline{\underline{A}}\underline{u}^n$. Let's talk about A-stability, do the solutions grow in time? Eigenvalues/vectors of $\underline{\underline{A}}$ computable:

$$\lambda_j(\underline{\underline{A}}) = \frac{-4}{h^2}\sin^2\left(\frac{\pi \check{j}}{2M}\right), \quad \check{j} = \begin{cases} j-1, & j \text{ odd} \\ j, & j \text{ even} \end{cases}, \quad j \in [M] \quad (1, 2, \dots, M)$$

all of which have negative real parts, as we'd hoped for. So solns should not grow in time. Well, $\lambda_{min}(\underline{\underline{A}}) = \frac{-4}{h^2} \sim -4M^2$, $\lambda_{max}(\underline{\underline{A}}) \approx -1$.
Note, $\lambda_{min} \ll \lambda_{max}$ suggests this problem may be stiff.
Some very quick decay, some slow decay. So, our choice of spatial discretization will impact our time discretization.

Recall for FE, the ROS given by $|z+1| \leq 1$ with $z = \lambda k$. All our λ's on negative real axis, so we'll require $z \geq -2 \rightarrow k|\lambda_{min}(\underline{\underline{A}})| \leq 2$



$$\Rightarrow \quad k \leq \frac{h^2}{2}.$$

So for A-stability, say $M = 100$, $h = 1e{-}2$, $k \leq \frac{1}{2}e{-}4$.

And note, you won't escape this with other explicit methods, say RK-4. So we need an A-stable implicit scheme. Recall for Backward Euler or Crank-Nicolson, $Re(z) < 0$ is all in ROS.

How about LTE?

$$LTE^n = D^+ u(x_j, t_n) - D_- D_+ u(x_j, t_n) \sim O(h^2 + k).$$

So the method is indeed consistent, $\lim_{k,h \to 0} LTE^n = 0$.

And how about convergence? Suppose $\underline{u}^{n+1} = \underline{\underline{B}}\,\underline{u}^n + \underline{f}^n$. For Forward Euler, $\underline{\underline{B}} = \underline{\underline{I}} + k\underline{\underline{A}}$, $\underline{f}^n = \underline{0}$. Suppose $\underline{u}$, numerical solution, solves it exactly, and the exact solution has LTE $\underline{U}^{n+1} = \underline{\underline{B}}\,\underline{U}^n + \underline{f}^n + k\underline{\tau}^n$. Subtracting the two, $\underline{e}^{n+1} = \underline{\underline{B}}\,\underline{e}^n + k\underline{\tau}^n$. Iterating this equation, we have that $\underline{e}^n = \underline{\underline{B}}^n \underline{e}_0 + k\sum_{j=1}^{n} \underline{\underline{B}}^{n-j}\underline{\tau}^{j-1}$. We'll suppose $\underline{e}^0 = \underline{0}$, we wish to control $\underline{\underline{B}}^n$.

Def: A numerical scheme $\underline{u}^{n+1} = \underline{\underline{B}}\,\underline{u}^n + \underline{f}^n$ for computing a solution up to time T is Lax-Richtmyer stable if $\|\underline{\underline{B}}^n\| \leq C(T)$ for k sufficiently small and all n s.t. $nk \leq T$.

In practice, show $\|\underline{\underline{B}}\| \leq 1 + Ck$ for a constant C independent of k.

$$\|\underline{\underline{B}}^N\| \leq \|\underline{\underline{B}}\|^N \leq \left(1 + \frac{T}{N}\cdot\frac{1}{T}C\right)^N \sim e^{C/T} \quad \text{as } n \text{ grows.}$$

Theorem: Lax-Richtmyer Equivalence: A linear scheme is convergent if and only if it is consistent and Lax-Richtmyer stable.

So for our former scheme, $\underline{\underline{B}} = \underline{\underline{I}} + k\underline{\underline{A}}$, we want its norm less than or equal to 1. Due to symmetry, requires $|1 + k\lambda_j(\underline{\underline{A}})| \leq 1$, ensured via $k|\lambda_{min}(\underline{\underline{A}})| \leq 2$ $\Rightarrow k \leq \frac{h^2}{2}$, same as A-stability requirement (when λ's real, negative).

So how do we verify convergence?

1) Fix h_{min}, fix $k \le h_{min}^2/2$, compare errors for $h = h_{min}$, $2h_{min}$, $4h_{min}$, $8h_{min}$, ...

2) If we let $k \ll h^2$, will only see error in h^2. So when refining k, choose h s.t. $h \sim \sqrt{2k}$.

3/1/23:

An alternative notion of stability (not necessarily sufficient) is Von Neumann stability.

- Ignore bdry conditions
- For linear differential equations, $e^{i\omega x}$ is an eigenfunction
  - ex. $(e^{i\omega x})_{xx} = -\omega^2 e^{i\omega x}$
- Want to make sure these eigenfunctions don't grow in time.
- Suppose $\underline{u}^n = e^{i\omega x} \to u_j^n = e^{i\omega jh}$, then $u_j^{n+1} = g(\omega) e^{i\omega jh}$
- $g(\omega)$ is the amplification factor
- Scheme is Von-Neumann stable if $|g(\omega)| \le 1$
- Again, note it is a necessary but not sufficient condition for stability

Now, we'll focus on hyperbolic problems

- ex: $u_t + a u_x = 0$, $u(x,0) = u_0(x)$, $a \in \mathbb{R}$, $x \in [0, 2\pi)$
  with periodic boundary conditions.
  Exact soln given by $u(x,t) = u_0(x-at)$
- Harder problems include variable wavespeed, nonperiodic boundary conditions, and nonlinearities, but we can mostly understand those thru the simple example

Define the domain of dependence $D(x,t)$ as the set of points s.t. $u(x,t)$ only depends on $u(D(x,t), 0)$.

- For above problem, $D(x,t) = \{x-at\}$, a single point
- For heat equation, $D(x,t) = \mathbb{R}$, depends on all other points

Note: Parabolic problems have infinite propagation speed, $D(x,t) = \mathbb{R}^{(n)}$, while hyperbolic methods do not.

$u_t + a u_x = 0$

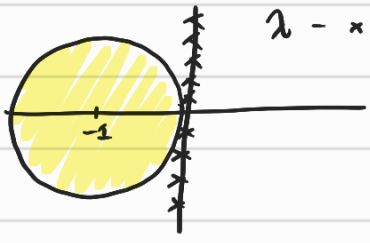
- Equidistant $h = \Delta x$, $k = \Delta t$, $u_j^n \approx u(x_j, t_n)$, $\underline{u}^n = (u_0^n, \dots, u_{M-1}^n)^T$ $j = 0, \dots, M$, $n = 0, \dots, N$
- First, $u_x(x_j, t_n) \mapsto D_0 u_j^n$

- Results in $\frac{\partial}{\partial t}\underline{u}(t) = \underline{\underline{A}}\,\underline{u}$, $\underline{u}(0) = \underline{u}_0$

$$\underline{\underline{A}} = \frac{-a}{2h}\begin{bmatrix} 0 & 1 & & & -1 \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \\ 1 & & & -1 & 0 \end{bmatrix}$$

$$\lambda_j(\underline{\underline{A}}) = \frac{-ia}{h}\sin\left(\frac{2\pi j}{M}\right), \quad j \in [M]$$

