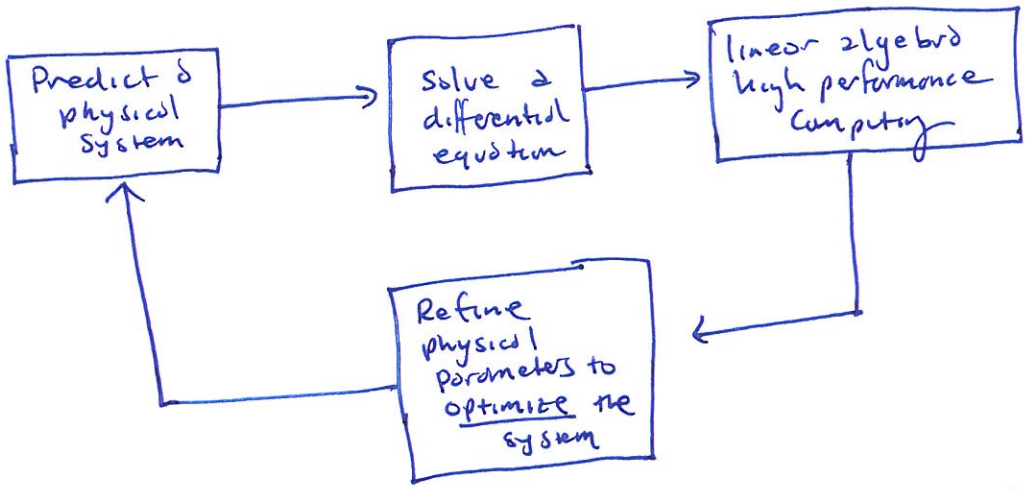
CMDA 4604: Intermediate Topics in Mathematical Modeling

FALL 2015 : Virginia Tech

Lecture 1 : Introduction and Overview

This is a course about Partial Differential Equations. (PDEs).

PDEs form the heart of computational science:



Some PDEs can be solved easily (this class); others are among the most difficult "grand challenge" problems in science and engineering. Progress on the solution of PDEs unlocks technology in many fields of application.

Historical overview

1600s ordinary differential equations of mechanics

$$\frac{dx}{dt}(t) = \alpha x(t), \quad \frac{d^2x(t)}{dt^2} = -\alpha^2 x(t) \quad \Big\} \text{ linear problems}$$

$$\frac{d^2r(t)}{dt^2} = -M\frac{r(t)}{\|r(t)\|^3} \quad \text{Newton's inverse square law of gravitation} \quad \Big\} \text{ nonlinear problems}$$

1700s Partial differential equations of mechanics

$$\frac{\partial^2 u(x,t)}{\partial t^2} = \frac{\partial^2 u(x,t)}{\partial x^2}$$

Wave equation – Vibrating strings
linear PDE

1800s Partial differential equations of heat, fluids, electricity + magnetism

$$\begin{cases} \dfrac{\partial u(\vec{x},t)}{\partial t} = \varepsilon \Delta u(\vec{x},t) - \underbrace{u \cdot \nabla u}_{\text{nonlinearity}} - \nabla p \\ 0 = \nabla \cdot u \end{cases}$$

Incompressible Navier-Stokes equations

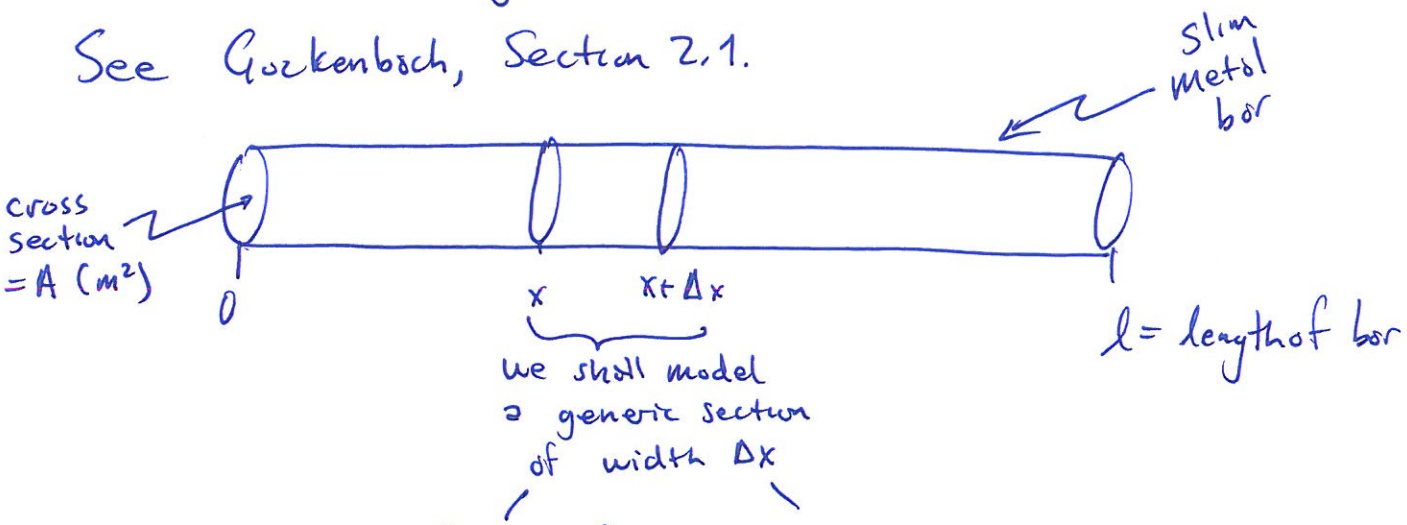
1900s+ Partial differential equations across science & engineering

Quantum mechanics
Porous media flow (groundwater contamination; oil)
neuroscience
Scattering of waves (radar, ultrasound)
traffic flow
financial modeling
black holes
.....

Lecture 2 Derivation of the heat equation.

We shall focus this semester on solving PDEs. However, we start by showing how they arise from physical processes, and how modeling assumptions work their way into the process.

See Gockenbach, Section 2.1.



$u(x,t)$ = temperature of bar at a point $x \in [0, \ell]$ at time $t \geq 0$.

Assume the bar is insulated along its length, so no energy escapes from the side of the bar. (We will consider different possibilities for the end of the bar later.)

A = cross-sectional area (m^2)

ℓ = length of the bar (m)

T_0 = some baseline/ambient temperature (K = Kelvin)

ρ = density of the metal (g/m^3)

Three Key "Lemmas"

Lemma 1 Let $f(x,t)$ be a function that maps $x \in [a,b]$ and $t \in [c,d]$ to a real number,

$$f: [a,b] \times [c,d] \longrightarrow \mathbb{R},$$

and suppose f is continuous on $[a,b]\times[c,d]$ and $\frac{\partial f}{\partial t}$ is also continuous on $[a,b]\times[c,d]$. Then

$$\frac{d}{dt}\int_a^b f(x,t)\,dx = \int_a^b \frac{\partial f}{\partial t}(x,t)\,dx$$

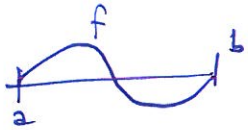
(Continuity of f and $\frac{\partial f}{\partial t}$ is essential.)