- Not stiff, $\max|\lambda_j| \sim |a| \cdot M$
- If use FE, absolute stability says $z = \lambda k$, require $|z+1| \leq 1$, but <u>not stable</u> for any $k > 0$ as eigenvalues pure imaginary



- For other explicit methods, will obtain stability arguments like $|a|k \leq h$
- Should not use RK4, we have $O(h^8)$ space LTE, won't see $O(k^4)$ improvements unless h very small. Should not use Crank-Nicolson, using implicit method too much for non-stiff problem
- Lax-Richtemeyer tells us we have stability if $k = \left(\frac{h}{a}\right)^2$
- <u>Leapfrog method</u>: $\underline{u}^{n+1} = \underline{u}^{n-1} + 2k\underline{\underline{A}}\,u^n$

One idea to get rid of accumulated errors, add dissipation term as $u_t + au_x = \varepsilon u_{xx}$ for $0 < \varepsilon \ll 1$, surprisingly effective in practice

- Choosing ε too small will make no difference
- Choosing ε too big will introduce too much dissipation, problem becomes stiff, introduce $k \leq h^2$ stability requirement
- Will insist on FE, want eigenvalues of $(\underline{\underline{A}} + \underline{\underline{\tilde{A}}}_\varepsilon)$ satisfy $|1+z| \leq 1$.
- This spectrum is <u>actually computable</u>

<u>Lax-Friedrichs scheme</u> - Take $\varepsilon = \frac{h^2}{2k}$, and you get scheme

$$D^+ u_j^n = -aD_0 u_j^n + \frac{h^2}{2k} D_+ D_- u_j^n$$

$$\text{or} \quad u_j^{n+1} = \frac{1}{2}(u_{j-1}^n + u_{j+1}^n) - ka\,D_0 u_j^n$$

<u>3/13/23</u>: Note that ↑ is a standard FE move plus an averaging term.

Another scheme called <u>Lax-Wendroff</u> where you choose "just enough" dissipation rather than the largest possible amount, $\varepsilon = \frac{a^2 k}{2}$:

$$D^+ u_j^n = -a D_0 u_j^n + \overbrace{\frac{a^2 k}{2}}^{\varepsilon} D_+ D_- u_j^n$$

which has a higher order LTE than Lax-Friedrichs: $O(k^2, h^2)$.

Another popular class are <u>upwind schemes</u>. Since the physics of the problem are not symmetric, try using D_+ or D_- instead of D_0 for the spacial derivative.

- If $a > 0$, solution travels to right, look at things behind the particle, <u>use D_-</u>
- If $a < 0$, <u>use D_+</u>

Upwind Scheme: $$D^+ u_j^n = -a D_{+/-} u_j^n,$$

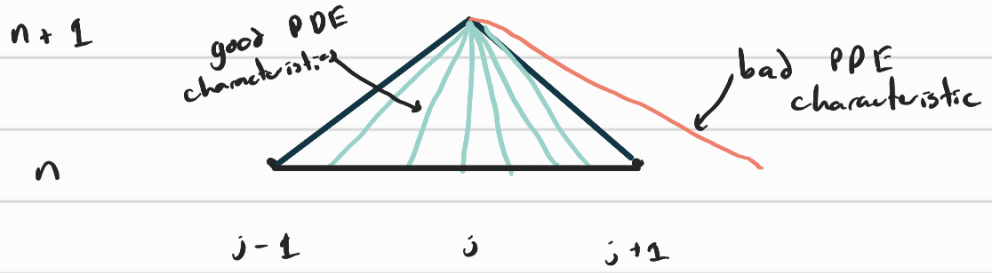
Facts:

- If use wrong sign, unconditionally unstable
- If use right sign, stable iff $\left|\frac{ak}{h}\right| \leq 1$

Note: We keep seeing the condition $\left|\frac{ak}{h}\right| \leq 1$. Why is that?

In the (x,t) plane the discretization step has slope $\frac{\Delta t}{\Delta x} = \frac{k}{h}$. And the PDE characteristics, $X(t) = at + x_0$, have slope $\frac{\Delta t}{\Delta x} = \frac{1}{|a|}$. So, $\left|\frac{ak}{h}\right| \leq 1$ is equivalent to $\frac{k}{h} \leq \frac{1}{|a|}$, numerical characteristic have smaller slope than the PDE characteristics. Or,

- The interval $[x_{j-2}, x_{j+2}]$ <u>contains</u> $D(x_j, t_{n+2})$ for time t_n



Define the <u>numerical domain of dependence</u> $\tilde{D}(x_j, t_n)$, as the spacial size of the stencil. Ex for D_0, $\tilde{D}(x_j, t_n) = [x_{j-1}, x_{j+1}]$

The <u>Courant-Friedrichs-Lewy (CFL) condition</u> says that the numerical domain of dependence contains the analytical domain of dependence. A necessary condition as $k, h \to 0$.

Spectral Methods

Recall, D_0 is a 3 point stencil. By playing w/ Taylor series, we can increase stencil size to increase LTE accuracy.

Consider a function $u: [0, 2\pi] \to \mathbb{C}$, the Fourier series decomposes u into

$$u(x) = \sum_{k=-\infty}^{\infty} \frac{\hat{u}_k}{\sqrt{2\pi}} e^{ikx} \qquad \left(\phi_k(x) = \frac{e^{ikx}}{\sqrt{2\pi}} \right)$$

where we find $\hat{u}_k = \langle u, \phi_k \rangle = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} u(x) e^{-ikx} \, dx.$

On a computer, truncate by $u(x) \simeq u_N(x) := \sum_{k=-N}^{N} \hat{u}_k \phi_k(x).$

Questions: How well does u_N approximate u? What if we approximate $\hat{u}_k$'s?

3/15/23:

Introduce the idea of a projector, a map $P: L^2 \to V \subseteq L^2$ s.t. $P^2 = P$. $u \mapsto Pu$ projects u onto V. $u \mapsto (1-P)u$ projects u onto W s.t. $V \oplus W = L^2$. P is orthogonal if $V \perp W$ or $P = P^*$ ($\langle Pu, v \rangle = \langle u, P^* v \rangle$). V is the range, W is the kernel.

Note: (oblique) projectors can have arbitrarily large norm.

Theorem: Define P_N as the operator

$$P_N u = u_N = \sum_{|k| \le N} \hat{u}_k \phi_k(x), \qquad u \overset{L^2}{=} \sum_{k \in \mathbb{Z}} \hat{u}_k \phi_k(x).$$

Then P_N is an orthogonal projection operator.