Lemma 2 (Fundamental Theorem of Calculus)

$$\int_a^b \frac{\partial f}{\partial x}(s)\,ds = f(b) - f(a)$$

Lemma 3 If $f: \mathbb{R} \to \mathbb{R}$ is continuous and $\int_a^b f(x)\,dx = 0$ for all choices of $a, b \in \mathbb{R}$, then $f(x) = 0$ for all x.

(You can find plenty of functions $f \neq 0$ for which $\int_a^b f(x)\,dx = 0$ for *some* $a, b \in \mathbb{R}$, but not *all* $a, b \in \mathbb{R}$.)



We seek to model the energy in $[x, x+\Delta x]$

E_0 = energy (in Joules = N·m)
in the segment $[x, x+\Delta x]$
when the bar has constant temperature T_0.

$u(x,t)$ = temperature of the bar at point x, time t,
(This will deviate slightly from T_0.)

Physical Properties of the Metal Bar

ρ = density (g/m^3)

c = heat capacity (specific heat)
= energy needed to raise 1 g of the metal
1 degree Kelvin from temperature T_0.
(J/gK)

Wikipedia: Aluminum $c = 0.897$ J/gk at $T_0 = 298$ K
$= 25°$ Centigrade
Iron $c = 0.450$ J/gk at $T_0 = 298$ K

Total energy in $[x, x+\Delta x]$

$$\underbrace{E_0}_{\text{energy in the bar if temp} = T_0} + \int_x^{x+\Delta x} \underbrace{(u(s,t) - T_0)}_{\text{deviation of the true temperature from the nominal value } T_0} (c)(\rho A)\, ds$$

Units

E_0	$u(s,t) - T_0$	c	ρ	A	ds	
J	K	$\frac{J}{gK}$	$\frac{g}{m^3}$	m^2	m	} = Joules

Regroup:

$$\underbrace{E_0 - \int_x^{x+\Delta x} T_0 c\rho A\, ds}_{\tilde{E}_0} + \int_x^{x+\Delta x} u(s,t) c\rho A\, ds$$

How does temperature evolve in time?

$$\frac{d}{dt}\left(\tilde{E}_0 + \int_x^{x+\Delta x} u(s,t) c\rho A\, ds\right) \qquad \left(\text{in } \frac{J}{sec}\right)$$

$$= \frac{d}{dt}\int_x^{x+\Delta x} u(s,t) c\rho A\, ds \qquad \left(\frac{d}{dt}\tilde{E}_0 = 0\right)$$

$$= \int_x^{x+\Delta x} \left(\frac{\partial}{\partial t} u(s,t)\right) c\rho A\, ds \qquad (\text{Lemma 1}) \quad (*)$$

(Do you expect the continuity assumption to hold here?)

<u>KEY IDEA</u> Find an entirely different way to describe the change in energy in the bar from x to $x+\Delta x$.

The <u>flux</u> of energy through the bar is described by the function $q(x,t)$ (in $J/m^2 sec$)

Thus (heat in at x) − (heat out at $x+\Delta x$)

$$= \left(q(x,t) - q(x+\Delta x, t)\right) A \qquad \left(\text{in } \frac{J}{sec}\right)$$

Lemma 2 $\Rightarrow$

$$= -\left(\int_x^{x+\Delta x} \frac{\partial}{\partial x} q(s,t)\, ds\right) A \qquad (**)$$

Equate (*) and (**) to obtain

$$\int_x^{x+\Delta x} \frac{\partial}{\partial t} u(s,t)\, c\rho A\, ds = -\left(\int_x^{x+\Delta x} \frac{\partial}{\partial x} q(s,t)\, ds\right) A$$

Since this holds for all choices of x, $x+\Delta x$, Lemma 3 implies

$$\frac{\partial u}{\partial t} u(x,t)\, c\rho A = -\frac{\partial}{\partial x} q(x,t)\, A$$

Cancel A to obtain (***)

$$\boxed{\frac{\partial u}{\partial t} u(x,t)\, c\rho = -\frac{\partial}{\partial x} q(x,t)} \quad \text{for all } x \in [0,L].$$

Now we need to invoke a physical law.

Fourier's Law of Heat Conduction gives

$$\boxed{q(x,t) = -K \frac{\partial u}{\partial x}(x,t)}$$

where K = thermal conductivity, $\frac{J}{m \cdot K \cdot sec}$

Wikipedia: Aluminum: $K = 237$ $J/m \cdot K \cdot sec$
Iron: $K = 80$ $J/m \cdot K \cdot sec$

Insert Fourier's Law into (***) to get

$$\frac{\partial u}{\partial t} u(x,t)\, c\rho = +\frac{\partial}{\partial x}\left(K(x) \frac{\partial u}{\partial x}(x,t)\right)$$

When c, ρ, k are constant (a "homogeneous bar") we have

$$\frac{\partial u}{\partial t}(x,t) = \left(\frac{k}{c\rho}\right) \frac{\partial^2 u}{\partial x^2}(x,t)$$

which we abbreviate

$$u_t = \frac{k}{c\rho} u_{xx}.$$

In many cases, we shall "choose units" such that

$$\boxed{u_t = u_{xx}}$$

This is typically called "the heat equation".

Lecture 3: Initial conditions, boundary conditions.

Some additional considerations to fully specify the partial differential equation:

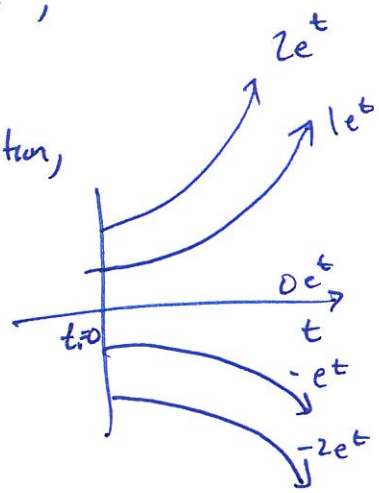
(a) Initial conditions

Recall the simple ordinary differential equation

$$w'(t) = \lambda w(t).$$

The exact solution is $w(t) = Ce^{\lambda t}$, where C is a constant that is specified by the initial condition,

$$C = w(0).$$



Similarly, we need to know the temperature distribution of the bar at time $t=0$ for all $x \in [0, \ell]$.

INITIAL CONDITION: $u(x,0) = u_0(x)$ $\quad x \in [0,\ell]$

(b) Boundary Conditions

What happens at the end of the bar?

This will significantly affect how the temperature changes.

- For example, fix the temperature at <u>both ends</u> of the bar: (e.g., via on ice bath):

$$u(0,t) = \alpha \qquad u(\ell,t) = \beta \quad \text{for all } t \geq 0.$$

These are "<u>Dirichlet boundary conditions</u>"

- Alternatively, we can insulate the bar at both ends, so no heat can escape:

$$\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(l,t) = 0 \qquad \text{for all } t \geq 0.$$

These are "Neumann boundary conditions." (homogeneous ⇒ zero)

Here is a neat side-effect of Neumann boundary conditions: If the bar is insulated on all its sides, the bar should not lose any heat energy.

What is the change in energy?

$$\frac{d}{dt}\int_0^l c\rho A\, u(s,t)\, ds$$

$$= \int_0^l A k \frac{\partial^2}{\partial x^2} u(s,t)\, ds \qquad \left(\begin{array}{c}\text{using the}\\ \text{heat equation}\end{array}\right)$$

$$= A k \left[u_x(l,t) - u_x(0,t)\right] \qquad \text{Lemma 2}$$

$$= 0$$

So, indeed energy is conserved!

(c) We could add energy to the bar from an external source.

Recall units:

$$c\rho u_t = \left(\frac{J}{g\,k}\right)\left(\frac{g}{m^3}\right)\left(\frac{k}{sec}\right) = \frac{J}{m^3\, sec}$$

This gives the full heat equation (constant c, ρ, k):

$$\boxed{c\rho u_t = k u_{xx} + f(x,t)}$$

$f(x,t)$: external heat source, $\frac{J}{m^3 \cdot sec}$

Steady-state behavior

Think about heat flowing in a bar, even with some fixed external source $f(x)$ (not time dependent). You might expect the heat to eventually distribute throughout the bar in some time-independent / steady state distribution, in which case $u_t(x,t) = 0$.

The heat equation simplifies then to

$$0 = u_{xx} + f$$

$$\Rightarrow \boxed{-u_{xx} = f}$$ This is called "Laplace's equation"

If the material properties vary in space, this instead takes the more general form

$$\boxed{-\frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}(x)\right) = f(x) \qquad x \in [0,\ell]}$$

We shall spend the first half of the semester studying these two equations, then introduce time dependence.

We shall draw an analogy between the linear algebra problem $Ax = b$ (solvable by Gaussian elimination) and the Differential equation $\frac{-d^2}{dx^2}u = f$.

A: generalizes; x: unknown; b: generalizes

We seek some general methods to solve this broad class of equations.

Lecture 4: Linear Operators and Inner Products

Section 3 Mathematical Landscape

We seek to generalize the tools of linear algebra (vectors on $\mathbb{R}^N$) to work on functions, say on $x \in [0,1]$.

First, we generalize $\mathbb{R}^N$ to more abstract objects.

Def A collection of objects $\mathcal{V}$ ("vectors") with which we associate vector addition ($v+w$ for $v, w \in \mathcal{V}$) and scalar multiplication (αv for $v \in \mathcal{V}$, $\alpha \in \mathbb{R}$) is a vector space provided

(i) If $v, w \in \mathcal{V}$, then $v+w \in \mathcal{V}$ ("closed under vector addition")

(ii) If $v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, then $\alpha v \in \mathcal{V}$. ("closed under scalar multiplication")

Note that vector addition and scalar multiplication must obey a set of axioms — we shall not go into details here, though this idea is emphasized in linear algebra courses...

Examples

1) $\mathcal{V} = \mathbb{R}^N$ holds, with standard vector addition and scalar multiplication.

2) $\mathcal{V} = C[a,b]$ = continuous functions of a real variable $x \in [a,b]$.

"Vector addition" $\Rightarrow$ adding functions $f(x)+g(x)$

"scalar multiplication $\Rightarrow$ scaling functions $\alpha f(x)$

3) $V = C^1[a,b] = \{$ continuous functions on $x \in [a,b]$ with a continuous first derivative $\}$

Some vector addition and scalar multiplication as for $V = C[a,b]$.

4) $V = C_D^2[a,b] = \{$ continuous functions for $x \in [a,b]$ with a continuous first and second derivative, with $f(a) = f(b) = 0$ $\}$

$D \Rightarrow$ "Dirichlet" — Dirichlet b.c.'s

5) $V = \{ f \in C[a,b]$ with $f(x) > 0$ for all $x \in [a,b] \}$

NOT A VECTOR SPACE!

If $f \in V$, e.g. $f(x) = 1$, then $-f = -1 \notin V$.

V is not closed under ~~vector~~ scalar multiplication.

Linear operators If linear spaces contain the generalizations of vectors in $\mathbb{R}^N$, what is the generalization of matrices in $\mathbb{R}^{N \times N}$ that act upon those vectors?

Def A map $L: V \to W$ is a **linear operator** from the vector space V to the vector space W provided

(i) $L(u+v) = Lu + Lv$ for all $u, v \in V$

(ii) $L(\alpha u) = \alpha(Lu)$ for all $u \in V$, $\alpha \in \mathbb{R}$

Examples

1) $V = C[a,b]$, $W = C[a,b]$.

$Lu = e^x u$

So $L(u+v) = e^x(u+v) = e^x u + e^x v = Lu + Lv$ ✓

$L(\alpha u) = e^x(\alpha u) = \alpha(e^x u) = \alpha Lu$ ✓

Thus L is a linear operator.

We call this L a "multiplication operator."

2) $V = C^1[a,b]$, $W = C[a,b]$

$Lu = u'$

Then for all $u, v \in V$: $L(u+v) = (u+v)' = u' + v' = Lu + Lv$

$\alpha \in \mathbb{R}$: $L(\alpha u) = (\alpha u)' = \alpha u' = \alpha Lu$.

Thus L is a linear operator. (A "differential operator")

3) $V = C_D^2[a,b]$, $W = C[a,b]$ (from the steady state heat equation...)

$Lu = -(k(x)u'(x))'$ $\quad k(x) > 0$

For all $u, v \in V$,

$L(u+v) = -(k(x)(u'(x)+v'(x))'$

$= -(k(x)u'(x))' - (k(x)v'(x))'$

For all $\alpha \in \mathbb{R}$,

$L(\alpha u) = -(k(x)((\alpha u)'))'$

$= -\alpha(k(x)u'(x))' = \alpha Lu$.

So L is linear. (Also a "differential operator")

Inner products Now we seek to generalize the dot product for vectors $x, y \in \mathbb{R}^N$: $x \cdot y = \sum x_j y_j = x^T y$.

For motivation, suppose we have a grid of uniformly spaced points, $x_1 = h$, $x_2 = 2h$, ..., $x_N = Nh$, for $h = \frac{1}{N}$. We want an "inner product" / dot product for functions in $C[0,1]$. Represent $f \in C[0,1]$ via the approximation

$$f = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{bmatrix} \in \mathbb{R}^N$$

Similarly, represent $g \in C[a,b]$ via the approximation

$$g = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_N) \end{bmatrix} \in \mathbb{R}^N.$$

Then $f \cdot g = \sum_{j=1}^{N} f(x_j) g(x_j)$.

Notice that this will not typically converge as $N \to \infty$.

If instead we consider

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^{N} f(x_j) g(x_j) &= h \sum_{j=1}^{N} f(x_j) g(x_j) \\ &= \sum_{j=1}^{N} \left(h f(x_j) g(x_j) \right) \\ &\approx \sum_{j=1}^{N} \int_{x_{j-1}}^{x_j} f(x) g(x) \, dx \\ &= \int_{x_0}^{x_N} f(x) g(x) \, dx. \\ &= \int_0^1 f(x) g(x) \, dx. \end{aligned}$$

This suggests that we <u>define</u> the inner product on $C[a,b]$ as

$$(f, g) = \int_0^1 f(x) g(x) \, dx$$

We will now collect properties that this (and any other) inner product must satisfy.

Def A function $(\cdot,\cdot): V \times V \to \mathbb{R}$ is an *inner product* provided

i) $(u+v, w) = (u,w) + (v,w)$
$(\alpha u, w) = \alpha (u,w)$ } linearity

ii) $(u,w) = (w,u)$ } symmetry

iii) $(u,u) \geq 0$, $(u,u)=0$ if and only if $u=0$ } Positivity.