Can we bound $\|u - P_N u\|_2$?

$$\|u - P_N u\|_2^2 = \sum_{|k| > N} |\hat{u}_k|^2.$$

And, $\hat{u}_k = \langle u, \phi_k \rangle = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} u(x) e^{-ikx} \, dx$

$$= \frac{i}{k\sqrt{2\pi}} u(x) e^{-ikx} \Big|_0^{2\pi} - \frac{i}{k\sqrt{2\pi}} \int_0^{2\pi} u'(x) e^{-ikx} \, dx$$

Now, assume $u(0) = u(2\pi)$ to get rid of first part. And

the remaining part is $\frac{-i}{k}\widehat{u_k'}$, Fourier coefficient for $u'(x)$, assuming $u'(x) \in L^2$. And $\left|\frac{-i}{k}\widehat{u_k'}\right| \to 0$ as $k \to \infty$. So

$$\|u - P_N u\|_2^2 = \sum_{|k|>N} |\hat{u}_k|^2 = \sum_{|k|>N} \frac{1}{|k|^2}|\widehat{u_k'}|^2 \le \frac{1}{N^2}\sum_{|k|>N}|\widehat{u_k'}|^2 \le \frac{\|u'\|_2^2}{N}$$

Theorem: Suppose $u, u' \in L^2$ and $u(0) = u(2\pi)$. Then

$$\|u - P_N u\|_2 \le \frac{1}{N}\|u'\|_2.$$

Given $s \in \mathbb{N}_0$, the (L_2 periodic) **Sobolev space** is given by

$$H_p^s([0,2\pi];\mathbb{C}) := \left\{ f:[0,2\pi]\to\mathbb{C} \;\middle|\; \begin{array}{l} f^{(k)} \in L^2([0,2\pi];\mathbb{C}),\ 0\le k\le s \\ f^{(k)}(0) = f^{(k)}(2\pi),\ 0\le k\le s-1 \end{array} \right\}$$

with norm $\|u\|_{H^s}^2 := \sum_{k=0}^{s} \|u^{(k)}\|_2^2$

Note: The higher s is, the more "smooth" it is.
And $H^r \subseteq H^s$ for $r > s \ge 0$.

Theorem: If $u \in H_p^s$, then $\|u - P_N u\|_{L^2} \le N^{-s}\|u\|_{H_p^s}$.

Our former result is $s = 1$, can show inductively.

Theorem: If $u \in H_p^s$, $\forall\ 0 \le r < s$, $\|u - P_N u\|_{H_p^r} \le N^{-(s-r)}\|u\|_{H_p^s}$.

So we've answered how well we can approximate a function by Fourrer series. Second question is what happens if we don't know Fourrier coefficients, have to approximate P_N.
How do we compute $\hat{u}_k$ in practice? We will likely know u at some points, but not have an explicit formula.

Quadrature: $\hat{u}_k = \frac{1}{\sqrt{2\pi}}\int_0^{2\pi} u(x)e^{-ikx}\,dx \approx \sum_{j=1}^{M} w_j u(x_j)$

with $w_{k,j} = \frac{\sqrt{2\pi}}{M} e^{-ikx_j}$, $x_j = \frac{2\pi(j-1)}{M}$

giving us Trapezoidal rule or equidistant, assume $M = 2N+1$.

So set $\hat{u}_k \approx \tilde{u}_k := \sum_{j=1}^{M} w_{k,j}\, u(x_j)$.

$\underline{u} = \begin{pmatrix} u(x_1) \\ \vdots \\ u(x_M) \end{pmatrix}$, $\underline{\tilde{u}} = \begin{pmatrix} \tilde{u}_{-N} \\ \tilde{u}_{-N+1} \\ \vdots \\ \tilde{u}_N \end{pmatrix} \Rightarrow \underline{\tilde{u}} = \underline{\underline{\tilde{V}}}^* \underline{u}$

where $\underline{\underline{\tilde{V}}} = \sqrt{\frac{2\pi}{M}} \begin{pmatrix} | & | & & | \\ \underline{v}_{-N} & \underline{v}_{-N+1} & \cdots & \underline{v}_N \\ | & | & & | \end{pmatrix}$, $\underline{v}_k = \sqrt{\frac{2\pi}{M}}\, \phi_k(\underline{x})$ ← $\underline{\underline{V}}$

A fairly straight forward computation shows that

$\langle \underline{v}_k, \underline{v}_\ell \rangle = 1$ for $k = \ell$, $\langle \underline{v}_k, \underline{v}_\ell \rangle = 0$ for $k \neq \ell$.

So $\underline{\underline{V}}$ is a unitary matrix, $\underline{\underline{V}}^* = \underline{\underline{V}}^{-1}$. So

$$\underline{u} = \frac{M}{2\pi} \underline{\underline{\tilde{V}}}\, \underline{\tilde{u}}.$$

This invertible map is called the <u>Discrete Fourier Transform</u> (DFT)

$$\underline{\tilde{u}} = \underline{\underline{\tilde{V}}}^*\, \underline{u} \quad , \quad \underline{u} = \frac{M}{2\pi} \underline{\underline{\tilde{V}}}\, \underline{\tilde{u}}.$$

The DFT/ inverse DFT is relatively expensive, $O(M^2)$. The FFT uses recursion to reduce to $O(M \log(M))$ cost.

Now, DFT is a projection operator, but in fact it can be an unstable oblique projector. Define I_N to be the DFT operator. How different is $\|u - I_N u\|$ from $\|u - P_N u\|$?

<u>3/20/23:</u>

Using the triangle inequality, $\|u - I_N u\| \leq \|u - P_N u\| + \|P_N u - I_N u\|$. We know $\|u - P_N u\| \leq N^{-s}$. Call $A_N = P_N - I_N$. If $u \in V_N$, then $A_N u = 0$, so in fact $A_N u = A_N (I - P_N) u$. We know $(I - P_N) u$ is "small", and can show that A_N is well-behaved on small inputs.

<u>Theorem:</u> Assume $u \in H^s_p$ with $s > \frac{1}{2}$. Then

$$\|u - I_N u\|_{L^2} \leq N^{-s} \|u\|_{H^s}$$

$$\|u - I_N u\|_{H^r} \leq N^{-(s-r)} \|u\|_{H^s} \quad \text{if } r < s.$$

Now, consider the abstract PDE $\mathcal{L}(u) = f$. Also define the residual $R(u) = \mathcal{L}(u) - f$. Also assume 2π periodicity. Make the ansatz

$$u(x) \approx u_N(x) = \sum_{|k| \le N} \hat{u}_k e^{ikx}.$$

Due to finiteness of N, we won't be able to get $R(u_N) = 0$. First, let's assume $R(u)$ is linear in u and that R and u are independent of time. $R(u) = 0$ means u is a strong solution. For example, $u_t + u_x = 0$, $u(x,0) = \sin(x)$ has a solution $\sin(x-t)$, so we can enforce the residual strongly. But for $u(x,0) = H(x)$, ($H(x) = 0$ for $x < 0$, $H(x) = 1$ for $x \ge 1$). We can enforce it if we ignore $t = 0$, $H(x)$ is a solution, if we ignore $x = t$, $H(x-t)$ is a solution. Both are valid almost everywhere, everywhere except along a line in x-t space. Hence, solutions are not unique here. Hence, enforcing $R(u) = 0$ strongly almost everywhere can be a bad plan.

Instead of requiring $R(u_N) = 0$, we'll require $\langle R(u_N), \phi_k \rangle = 0 \ \forall k$ or that the residual is orthogonal to Fourier space.

Let H be a Hilbert space, and H^* be the topological dual space or H, continuous bounded linear functionals on H. An example in H^* is $h \mapsto \langle h, h_\phi \rangle$ for some $h_\phi \in H$. The Riesz Representation Theorem essentially tells us that all elements of H^* are of the form $h \mapsto \langle h, h_\phi \rangle$ or that $\exists$ an isomorphism H to H^*.

Now, consider a subspace $V \subseteq H$. The dual of V with respect to the inner product on H is the collection of objects w such that $v \mapsto \langle v, w \rangle$ is continuous for every $v \in V$. This condition is looser, so $H^* \subseteq V^*$. So, we have that $V \subseteq H = H^* \subseteq V^*$. We call (V, H, V^*) a Gelfand triple or a rigged Hilbert space. w need not lie in H, but we will consider $\langle v, w \rangle$, "throw smoothness of v onto w".