(Check that
$(\cdot,\cdot): C[a,b] \times C[a,b] \to \mathbb{R}$
$(f,g) = \int_0^1 f(x) g(x)\, dx$
is an inner product.)

An inner product is linear in both the first component

$(\alpha u + v, w) = \alpha (u,w) + (v,w)$

and, using symmetry, the second component:

$(u, \alpha v + w) = \alpha (u,v) + (u,w)$.

Thus we say that the inner product is a "bilinear form".

Lecture 5: Inner products and Norms

LAST TIME WE INTRODUCED ABSTRACT INNER PRODUCTS.

Let $(\cdot,\cdot)$ be an inner product on the vector space V.

Def The <u>norm</u> of a vector $v \in V$ is given by

$$\|v\| = \sqrt{(v,v)}$$

The <u>angle</u> between nonzero vectors $v, w \in V$ is defined via

$$\cos \measuredangle(v,w) = \frac{(v,w)}{\|v\|\,\|w\|}$$

The norm satisfies the following properties:

(i) $\|v\| \geq 0$ for all $v \in V$, and $\|v\| = 0$ if and only if $v = 0$. ("positivity")

(ii) $\|\alpha v\| = |\alpha|\,\|v\|$ for all $v \in V$, $\alpha \in \mathbb{R}$ ("scaling")

(iii) $\|v+w\| \leq \|v\| + \|w\|$ for all $v, w \in V$ ("triangle inequality")

Properties (i) and (ii) follow immediately from the definition and properties of inner products. We shall prove property (iii). Before doing so, we first establish a vital result.

<u>Theorem</u> (Cauchy-Schwarz Inequality)

For any $v, w \in V$,

$$|(v,w)| \leq \|v\|\,\|w\|$$

There are many proofs of Cauchy-Schwarz — see the book "The Cauchy-Schwarz Master Class" or N. Young, "Introduction to Hilbert Space"

Proof Let $v, w \in V$.

CASE 1 If $v = \alpha w$ for some $\alpha \in \mathbb{R}$, then

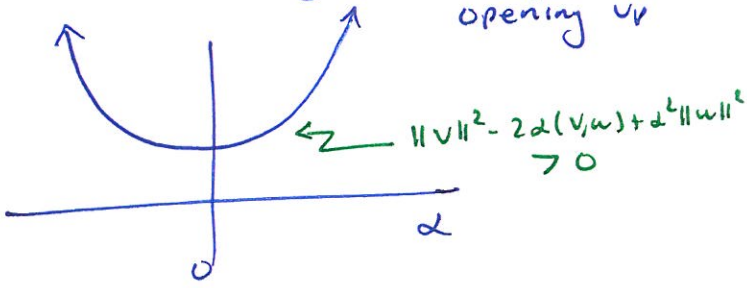
$$|(v,w)| = |(\alpha w, w)| = |\alpha (w,w)|$$
$$= |\alpha|\,|(w,w)| = |\alpha|\,\|w\|^2$$
$$= \|\alpha w\|\,\|w\| = \|v\|\,\|w\|.$$

Thus Cauchy-Schwarz holds with equality.

CASE 2 Suppose $v \neq \alpha w$ for all $\alpha \in \mathbb{R}$. Then $v - \alpha w \neq 0$ for all $\alpha \in \mathbb{R}$.

$$0 < (v - \alpha w, v - \alpha w)$$
$$= (v,v) - \alpha(v, \alpha w) - (\alpha w, v) + (\alpha w, \alpha w)$$
$$= \|v\|^2 - 2\alpha(v,w) + \alpha^2 (w,w)$$
$$= \|v\|^2 - 2\alpha(v,w) + \alpha^2 \|w\|^2$$

} This is a quadratic in α, opening up



Since this quadratic in α positive for all real α, the roots must be Complex.

Thus the discriminant $(b^2 - 4ac)$ in the quadratic formula must be negative:

"$b^2 - 4ac$" $= \underbrace{(-2(v,w))^2}_{b} - 4\,\underbrace{\|w\|^2}_{a}\,\underbrace{\|v\|^2}_{c} < 0$

Recall the quadratic formula

$$ax^2 + bx + c = 0$$
$$\Rightarrow x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Complex roots for $a, b, c \in \mathbb{R}$
$\Rightarrow \sqrt{b^2 - 4ac}$ purely imaginary
$\Rightarrow b^2 - 4ac < 0$

$$\Rightarrow \quad 4(v,w)^2 - 4\|v\|^2\|w\|^2 \leq 0$$

$$\Rightarrow \quad (v,w)^2 \leq \|v\|^2\|w\|^2$$

$$\Rightarrow \quad |(v,w)| \leq \|v\|\,\|w\|, \quad \text{establishing the result.}$$

Now we can prove the triangle inequality.

For all $v, w \in V$, $\quad \|v+w\| \leq \|v\| + \|w\|$

Proof $\quad \|v+w\|^2 = (v+w,\, v+w)$

$$= \|v\|^2 + (v,w) + (w,v) + \|w\|^2$$

$$= \|v\|^2 + 2(v,w) + \|w\|^2$$

$$\leq \|v\|^2 + 2|(v,w)| + \|w\|^2 \qquad \left((v,w) \leq |(v,w)|\right)$$

Cauchy-Schwarz ⤷

$$\leq \|v\|^2 + 2\|v\|\,\|w\| + \|w\|^2$$

$$= \left(\|v\| + \|w\|\right)^2$$

Thus $\quad \|v+w\| \leq \|v\| + \|w\|.$

Note we defined

$$\cos \measuredangle(v,w) = \frac{(v,w)}{\|v\|\,\|w\|}.$$

Cauchy-Schwarz ensures that

$$\frac{|(v,w)|}{\|v\|\,\|w\|} \in [0,1], \quad \text{so} \quad \frac{(v,w)}{\|v\|\,\|w\|} \in [-1,1]$$

and hence $\quad \cos^{-1}\left(\dfrac{(v,w)}{\|v\|\,\|w\|}\right) = \measuredangle(v,w)$

is sensible.

Lecture 6: Intro to Best Approximation

Section 4 Best Approximation.

Let V be a vector space.

Def Two vectors $u, v \in V$ are <u>orthogonal</u> provided $(u,v) = 0$.

Def A set of vectors $\emptyset \neq U \subseteq V$ is a <u>subspace</u> of V provided
- If $u, w \in U$, then $u + w \in U$
- If $u \in U$, $\alpha \in \mathbb{R}$, then $\alpha \in \mathbb{R}$. (empty set)

It is customary to require that $U \neq \emptyset$ i.e., U must contain at least one vector. (The zero vector is always in U: $0 \in U$.)

BEST APPROXIMATION PROBLEM

Let U be a subspace of the vector space V.

Given some $v \in V$, find the vector u_* that best approximates v over all vectors in U:

$$\|v - u_*\| \leq \|v - u\| \quad \text{for all } u \in U.$$

In this lecture, we simply take

$$U = \text{span}\{u\} = \{c\phi : c \in \mathbb{R}\} \qquad (\phi = \text{phi})$$

for some nonzero vector $\phi \in V$. Thus, U is a one-dimensional subspace. (In the next lecture we will tackle n-dimensional subspaces.)

We will use calculus to find the best approximation:

Given $v \in V$, consider any element $c\phi \in \mathcal{U}$.

We measure the mismatch between v and $c\phi$ as:

$$E(c) = \|v - c\phi\|^2$$

$$= (v - c\phi,\ v - c\phi)$$

$$= (v,v) - (v, c\phi) - (c\phi, v) + (c\phi, c\phi)$$

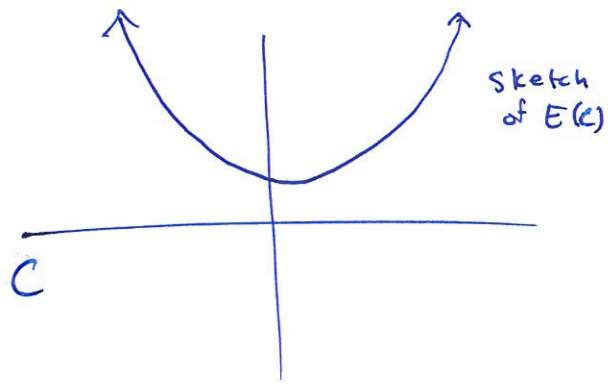
$$= \|v\|^2 - 2(v, c\phi) + c^2\|\phi\|^2$$

$$= \|v\|^2 - 2c(v,\phi) + c^2\|\phi\|^2$$

} using properties of inner products

$\Rightarrow$ We seek to minimize this error function: $\Leftarrow$

First note that $E(c)$ is a quadratic function in c, opening up, and $E(c) = \|v - c\phi\|^2 \geq 0$.



Sketch of $E(c)$

What c minimizes $E(c)$?

$$\frac{d}{dc} E(c) = -2(v,\phi) + 2c\|\phi\|^2$$

Set $\frac{d}{dc} E(c) = 0 \Rightarrow 0 = -2(v,\phi) + 2c\|\phi\|^2$

$$\Rightarrow \boxed{c = \frac{(v,\phi)}{\|\phi\|^2}}$$

No division by zero because $\phi \neq 0$, by assumption.

The best approximation to $v \in V$ from $\mathcal{U} = \text{span}\{\phi\}$ is thus

$$\boxed{u_* = \frac{(v,\phi)}{\|\phi\|^2}\phi = \frac{(v,\phi)}{(\phi,\phi)}\phi}$$

It is helpful to see this best approximation in terms of projection.

Def A linear operator $P: V \to V$ is a projector provided $P^2 = P$, which means that for all $v \in V$,

$$P(Pv) = Pv.$$

We say P is "idempotent".

Define $Pv = \dfrac{(v,\phi)}{(\phi,\phi)}\phi$ = best approx to v from $\operatorname{span}\{\phi\}$.

Theorem P is a projector.

Proof Let $v \in V$. Then

$$Pv = \frac{(v,\phi)}{(\phi,\phi)}\phi \quad \text{and} \quad P(Pv) = \frac{(Pv,\phi)}{(\phi,\phi)}\phi.$$

Hence
$$P(Pv) = \frac{\left(\frac{(v,\phi)}{(\phi,\phi)}\phi, \phi\right)}{(\phi,\phi)}\phi$$

$$= \frac{\frac{(v,\phi)}{(\phi,\phi)}(\phi,\phi)}{(\phi,\phi)}\phi \qquad \text{Using properties of inner products}$$

$$= \frac{(v,\phi)}{(\phi,\phi)}\phi.$$

$$= Pv.$$

Hence P is a projector.

One more (vital) property:

Theorem The error $v - u_*$ between v and its best approximation u_* from $\mathcal{U} = \text{span}\{\phi\}$ is orthogonal to the approximating subspace $\mathcal{U}$.

Proof $u_* = \dfrac{(v,\phi)}{(\phi,\phi)}\phi$. Check the orthogonality:

Given any $u \in \text{span}\{\phi\}$, there exists $\gamma \in \mathbb{R}$ such that $u = \gamma\phi$.

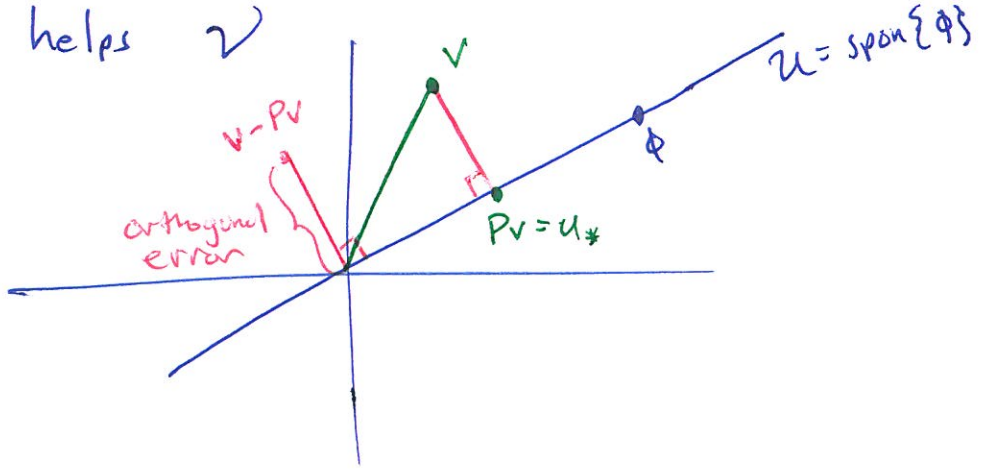
$$(v - u_*, u) = \left(v - \frac{(v,\phi)}{(\phi,\phi)}\phi, \gamma\phi\right)$$

$$= \gamma\left(v - \frac{(v,\phi)}{(\phi,\phi)}\phi, \phi\right)$$

$$= \gamma\left[(v,\phi) - \frac{(v,\phi)}{(\phi,\phi)}(\phi,\phi)\right] = \gamma(0) = 0.$$

Hence $v - u_*$ is orthogonal to every vector in $\mathcal{U}$. (We say $v - u_* \perp \mathcal{U}$.) ∎

A picture helps ✓



Lecture 7: Best approximation from general subspaces

In this lecture we generalize the result from lecture 6 to obtain best approximations from a general N-dimensional subspace $\mathcal{U} = \text{Span}\{\phi_1, \phi_2, \ldots, \phi_N\}$, where the <u>basis vectors</u> $\phi_1, \phi_2, \ldots, \phi_N$ are assumed to be linearly independent.

<u>Def</u> The span of vectors $\phi_1, \ldots, \phi_N \in \mathcal{V}$ is the set of all linear combinations (weighted sums) of $\phi_1, \ldots, \phi_N$:

$$\text{Span}\{\phi_1, \ldots, \phi_N\} = \left\{ c_1\phi_1 + \cdots + c_N\phi_N : \text{for any } c_1, \ldots, c_N \in \mathbb{R} \right\}.$$

<u>Theorem</u> $\text{Span}\{\phi_1, \ldots, \phi_N\}$ is a subspace of $\mathcal{V}$.

First we handle the case of $N=2$. This will make the general pattern clearer.