3/22/23:

In our Gelfand triple (V, H, V^*), think of V as fns, and V^* our derivative space, contains L^2 but may have elements not in L^2.

Ex: Consider $H = L^2([0, 2\pi]; \mathbb{C})$ with it's standard inner product $\langle \cdot, \cdot \rangle$. Define

$$V := H_p^1([0, 2\pi], \mathbb{C}) = \left\{ f = \sum_{k \in \mathbb{N}} c_k \phi_k(x) \in H \;\middle|\; f' \in H,\ f(0) = f(2\pi) \right\}$$

is a subspace of H.

Claim: $H_p^{-1} := V^*$ is the space whose first "Fourier" antiderivative is in L^2.

Why? Let $v \in V$, and some w satisfying $\langle w, \phi_0 \rangle = 0$ with antiderivative W.

$$(v, w) = \langle v, w \rangle \overset{\text{int. by parts}}{=} -\langle v', W \rangle.$$

Hence, if $W \in L^2$, then

$$|(v, w)| = |\langle v', W \rangle| \leq \|v'\|_{L^2} \|W\|_{L^2} \leq C(w) \|v\|_{H_p^1},$$

hence $v \mapsto (v, w)$ is a bounded map, thus $w \in V^*$ given $\langle w, \phi_0 \rangle = 0$ and $W \in L^2$ where $W' = w$.

What kind of functions are in V^*? Consider

$$w(x) = \sum_{k \in \mathbb{Z}} \overline{\phi_k(0)}\, \phi_k(x) = \overline{\phi_0(0)}\, \phi_0(x) + \sum_{k \geq 0} \overline{\phi_k(0)}\, \phi_k(x)$$

$$= \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} e^{ikx}.$$

Note this is nowhere convergent. But, the Fourier antiderivative is

$$W_1(x) = \sum_{|k| > 0} \frac{\overline{\phi_k(0)}}{ik} \phi_k(x) \in L^2$$

Fourier antiderivative: Ignore $k = 0$ term

Thus, $$(v, w) = \langle v, \overline{\phi_0(0)}\,\phi_0(x)\rangle + \langle v, w_1\rangle$$
$$= \hat{v}_0 \phi_0(0) - \langle v', w_1\rangle$$
$$= \hat{v}_0 \phi_0(0) - \left\langle \sum_{k\in\mathbb{Z}} ik\hat{v}_k\phi_k, \sum_{|\ell|>0} \frac{\overline{\phi_\ell(0)}}{i\ell}\phi_\ell \right\rangle$$
$$= \hat{v}_0\phi_0(0) + \sum_{|k|,|\ell|>0} \frac{k}{\ell}\phi_\ell(0)\hat{v}_k\langle\phi_k,\phi_\ell\rangle$$
$$= \sum_{k\in\mathbb{Z}} \hat{v}_k\phi_k(0) = v(0).$$

This tells us that $w = \delta_0$, the Dirac delta centered at 0, is an element of $H_p^{-1} = V^*$. And inner product is point evaluation.

Let's consider an example, $-u''(x) + u(x) = f(x)$, $u(0) = u(2\pi)$. We'll find $u \in V \subseteq L^2$ s.t. $\langle \mathcal{L}u, v\rangle = \langle f, v\rangle$ $\forall$ $v \in V$ where $\mathcal{L}u = -u''(x) + u(x)$. It is not necessary that $u \in V \Rightarrow u'' \in V$, so consider a Gelfand triple (V, L^2, V^*). Now,

$$v \in V,\ \mathcal{L}u \in V^* \Rightarrow \langle \mathcal{L}u, v\rangle = (-u'', v) + \langle u, v\rangle$$
$$\overset{\text{I.B.P \& BCs}}{=} \langle u', v'\rangle + \langle u, v\rangle.$$

So, our new PDE statement is "Find $u \in V$ s.t. $\langle u', v'\rangle + \langle u, v\rangle = \langle f, v\rangle$ for every $v \in V$". This is a <u>weak form</u> or <u>variational form</u> for the PDE. And a solution u is called a <u>weak solution</u>. In addition, if u is a bona fide strong solution (satisfied $\forall x$), it must also be a weak solution. Converse need not be true.

In this example, u need not be differentiable twice, it only must be differentable once in a weaker (averaging) integral sense.

<u>Q</u>: Do weak solns exist in general? Is this well-posed?

Let $a(\cdot,\cdot) : V \times V \to \mathbb{C}$ be a <u>sesquilinear form</u> (linear in first arg, conjugate linear in second). It is <u>coercive</u> or <u>elliptic</u> if $\exists$ $c > 0$ such that
$$|a(v,v)| \geq c\|v\|_V^2 \quad \forall \quad v \in V.$$

It is bounded or continuous if $\exists\ C>0$ s.t. $|a(u,v)| \le C\|u\|\cdot\|v\|$ $\forall\ u,v \in V$.

Lax-Milgram Theorem: Let $a(\cdot,\cdot)$ be a (w/ constant c) coercive and bounded sesquilinear form on a Hilbert space V, and let (V, H, V^*) be a Gelfand triple w/ an H inner-product $\langle\cdot,\cdot\rangle$. Let $f \in V^*$. Then there exists a unique solution $u \in V$ to the problem

"Find $u \in V$ such that $a(u,v) = \langle f, v\rangle$ for every $v \in V$."

Moreover, the solution is well-posed with respect to f.

$$\|u\|_V \le \frac{1}{c}\|f\|_{V^*}.$$

This theorem gives us existence & uniqueness for an infinite-dimensional analogue of

$$Au = f, \quad A = A^*.$$

- $a(u,v)$ is ∞-dim version of v^*Au
- coerciveness & boundedness are analogues to $\sigma_{min}(A) > 0$, $\sigma_{max}(A) < \infty$ (singular values) (cont. map, inverse exists)
- $f \in V^*$ equivalent to $v \mapsto f^*v$ bounded
- In finite-dim, since norms equivalent, $V = H = V^*$.
- $a(u,v) = \langle f, v\rangle$ is analogous to $v^*Au = v^*f$.

In former problem, $-u'' + u = f$, $u(0) = u(2\pi)$ we let

$$H = \left\{u(x) = \sum_{k\in\mathbb{Z}} c_k \phi_k(x) \;\middle|\; \|u\|_{L^2} < \infty\right\}, \quad V = \{u \in H \mid u' \in H\}$$

with $\langle\cdot,\cdot\rangle$ standard inner-product on $L^2[0, 2\pi]$, $\|u\|_H^2 = \langle u, u\rangle$, $\|u\|_V^2 = \langle u', u'\rangle + \langle u, u\rangle$. Define the bilinear form $a(u,v) = \langle u', v'\rangle + \langle u, v\rangle$ which we can show to be coercive ($c=1$) and bounded. Then, by Lax-Milgram, there exists a unique weak solution $u \in V$ s.t. $a(u,v) = \langle f, v\rangle$ $\forall\ v \in V$ if $f \in V^*$ and $\|u\|_V \le \|f\|_{V^*}$.

Now, let V_N be a finite dimensional subspace of V. Now, Lax-Milgram yields a unique solution u_N.