Given $v \in \mathcal{V}$, we seek $u_* \in \text{span}\{\phi_1, \phi_2\}$ to minimize $\|v - u_*\|$. Any $u \in \text{span}\{\phi_1, \phi_2\} = \mathcal{U}$ can be written as $u = c_1\phi_1 + c_2\phi_2$. Define the error function

$$\begin{aligned} E(c_1, c_2) &= \|v - u\|^2 \\ &= \|v - (c_1\phi_1 + c_2\phi_2)\|^2 \\ &= (v - c_1\phi_1 - c_2\phi_2,\ v - c_1\phi_1 - c_2\phi_2) \end{aligned}$$

Expanding the inner product:

$$E(c_1,c_2) = \|v\|^2 - 2c_1(v,\phi_1) - 2c_2(v,\phi_2) + c_1^2(\phi_1,\phi_1) + 2c_1c_2(\phi_1,\phi_2) + c_2^2(\phi_2,\phi_2)$$

This function is now a <u>paraboloid</u> in c_1, c_2 opening up.

To find its minimum, take partial derivatives:

$$\frac{\partial E}{\partial c_1} = 0 - 2(v,\phi_1) - 0 + 2c_1(\phi_1,\phi_1) + 2c_2(\phi_1,\phi_2) + 0$$

$$\frac{\partial E}{\partial c_2} = 0 - 0 - 2(v,\phi_2) + 0 + 2c_1(\phi_1,\phi_2) + 2c_2(\phi_2,\phi_2)$$

and set them simultaneously to zero:

$$0 = 2c_1(\phi_1,\phi_1) + 2c_2(\phi_1,\phi_2) - 2(v,\phi_1)$$
$$0 = 2c_1(\phi_1,\phi_2) + 2c_2(\phi_2,\phi_2) - 2(v,\phi_2)$$

Re-arrange and cancel the 2:

$$c_1(\phi_1,\phi_1) + c_2(\phi_1,\phi_2) = (v,\phi_1)$$
$$c_1(\phi_1,\phi_2) + c_2(\phi_2,\phi_2) = (v,\phi_2)$$

Notice: this is a system of 2 linear equations in the 2 unknowns c_1, c_2.

Set up as a linear system.

$$\underbrace{\begin{bmatrix} (\phi_1,\phi_1) & (\phi_1,\phi_2) \\ (\phi_2,\phi_1) & (\phi_2,\phi_2) \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} c_1 \\ c_2 \end{bmatrix}}_{c} = \underbrace{\begin{bmatrix} (V,\phi_1) \\ (V,\phi_2) \end{bmatrix}}_{b}$$

Solve this (via Gaussian elimination) for the unknowns c_1 and c_2.

Then the best approximation is

$$U_* = c_1\phi_1 + c_2\phi_2.$$

CASE OF GENERAL N: $\mathcal{U} = \text{span}\{\phi_1, \dots, \phi_N\}$.

Now $E(c_1, \dots, c_N) = \|V - (c_1\phi_1 + \cdots + c_N\phi_N)\|^2$

we can expand

$$E(c_1, \dots, c_N) = \left(V - (c_1\phi_1 + \cdots + c_N\phi_N),\ V - (c_1\phi_1 + \cdots + c_N\phi_N)\right)$$

$$= (V,V) - 2\sum_{j=1}^{N} c_j (V,\phi_j) + \sum_{j=1}^{N}\sum_{k=1}^{N} c_j c_k (\phi_j,\phi_k)$$

The partial derivatives are a bit trickier to compute: We shall isolate $\frac{\partial E}{\partial c_1}$: Then the other cases will be evident.

7.4

$$\frac{\partial E}{\partial c_1}(c_1,\dots,c_N) = \frac{\partial}{\partial c_1}\left(\|v\|^2 - 2\sum_{j=1}^{N} c_j(v,\phi_j) + \sum_{j=1}^{N}\sum_{k=1}^{N} c_j c_k(\phi_j,\phi_k)\right)$$

$$= 0 - \frac{\partial}{\partial c_1}\left(2\sum_{j=1}^{N} c_j(v,\phi_j)\right)$$

Break the double sum into four parts

$$\begin{cases} + \frac{\partial}{\partial c_1}\left(\sum_{\substack{k=2 \\ (j=1)}}^{N} c_1 c_k(\phi_1,\phi_k)\right) & j=1,\ k\neq 1 \\ + \frac{\partial}{\partial c_1}\left(\sum_{\substack{j=2 \\ (k=1)}}^{N} c_j c_1(\phi_j,\phi_1)\right) & j\neq 1,\ k=1 \\ + \frac{\partial}{\partial c_1}\left(\sum_{j=2}^{N}\sum_{k=2}^{N} c_j c_k(\phi_j,\phi_k)\right) & j\neq 1,\ k\neq 1 \\ + \frac{\partial}{\partial c_1}\left(c_1 c_1(\phi_1,\phi_1)\right) & j=1,\ k=1 \end{cases}$$

$$= -2(v,\phi_1) + \sum_{k=2}^{N} c_k(\phi_1,\phi_k) + \sum_{j=2}^{N} c_j(\phi_j,\phi_1) + 0 + 2c_1(\phi_1,\phi_1)$$

Consolidate into $2\sum_{j=2}^{N} c_j(\phi_j,\phi_1)$

$$= 2c_1(\phi_1,\phi_1) + 2\sum_{j=2}^{N} c_j(\phi_j,\phi_1) - 2(v,\phi_1)$$

$$= 2\sum_{j=1}^{N} c_j(\phi_j,\phi_1) - 2(v,\phi_1)$$

$$\frac{\partial E}{\partial c_1} = 0 \Longrightarrow \boxed{\sum_{j=1}^{N} c_j(\phi_j,\phi_1) = (v,\phi_1)}$$

In the same way,

$$\frac{\partial E}{\partial c_\ell}(c_1,\dots,c_N)=0 \Longrightarrow \sum_{j=1}^{N} c_j(\phi_j,\phi_\ell)=(v,\phi_\ell)$$

for all $\ell=1,\dots,N$.

This again gives N linear equations in the N unknowns $c_1,\dots,c_N$:

$$\underbrace{\begin{bmatrix}(\phi_1,\phi_1) & \cdots & (\phi_1,\phi_N)\\ \vdots & & \\ (\phi_N,\phi_1) & \cdots & (\phi_N,\phi_N)\end{bmatrix}}_{G}\underbrace{\begin{bmatrix}c_1\\ c_2\\ \vdots\\ c_N\end{bmatrix}}_{c}=\underbrace{\begin{bmatrix}(v,\phi_1)\\ (v,\phi_2)\\ \vdots\\ (v,\phi_N)\end{bmatrix}}_{b}$$

G = "Gram matrix".

Symmetry of the inner product $\Longrightarrow$

$$G=G^T \quad (G \text{ is a symmetric matrix})$$

Is G invertible?

We will address this question in the next lecture.

Lecture 8 : Properties of best approximations

In this lecture we will prove that:

- If $\phi_1, \dots, \phi_N$ are linearly independent, then G = Gram matrix is invertible.
- The error $v - u_*$ in the best approximation is always orthogonal to the approximating subspace $\mathcal{U}$
- If the basis vectors $\phi_1, \dots, \phi_N$ are orthogonal, then G is <u>diagonal</u> and the best approximation is easy to compute.

Invertibility of the Gram matrix

If G is not invertible, it must have a nontrivial null space, i.e., there exists $z \neq 0$ such that $Gz = 0$.

Suppose such a z exists. Then $Gz = 0 \Longrightarrow z^T G z = 0$

$$Gz = \begin{bmatrix} z_1(\phi_1,\phi_1) + \cdots + z_N(\phi_1,\phi_N) \\ \vdots \\ z_1(\phi_N,\phi_1) + \cdots + z_N(\phi_N,\phi_N) \end{bmatrix} = \begin{bmatrix} (\phi_1, \sum_{j=1}^N z_j \phi_j) \\ \vdots \\ (\phi_N, \sum_{j=1}^N z_j \phi_j) \end{bmatrix}$$

$$z^T G z = [z_1 \cdots z_N] \begin{bmatrix} (\phi_1, \sum_{j=1}^N z_j \phi_j) \\ \vdots \\ (\phi_N, \sum_{j=1}^N z_j \phi_j) \end{bmatrix} = \sum_{k=1}^N z_k \left(\phi_k, \sum_{j=1}^N z_j \phi_j\right)$$

$$= \left(\sum_{k=1}^N z_k \phi_k, \sum_{j=1}^N z_j \phi_j\right)$$

Thus

$$z^T G z = \left(\sum_{k=1}^{N} z_k \phi_k, \sum_{j=1}^{N} z_j \phi_j \right) = \left\| \sum_{j=1}^{N} z_j \phi_j \right\|$$

Some vector - just different counting indices

So if $z^T G z = 0$, then $\sum_{j=1}^{N} z_j \phi_j = 0$.

If $z \neq 0$, then we have a nontrivial linear combination of $\phi_1, \dots, \phi_N$ that equals 0, thus contradicting the linear independence of $\phi_1, \dots, \phi_N$.

So, $\{\phi_1, \dots, \phi_N\}$ linearly independent $\Rightarrow$ G is invertible.

(In fact, since $z^T G z > 0$ for all $z \neq 0$, by positivity of the norm of nonzero vectors, we say that G is a positive definite matrix.)

Best approximation and orthogonality.

Let $u_* = c_1 \phi_1 + \cdots + c_N \phi_N$ be the best approximation to $v \in V$ from $\mathcal{U} = \text{span}\{\phi_1, \dots, \phi_N\}$, with $c_1, \dots, c_N$ arrived at by solving $Gc = b$.

First test

$$(v - u_*, \phi_k) = (v, \phi_k) - (u_*, \phi_k) \qquad k \in \{1, \dots, N\}$$

$$= (v, \phi_k) - \left(\sum_{j=1}^{N} c_j \phi_j, \phi_k\right)$$

$$= (v, \phi_k) - \sum_{j=1}^{N} c_j (\phi_j, \phi_k) \quad \Big\} \; c_1, \dots, c_N \text{ must satisfy } \sum_{j=1}^{N} c_j (\phi_j, \phi_k) = (v, \phi_k) \text{ from the } k^{\text{th}} \text{ row of } Gc = b$$

$$= (v, \phi_k) - (v, \phi_k)$$

$$= 0$$

Now, for arbitrary $u \in \mathcal{U}$, write

$$u = \sum_{k=1}^{N} d_k \phi_k.$$

Then

$$(v - u_*, u) = \left(v - u_*, \sum_{k=1}^{N} d_k \phi_k\right)$$

$$= \sum_{k=1}^{N} (v - u_*, \phi_k)$$

$(v - u_*, \phi_k) = 0$ by the previous calculation

$$= 0.$$

Hence: $u_* =$ best approximation $\Rightarrow v - u_* \perp \mathcal{U}$.

What about the opposite implication?

Suppose $\hat{u} = \sum_{j=1}^{N} \gamma_j \phi_j$ is some vector in $\mathcal{U}$ such that $v - \hat{u} \perp \mathcal{U}$. Is $\hat{u}$ a best approximation?

Note that $v - \hat{u} \perp \mathcal{U}$ implies that $(v - \hat{u}, \phi_k) = 0$ for all $k = 1, \ldots, N$, since $\phi_k \in \mathcal{U}$.

Thus

$$0 = (v - \hat{u}, \phi_k)$$
$$= (v - \sum_{j=1}^{N} \gamma_j \phi_j, \phi_k)$$
$$= (v, \phi_k) - \sum_{j=1}^{N} \gamma_j (\phi_j, \phi_k) \qquad k = 1, \ldots, N$$

But then we have N equations for the N variables $\gamma_1, \ldots, \gamma_N$ to satisfy:

$$\begin{bmatrix} (\phi_1, \phi_1) & \cdots & (\phi_1, \phi_N) \\ \vdots & \ddots & \vdots \\ (\phi_N, \phi_1) & \cdots & (\phi_N, \phi_N) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_N \end{bmatrix} = \begin{bmatrix} (v, \phi_1) \\ (v, \phi_2) \\ \vdots \\ (v, \phi_N) \end{bmatrix}$$

So $\gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_N \end{bmatrix}$ must satisfy $G\gamma = b$.

But if $\phi_1, \ldots, \phi_N$ is linearly independent, G is invertible, and we have $\gamma = G^{-1}b$, the same solution $c = G^{-1}b$ for the best approximation. Thus, <u>if $v - \hat{u} \perp \mathcal{U}$, $\hat{u}$ must be the best approximation!</u>

We summarize these results in a theorem.

Theorem Let $\phi_1, \dots, \phi_N \in V$ be linearly independent. Then $u_* \in \mathcal{U} = \text{span}\{\phi_1, \dots, \phi_N\}$ is the best approximation to $v \in V$ if and only if

(1) $v - u_*$ is orthogonal to all $u \in \mathcal{U}$, (We write $v - u_* \perp \mathcal{U}$.)

Example To find the best approximation to $v(x) = e^x$ from $\text{span}\{1, x\}$ over $C[0,1]$, we compute: $\phi_1(x) = 1$, $\phi_2(x) = x$

$$(\phi_1, \phi_1) = \int_0^1 1 \cdot 1 \, dx = 1$$

$$(\phi_2, \phi_1) = (\phi_1, \phi_2) = \int_0^1 1 \cdot x \, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

$$(\phi_2, \phi_2) = \int_0^1 x \cdot x \, dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}$$

$$G = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix} \qquad b = \begin{bmatrix} e-1 \\ 1 \end{bmatrix}$$

$$(v, \phi_1) = \int_0^1 e^x 1 \, dx = \left[e^x\right]_0^1 = e - 1$$

$$(v, \phi_2) = \int_0^1 x e^x \, dx = \left[e^x x\right]_0^1 - \int_0^1 e^x dx = 1$$

Solve $Gc = b \Rightarrow$

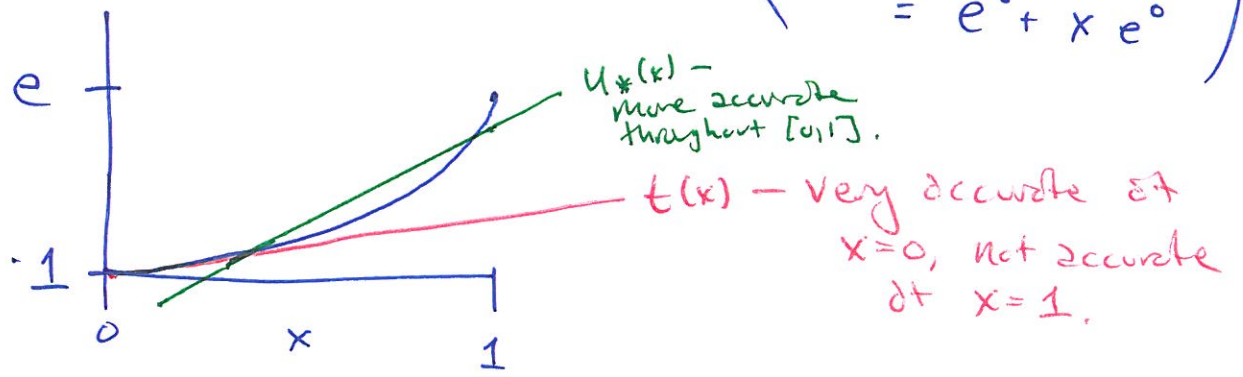
$$c = \begin{bmatrix} 4e - 10 \\ 18 - 6e \end{bmatrix} \approx \begin{bmatrix} 0.8731 \\ 1.6903 \end{bmatrix}$$