How accurate is this solution? What is $\|u - u_N\|_V$?

Céa Lemma: $$\|u - u_N\|_V \leq \frac{C}{c} \inf_{v \in V_N} \|u - v\|_V$$

where C & c are continuity & coercive constants respectively of a.

Note, from linear algebra, $\frac{C}{c} = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$ = condition number of A.

3/27/23:

Gelfand Aside: We choose $V = H^1_p$, so $v \in V$ must satisfy $\|v\|_{H^1} = \|v\|_{L^2} + \|v'\|_{L^2} < \infty$, or writing as a fourier series, $\|v\|_{H^1} = \sum (1 + k^2) \hat{v}_k^2 < \infty$. This means v must be smoother than just $u \in L^2$ periodic where $\sum \hat{u}_k^2 < \infty$. Now, $w \in H^*$ $\Rightarrow v \mapsto \langle v, w \rangle$ bounded, essentially implies $\sum \frac{\hat{w}_k^2}{1+k^2} < \infty$, so w can be very unsmooth, ex. $\hat{w}_k = 1 \ \forall k \Rightarrow w = \delta_0$. We can let $f \in V^* = H^{-1}$ so $\langle f, v \rangle$ makes sense. So, it kind of means $V = H^1_p$ = fns s.t. fourier derivative is in L^2, $V^* = H^{-1}_p = L^2$ = fns s.t. fourier integral is in L^2.

This theory is limited to linear PDEs, typically for elliptic or parabolic PDEs. Can be extended but more complicated.

- Hilbert structure not required, V can be Banach w/ dual V^*
- $a(\cdot, \cdot)$ can operate on different spaces $U \times V$

Problem: Given $\mathcal{L}u = f$, define residual $R(u) = \mathcal{L}u - f$, a generic weak formula asks: "Find $u \in V$ (or U) satisfying $\langle R(u), v \rangle = 0$ $\forall v \in V$" where V is some Hilbert (or Banach) space with some inner-product $\langle \cdot, \cdot \rangle$ (or duality pairing).

This is called a weighted residual method as we enforce the residual vanishes in some weighted sense.

Call V (or U) the trial space from which we pick u, the space in which we enforce zero residual (also V here) is the test space.

Trial = Test space $\Rightarrow$ Galerkin method, otherwise Petrov-Galerkin.

Ex: Choose $V = \text{span}\{\delta_{x_0}, \delta_{x_1}, \ldots, \delta_{x_m}\}$, called a collocation scheme. It essentially only forces residual = 0 at collocation points ($R(u)|_{x=x_m} = 0$).

Our freedoms then are, what kinds of function is our solution? What is our trial space U? And, how do we want to satisfy the PDE, what is the test space V?

Now, we'll move onto Fourier Spectral Methods. Consider time independent 2. We'll choose trial space U, test space V, as spaces of Fourier functions, and (almost always) L^2 inner-product.

Fourier-Galerkin Method:

- $U = V$, Galerkin
- $V = \text{span}\{e^{ikz} : k \in \mathbb{Z}\}$ in 1D w/ periodic BCs on $0, 2\pi$
- Given $\mathcal{L}u = f$, choose $a(u,v) = \langle f, v\rangle$
- Prob: Find $u_N \in V_N = \text{span}\{e^{ikz} : k \in \mathbb{Z}, |k| \le N\} \subset V$
- $\phi_k(z) = \frac{1}{\sqrt{2\pi}} e^{ikz}$
- $\langle -u'' + u, v\rangle = \langle u', v'\rangle + \langle u, v\rangle$ from I.B.P., so

  $a(u,v) = \langle u', v'\rangle + \langle u, v\rangle$
- Choose $V = H^1_p$, then $V^* = H^{-1}_p$.
- We have Gelfand triple, this problem well-posed by Lax-Milgram
- $\sup_{u,v \in V} \frac{|a(u,v)|}{\|u\|_V \|v\|_V} \le \sqrt{2} := C$, $\inf_{v \in V} \frac{|a(v,v)|}{\|v\|_V^2} = 1 = c$.
- Equivalent to ansatz $u_N(x) = \sum_{|k| \le N} \hat{u}_k \phi_k(x)$.
- Weak prob: $\left\langle \sum_{|l| \le N} (il)\hat{u}_l \phi_l, (ik)\phi_k \right\rangle + \langle u_N, \phi_k\rangle = \langle f, \phi_k\rangle$

$$\Rightarrow \sum_{|k| \le N} (k^2 + 1)\phi_k^2(x) = \hat{f}_k := \langle f, \phi_k\rangle$$

$$\Rightarrow \underline{\underline{D_2}}\,\underline{u} = \underline{\hat{f}}, \quad \underline{\underline{D_2}} = \text{diag}\left((-N)^2+1, (-N+1)^2+1, \dots, (N)^2+1\right)$$

$$\Rightarrow \underline{u} = \underline{\underline{D_2}}^{-1}\underline{\hat{f}}.$$

- Céa Lemma: $\|u - u_N\|_V \le \sqrt{2} \inf_{v \in V_N} \|u - v\|_V \le \sqrt{2}\|u - P_N u\|_V \le \sqrt{2}\, N^{1-s}$

(C/c); Projection; L^2 orthogonal projection on $V = H^1_p$; Fourier approx error

> $f \in H_p^{-1} \Rightarrow u \in H_p^{-1+2} = H_p^1 \Rightarrow s = r+2$ (space u lives in)

> $f \in H_p^r \Rightarrow \|u - u_N\|_{H^2} \leq \sqrt{2}\, N^{1-(r+2)} \sim N^{-(r+1)}$

> The smoother f is, the faster Fourier series converge.

3/29/23

Now, we'll use a collocation method for

$$-u'' + u = f, \quad 2\pi \text{ periodic}, \quad \text{Fourier series}$$

We'll define our test space

$$V = \text{span}\{\delta_{x_m}\}_{m \in [2N+1]}, \quad x_m = \frac{2\pi(m-1)}{2N+1}.$$

We essentially enforce $-u''(x_m) + u(x_m) = f(x_m) \quad \forall\ m$.

Define $$\tilde{D}_2 = \frac{M}{2\pi} \tilde{V} \begin{bmatrix} (-N)^2 & & & \\ & (-N+1)^2 & & \\ & & \ddots & \\ & & & N^2 \end{bmatrix} \tilde{V}^*$$

from which $\tilde{D}_2 \underline{u} = \underline{u}''$, a pointwise second derivative operator.

Note: This matrix is dense. Instead of storing all of $\tilde{D}_2$, we can use FFT, and do a diagonal multiplication, for $O(N \log N)$.

Our scheme reads $$(-\underline{\underline{D}}_2 + \underline{\underline{I}})\,\underline{u} = \underline{f},$$ linear system, can be solved for $\underline{u}$ easily.

> Easy to implement, more theory, dense matrix.

How about error?

Galerkin	Collocation
$P_N(-u_{N,G}'' + u_{N,G}) = P_N f$	$I_N(-u_{N,C}'' + u_{N,C}) = I_N f$

Note, $I_N(-u_{N,C}'' + u_{N,C}) = P_N(-u_{N,C}'' + u_{N,C})$. So subtracting the two,

$$P_N(-\Delta u_N'' + \Delta u_N) = (P_N - I_N) f = \underbrace{A_N f}_{\text{aliasing error}}$$

Since $\|P_N\| = 1$ (orthogonal), assuming $f \in H_p^r$.

$$\|\Delta u_N\|_2 \leq \|A_N f\|_{L^2} \leq N^{-r}$$

Recall, I_N is interpolating Fourier series (DFT), and P_N is orthogonal projection onto $V_N = \text{span}\{\phi_k\}_{|k|\le N}$.

Notes between Galerkin & Collocation:

- Both behave very similarly for smooth u.
- In general, easier to prove Galerkin Estimates
- In some cases, collocation is easier

ex: $$-u'' + \frac{u}{2+\sin x} = f(x)$$

pointwise evaluation of above vs. inner products. A pseudospectral method looks like

$$P_N(-u_N''(x)) + I_N\left(\frac{u_N(x)}{2+\sin(x)}\right) = P_N(f(x)).$$

Now, we move into time-dependent problems. Consider

$$\frac{\partial u}{\partial t} = \mathcal{L}u, \quad \text{periodic BCs}$$

where $\mathcal{L}$ is time independent, typically linear. (wave eq $\mathcal{L}u = cu_x$, heat eq $\mathcal{L}u = ku_{xx}$). We'll focus on stability, and strong-form or collocation methods.

Ex: $\frac{\partial u}{\partial t} = c\frac{\partial u}{\partial x} + \nu\frac{\partial^2 u}{\partial x^2}$, $c \in \mathbb{R}$, $\nu > 0$, assume ICs smooth.

This PDE satisfies an energy stability property

$$\int_0^{2\pi} u u_t \, dx = c\int_0^{2\pi} u u_x \, dx + \nu \int_0^{2\pi} u u_{xx} \, dx$$

$uu_t = \frac{1}{2}(u^2)_t$, $uu_x = \frac{1}{2}(u^2)_x$, BCs $\Rightarrow$

$$\Rightarrow \frac{d}{dt}\|u\|_{L^2}^2 = 0 - \nu\|u_x\|_{L^2}^2 \le 0.$$

And we'll use a Fourier-Galerkin scheme in space, discretize time.

$u_N(\cdot, t) \in V_N \quad \forall\ t.$

Residual equality ($\forall\ v \in V_N$): $$\left\langle \frac{\partial u_N}{\partial t}, v \right\rangle = c\left\langle \frac{\partial u_N}{\partial x}, v \right\rangle + \nu\left\langle \frac{\partial^2 u_N}{\partial x^2}, v \right\rangle.$$

Choosing $v = \phi_k \ \forall \ |k| \le N$, we get the $(2N+1)$ size ODE

$$\frac{d}{dt}\hat{\underline{u}} = c\,\hat{\underline{\underline{D_1}}}\,\hat{\underline{u}} + \nu\,\hat{\underline{\underline{D_2}}}\,\hat{u}$$

with $\hat{\underline{\underline{D_1}}} = \text{diag}(-iN, -i(N-1), \dots, iN)$, $\hat{\underline{\underline{D_2}}} = \text{diag}(-N^2, \dots, -N^2)$.

Yields exact solution $\hat{u}_k(t) = \hat{u}_k(0)e^{(-\nu k^2 + ick)t}$

which is nonincreasing in time. $\hat{u}_k(t) = \langle u, \phi_k \rangle$.

Ex: $u_t = \sin(x)\,u_x$

Facts: $\phi_k \phi_\ell = \frac{1}{\sqrt{2\pi}}\phi_{k+\ell}$, so $\sin x\,\phi_k = -i\sqrt{\frac{\pi}{2}}(\phi_1 - \phi_{-1})(ik)\phi_k$
$= \frac{k}{2}(\phi_{k+1} - \phi_{k-1})$

Now, we can find $\langle \sin x \cdot u, v \rangle$ for $u, v \in V_N$.

Yields ODE

$$\frac{d}{dt}\hat{\underline{u}} = \underline{\underline{A}}\,\hat{\underline{u}}, \qquad \underline{\underline{A}} = \begin{bmatrix} 0 & \frac{-N}{2} & & \\ \frac{N-1}{2} & 0 & \frac{-(N-1)}{2} & \\ & \ddots & \ddots & \ddots \\ & & \frac{N}{2} & 0 \end{bmatrix}.$$

Now, consider $u_t = uu_x$. let $u_N \in V_N$,

Fourier Galerkin: $\left\langle \frac{\partial u_N}{\partial t}, v \right\rangle = \left\langle u_N \frac{\partial u_N}{\partial x}, v \right\rangle \quad \forall\, v \in V_N.$

We then need to be able to compute a triple product of ϕs here. Can be done, just get ugly coefficients. End with $\frac{d}{dt}\hat{\underline{u}} = \underline{a}(\hat{\underline{u}})$ where $\underline{a}$ is nonlinear, quadratic in fact. End up w/ convolutions, so $\hat{\underline{u}} \to \underline{a}(\hat{\underline{u}})$ is an order N^2 operation.

Ex: $\frac{\partial u}{\partial t} = \sin(u)\frac{\partial u}{\partial x}$, in this case, cannot exactly compute anymore.

Takeaway: Things get difficult quick with full Fourier-Galerkin. For previous example, may use pseudospectral method. Evaluate $\sin(u_N)$ pointwise, and use DFT to get coefficients.

The (Petrov-Galerkin) collocation setup involves choosing M points $x_1, \cdots, x_M$, and using ansatz

$$u_N(x,t) = \sum_{k=1}^{M} u_k(t)\, l_k(t)$$

(in V_N, $l_k(x_j) = \delta_{kj}$, don't need explicitly)

where $l_k(t)$ is the cardinal Lagrange function at x_k.
Our scheme then asks for zero residual at $x_1, \cdots, x_K$.

$$\frac{d}{dt}\underline{u} = c\,\underline{\underline{\tilde{D}_1}}\,\underline{u} + r\,\underline{\underline{\tilde{D}_2}}\,\underline{u}$$

where $\underline{\underline{\tilde{D}_1}}$, $\underline{\underline{\tilde{D}_2}}$ are dense, discrete differentiation, collocation matrices corresponding to $u \xrightarrow{\text{DFT}} \hat{u}_k \xrightarrow{\text{Fourier diff}} \widehat{u_k^{(1,2)}} \xrightarrow{\text{IDFT}} u^{(1,2)}$.

Ex: $u_t = \sin(u)\, u_x$ is much easier w/ collocation.

$$\frac{d}{dt}\underline{u} = (\sin(\underline{u})) \odot \left(\underline{\underline{\tilde{D}_1}}\,\underline{u}\right)$$

where $\odot$ is elementwise multiplication.

Note: As long as have $\underline{\underline{\tilde{D}}}$ matrices, easier to implement collocation.
But much more difficult to prove things like stability.

Consider $u_t = \mathcal{L}u$ with periodic BCs on $[0, 2\pi]$ with $\mathcal{L}$ linear.
We call $\mathcal{L}$ semibounded if $\mathcal{L} + \mathcal{L}^* \le CI$ for some $C \in \mathbb{R}$,
i.e. $\langle (L+L^*)u, u\rangle \le \langle Cu, u\rangle = C\|u\|_{L^2}^2$. Given $\mathcal{L}$ semi-bounded,
we had shown the problem is well-posed w/ $\|u\|_{L^2} \le e^{(1/2)Ct}\|u(0)\|_{L^2}$.

Ex: $u_t = \mathcal{L}u = c(x)\,u_x$, $c(x)$ real & periodic.
We can show thru int. by parts, if $c(x)$ also
differentiable, and has bounded derivative, $\mathcal{L}$ semi-bounded
w/ constant $C = \max|c'(x)|$.

Ex: $\mathcal{L} = \frac{\partial}{\partial x} K(x) \frac{\partial}{\partial x}$ is self-adjoint, so have that
$\mathcal{L} + \mathcal{L}^* = 2\frac{\partial}{\partial x} K(x) \frac{\partial}{\partial x}$, using int. by parts, given K pos, real, semibounded.

Theorem: Assume $\mathcal{L}$ is semi-bounded with $\mathcal{L} + \mathcal{L}^* \leq CI$, and consider the PDE $u_t = \mathcal{L}u$. Then the Fourier-Galerkin method is stable and obeys the bound

$$\|u_N(t)\|_{L^2} \leq e^{(C/2)t} \|u_N(0)\|_{L^2}$$

Comes down to fact P_N is semi-bdd, self-adjoint.

4/5/03:

Now, suppose we have a non-periodic problem. If using above methods, must find new basis $\{\phi_k\}$. Most straightforward choice is polynomials. Define $P_n := \text{span}\{x^j : 0 \leq j \leq n\}$. We'll work in L^∞ w/ norm $\|u\|_{L^\infty} = \sup_{x \in D} |u(x)|$.

Weierstrass Theorem: Assume $u \in C[-1,1]$, then given $\varepsilon > 0$, $\exists\, n \in \mathbb{N}$ and $p_n \in P_n$ s.t. $\|u - p_n\|_{L^\infty} \leq \varepsilon$.

Theorem: Let $u \in C[0,\infty)$, and $\exists\, \delta > 0$ s.t. $\lim_{x \to \infty} u(x)e^{-\delta x} = 0$, then $\forall\, \varepsilon > 0$, $\exists\, n \in \mathbb{N}$, $p_n \in P_n$ s.t. $\|(u(x) - p_n(x))e^{-\delta x}\|_{L^\infty} < \varepsilon$.

i.e.: $u(x)$ cannot grow superexponentially

However, constructing such a polynomial is not feasible computationally because we are working in Banach space L^∞.

We want to work in L^2 instead, introduce notation

$$L^2_\omega(D) := \{u : D \to \mathbb{R} \mid \|u\|_{L^2_\omega(D)} < \infty\}, \quad \|u\|_{L^2_\omega(D)} = \int_D u^2(x)\omega(x)\,dx$$

(inner product)

And with an inner-product, we can minimize algorithmically.

$$p_n(x) = \sum_{j=0}^{n} \hat{u}_j \phi_j(x), \quad \hat{u}_j = \langle u(x), \phi_j(x) \rangle_{L^2_\omega}$$

given a complete orthonormal basis. They can be designed through a **Gram-Schmidt** approach.

$$\tilde{p}_k(x) = x^k - \sum_{j=0}^{k-1} \langle x^k, \tilde{p}_j(x) \rangle_{L^2_\omega} \tilde{p}_j(x),$$

$$p_k(x) = \frac{\tilde{p}_k(x)}{\|\tilde{p}_k(x)\|_{L^2_\omega}}.$$

However, this gets very unstable for $k > 10$, even with improved algorithms like modified Gram-Schmidt.

3-Term Lemma: Suppose $\{p_n\}_{n\geq 0}$ are L^2_ω-orthonormal with $n = \deg(p_n)$. Then there exist constants $a_n, b_n \in \mathbb{R}$ with $b_n > 0$ s.t.

$$x p_n = b_{n+1} p_{n+1} + a_n p_n + b_n p_{n-1}, \quad n \geq 0$$

where $p_{-1} \equiv 0$, $p_0 = \frac{1}{b_0}$, $b_0^2 = \int_D \omega(x)\,dx$.

Proof: Fix n,

$$x p_n = \sum_{j=0}^{n+1} c_j x^j = \sum_{j=0}^{n+1} \langle x p_n, p_j \rangle_{L^2_\omega} x^j = \sum_{j=0}^{n+1} \langle p_n, x p_j \rangle_{L^2_\omega} x^j$$

$$= \langle p_n, x p_{n-1} \rangle_{L^2_\omega} x^{n-1} + \langle p_n, x p_n \rangle x^n + \langle p_n, x p_{n+1} \rangle x^{n+1}$$

$$= \text{——————} \;||\; \text{——————} + \langle x p_n, p_{n+1} \rangle x^{n+1}$$

$$= b_{n-1} p_{n-1} + a_n p_n + b_{n+1} p_{n+1}.$$

And this recursion process is way more stable.

See Christoffel–Darboux Theorem for direct representations of $\sum_{k=0}^{n} p_k(x) p_k(y)$ for $x = y$ or $x \neq y$.

Now, we'd like to integrate a degree m polynomial w/ quadrature $\sum_{j=1}^{n} w_j p_n(x_j)$. Note, p_m has $m+1$ degrees of freedom, $2n$ unknowns.

Define the node polynomial $d(x) = \prod_{j=1}^{n} (x - x_j)$.

Jacobi Theorem: Let $d(x) \in P_n$ be defined as above for nodes $\{x_j\}_{j=1}^{n}$. Let $0 \leq k \leq n$, then the following statements are equivalent.

1) $\int_0 p(x)\omega(x)\,dx = \sum_{j=1}^{n} w_j p(x_j) \quad \forall\ p \in P_{n+k-1}$

2) $\langle d, p \rangle_{L^2_\omega} = 0 \quad \forall\ p \in P_{k-1}$. (quasi-orthogonality condition)

(↖ i.e. $d = p_n(x)$.)

Gaussian Quadrature is for $k = n$, integrate any $p \in P_{2n-1}$.

Proof from writing $p(x) = k(x)q(x) + r(x)$, and we have

$$\int_D p(x)\omega(x)\,dx = \int k(x)q(x)\omega(x)\,dx + \int r(x)\omega(x)\,dx$$

$$= \underbrace{0}_{\text{quad. rule}} + \text{exact evaluation by quad.}$$

Theorem: $L^2_\omega(D)$-orthogonal polynomials on D have (n) simple roots inside D.

So, $\{x_j\}_{j=1}^n = p_n^{-1}(0)$ are the unique sets of nodes for quadrature. weights given by

$$\omega_j = \frac{1}{K_n(x_j)}, \qquad K_n(x) = \sum_{j=0}^{n-1} p_j^2(x).$$

Takeaway: For integrating polynomial up to degree $2n-1$, the nodes given by roots of p_n, and weights given as above.
How to get roots? Recurrence relation says

$$x\vec{p}(x) = \begin{pmatrix} a_0 & b_1 & & & \\ b_1 & a_1 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & b_{n-1} & a_{n-1} \end{pmatrix}\vec{p}(x) + b_n p_n(x)\vec{e_n}, \quad \left(\vec{p}(x) = \begin{pmatrix} p_0(x) \\ \vdots \\ p_{n-1}(x) \end{pmatrix}\right)$$

J, Jacobi Matrix

So $p_n(x_0) = 0 \Rightarrow x_0\vec{p}(x_0) = J\vec{p}(x_0)$. I.e. $p_n^{-1}(0) = \lambda(J)$.

4/10/23:

In fact, given $J\vec{v} = x_j\vec{v}$ w/ $\|\vec{v}\| = 1$, $\omega_j = \vec{v}[1]$, first component of the normalized eigenvectors.

In summary, given $(a_n), (b_n)$, we can use Linear Algebra to compute x_j, ω_j. However, finding eigenvalues for general square matrix is difficult and unstable. If the matrix is hermitian or symmetric, becomes much nicer.

Consider the Sturm-Liouville problem

$$-\frac{\partial}{\partial x}\left[Q(x)\omega(x)\frac{\partial y}{\partial x}\right]-\lambda\omega(x)y(x)=0,$$

special choices of $\omega(x)$, $Q(x)$ yields families of orthogonal polynomials. Let

$$S(y)=\frac{-1}{\omega(x)}\frac{\partial}{\partial x}\left[Q(x)\omega(x)\frac{\partial y}{\partial x}\right],$$

then the above problem becomes $S(y)=\lambda y$, so λ an eigenvalue of S. S also a self-adjoint operator, and positive semi-definite. So eigenvalues real, non-negative. These facts w/ some extra details yield that linearly-independent solutions are L^2_ω-orthogonal.

The Rodrigues' formula says that if ω satisfies the ODE $\omega'=\frac{L(x)}{Q(x)}\omega$, L at most linear, then we have that

$$p_n(x)\propto\frac{1}{\omega(x)}\frac{\partial^n}{\partial x^n}\left(Q(x)^n\omega(x)\right)$$

which lets us explicitly calculate (a_n), (b_n).

Legendre Polynomials:

$\omega(x)=1$, $Q(x)=(1-x^2)$, and $a_n=0$, $b_0=\frac{1}{\sqrt{2}}$, $b_n=\frac{n}{\sqrt{4n^2-1}}$.

Can do same w/ Hermite, $\omega(x)=e^{-x^2}$, get $a_n=0$, $b_0=\frac{1}{\pi^{1/4}}$, $b_n=\sqrt{\frac{n}{2}}$.

Or Jacobi, Laguerre, Chebyshev, ...

Consider the ODE/eigenvalue problem, $-y''(x)=\lambda y(x)$ with periodic BCs.

Denote $S(y)=-\frac{\partial^2}{\partial x^2}y$. Can show we have eigenpairs $\phi_n(x)=\frac{1}{\sqrt{2\pi}}e^{inx}$, $\lambda_n=n^2$. Define projection coefficients $\hat{u}_n=\langle u,\phi_n\rangle$.

Note that S is self-adjoint, so

$$|\hat{u}_n|=|\langle u,\phi_n\rangle|=\frac{1}{\lambda_n}|\langle u,\lambda_n\phi_n\rangle|=\frac{1}{\lambda_n}|\langle u,S(\phi_n)\rangle|$$
$$=\frac{1}{\lambda_n}|\langle S(u),\phi_n\rangle|=\frac{1}{\lambda_n}|\langle u''(x),\phi_n\rangle|=\frac{1}{\lambda_n}|\widehat{(u'')}|.$$ We

have $\lambda_n=\frac{1}{n^2}>\frac{1}{(n+1)^2}=\lambda_{n+1}$, and w/ Parseval, $\sum_{|n|>N}|\hat{u}_n|^2\le\frac{1}{\lambda_N^2}\|u''\|_{L^2}^2$

and if still differentiable, can continue, and we can attain same error bound for $\|u-u_N\|$ as before.

$\|u - u_N\|_{L^2} \le \frac{1}{\lambda_N^s} \|u\|_{H^s_p}$. And $\lambda_N^{-s} = N^{-s}$ for Fourier.
Can do similar procedure for polynomials.

4/12/23:

Define weight function $w(x)$, use L^2_w inner-product, trial/test given by $\text{span}\{p_0, \dots\}$ where p_n's are L^2_w-orthogonal.
But now, we must be careful for boundary conditions.

Ex: $-u''(x) = f(x)$ on $(-1,1)$, $u(-1) = \alpha$, $u(1) = \beta$.

- If $\alpha \neq 0$, nonhomogeneous, solve $\alpha = \beta = 0$ and use lifting.
- Essential handling of BCs says all trial basis functions satisfy the BCs (homogenous)
- The natural handling weakly handles BCs.
- The fav handling involves tweaking high-frequency coefficients, devote higher degrees to freedom to BCs instead of PDE.

Consider $\alpha = 3$, $\beta = -2$, lifting notes that $w(x) = \frac{-5}{2}x + \frac{1}{2}$ satisfies the BCs and $w''(x) = 0$. So given a homogenous solution $U(x)$, a nonhomogeneous solution is $U(x) + w(x)$, lifting.
For essential handling, define $P_{N,0} \subset P_N$, reducing dim by 2, and then use $P_{N,0}$ as trial/test space. ($P_{N,0} = \{p \in P_N : p(-1) = \alpha, p(1) = \beta\}$).
We create a new basis $q_n(x) = p_n(x) - p_n(-1)\frac{1-x}{2} - p_n(1)\frac{1+x}{2}$ for $1 \le n \le N-1$; $\alpha = \beta = 0$. Another common choice is $r_n(x) = p_n(x) - \gamma_n p_{n-2}(x) - \delta_n p_{n-2}(x)$ w/ δ_n, γ_n carefully chosen. Or $s_n(x) = (1-x^2)p_n(x)$, $n = 0, 1, \dots, N-2$.

Ex: $-u'' + u = f$, $u(-1) = u(1) = 0$
Consider $H^1_0[-1,1] = \{u \in H^1[-1,1] : u(\pm 1) = 0\}$.
$\langle -u'', v\rangle + \langle u, v\rangle = \langle f, v\rangle \Rightarrow \langle u', v'\rangle + \langle u, v\rangle = \langle f, v\rangle$
because $u, v \in H^1_0$. So weak-form is to satisfy above $\forall\, v \in H^1_0$.
To discretize, use $P_{N,0}$ instead of H^1_0. Let $u(x) = \sum_{j=0}^{n} \hat{u}_j q_j(x)$.
Obtain $(\underline{\underline{S}} + \underline{\underline{M}})\underline{u} = \underline{f}$, $(S)_{n,j} = \langle q_n', q_j'\rangle$, $(M)_{n,j} = \langle q_n, q_j\rangle$ are stiffness & mass matrices.

- Analytical methods use 3-term recurrence for $\underline{\underline{S}}$, $\underline{\underline{M}}$
- Or use Legendre-Gauss quadrature for $\underline{\underline{S}}$, $\underline{\underline{M}}$. Integrand

is degree at most $2N$, so use $N+1$ point quadrature. q' obtained thru p'. Using r_n instead of q_n makes $\underline{\underline{S}}$ diagonal.

For Robin BCs, natural treatment is easier

$$\alpha u(\pm 1) + \beta u'(\pm 1) = 0$$

$$\Rightarrow \langle u'', v \rangle = \langle u', v' \rangle - u'(x) v(x)$$

replace $u'(\pm 1)$ with $\frac{\alpha}{\beta} u(\pm 1)$.

Tau method nice w/ time-dependent BCs, $u(1,t) = h(t)$. But not sure about proving convergence.

For collocation, place nodes at ± 1, but then cannot use Gauss-Quadrature, use Gauss-Lobatto, slight loss in accuracy. So enforce zero residual system inside, BCs at edges.