Compare this best approximation over $[0,1]$

$$u_*(x) \approx 0.8731 + 1.6903\,x$$

to the first two terms in the Taylor series for $v(x) = e^x$ expanded about $x=0$:

$$t(x) = 1 + x \qquad \left(t(x) = v(0) + x\,v'(0) = e^0 + x\,e^0\right)$$



- **Orthogonal bases.**

There is a very nice special case of best approximation: What if $\phi_1, \dots, \phi_N$ are mutually orthogonal: $(\phi_j, \phi_k) = \begin{cases} 0 & j \neq k \\ \neq 0 & j = k \end{cases}$?

In this case, $G = \begin{bmatrix} (\phi_1,\phi_1) & & 0 \\ & (\phi_2,\phi_2) & \\ 0 & & \ddots \\ & & & (\phi_N,\phi_N) \end{bmatrix}$, so

$$G^{-1} = \begin{bmatrix} \frac{1}{(\phi_1,\phi_1)} & & 0 \\ & \frac{1}{(\phi_2,\phi_2)} & \\ 0 & & \ddots \\ & & & \frac{1}{(\phi_N,\phi_N)} \end{bmatrix}$$

Inverses of diagonal matrices are easy to compute!

$$\Rightarrow c = G^{-1}b = \begin{bmatrix} \frac{1}{(\phi_1,\phi_1)} & & \\ & \ddots & \\ & & \frac{1}{(\phi_N,\phi_N)} \end{bmatrix} \begin{bmatrix} (v,\phi_1) \\ \vdots \\ (v,\phi_N) \end{bmatrix} = \begin{bmatrix} \frac{(v,\phi_1)}{(\phi_1,\phi_1)} \\ \vdots \\ \frac{(v,\phi_N)}{(\phi_N,\phi_N)} \end{bmatrix}$$

Thus If $\phi_1, \dots, \phi_N$ are orthogonal, the best approximation of $v \in V$ from $\mathcal{U} = \text{Span}\{\phi_1, \dots, \phi_N\}$ is given by

$$\boxed{u_* = \sum_{j=1}^{N} \frac{(v, \phi_j)}{(\phi_j, \phi_j)} \phi_j} \qquad (*)$$

** Beware: this only holds if $\phi_1, \dots, \phi_N$ are orthogonal. It is a common mistake to use this formula when $\phi_1, \dots, \phi_N$ are not orthogonal. Bogus! **

Note: the formula (*) is very nice: it says that, if $\phi_1, \dots, \phi_N$ are orthogonal, then u_* is the sum of the individual best approximations onto $\phi_1, \phi_2, \dots, \phi_N$ independently (See lecture 6).

$$u_* = \sum_{j=1}^{N} P_j v \quad \text{where} \quad P_j v = \frac{(v, \phi_j)}{(\phi_j, \phi_j)} \phi_j$$

is the projector onto ϕ_j.

We can write $u_* = \underbrace{\left(\sum_{j=1}^{N} P_j \right)}_{= P = \text{projector onto } \text{Span}\{\phi_1, \dots, \phi_N\}.} v$

Lecture 9: Symmetric Linear Operators

Section 5 Eigenvalues and Eigenfunctions

Recall that a matrix $A \in \mathbb{R}^{N\times N}$ has an eigenvalue λ corresponding to the eigenvector $V \neq 0$ if $\boxed{AV = \lambda V}$.

In this section, we generalize this notion to linear operators. (This will lead to a technique for solving differential equations in Section 6.)

We shall focus on a generalization of symmetric matrices.

Def A linear operator $L: \mathcal{U} \to \mathcal{V}$ is *symmetric* provided $(Lu, v) = (u, Lv)$ for all $u, v \in \mathcal{U}$.

First we shall see that this is consistent with the usual definition that $A \in \mathbb{R}^{N\times N}$ is symmetric if $A = A^T$.

In this case, $(x, y) = y^T x$.

If $A = A^T$, then $a_{j,k} = a_{k,j}$ for all $j, k \in \{1, \dots, N\}$.

We will show that $(Au, v) = (u, Av)$ for all $u, v \in \mathbb{R}^N$ implies that $a_{j,k} = a_{k,j}$ for all $j, k \in \{1, \dots, N\}$

Take $u = e_k$ = k^{th} column of the identity

$v = e_j$ = j^{th} column of the identity

$(Au, v) = e_j^T A e_k = a_{j,k}$

$(u, Av) = (Av)^T u = \left(1 \times (Av)^T u\right)^T = u^T A v$

$= e_k^T A e_j = a_{k,j}$

So $(Au, v) = (u, Av)$ $\Rightarrow a_{j,k} = a_{k,j}$ $\forall j, k$ $\Rightarrow A = A^T$.

One can similarly show that $A = A^T$ implies $(Au, v) = (u, Av)$ for all $u, v \in \mathbb{R}^N$.

Write $u = \sum_{k=1}^{N} u_k e_k$, $v = \sum_{j=1}^{N} v_j e_j$.

$$(Au, v) = \sum_{j=1}^{N}\sum_{k=1}^{N} v_j u_k e_j^T A e_k \quad \text{(linearity of the inner product)}$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{N} v_j u_k a_{j,k}$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{N} v_j u_k a_{k,j} \quad (\text{since } A = A^T,\ a_{j,k} = a_{k,j})$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{N} v_j u_k e_k^T A e_j$$

$$= (Av, u) = (u, Av) \quad \text{(by symmetry of the inner product.)}$$

Our goal is to apply the notion of symmetry to more general settings.

Theorem Let $L: C_0^2[0,1] \longrightarrow C[0,1]$ be defined by $Lu = -u''$. Then L is symmetric.

Proof Let $u, v \in C_0^2[0,1]$. Then

$$(Lu, v) = \int_0^1 -u''(x)\, v(x)\, dx$$

$$\overset{\text{IBP}}{=} \underbrace{\left[-u'(x)v(x)\right]_0^1}_{=0 \text{ since } v(0)=v(1)=0} + \int_0^1 u'(x)\, v'(x)\, dx$$

(Integration By Parts)

$$\overset{\text{IBP}}{=} \underbrace{\left[u(x)v'(x)\right]_0^1}_{=0 \text{ since } u(0)=u(1)=0} - \int_0^1 u(x)\, v''(x)\, dx$$

$$= \int_0^1 u(x)\,(-v''(x))\, dx = (u, Lv). \qquad \blacksquare$$

This proof follows the usual pattern for proving symmetry of a linear differential operator: use integration by parts to move derivatives from u to v, while using boundary conditions to see that the boundary terms in IBP are zero. We will use variations of this technique often!

Note that the same argument applies to the more general operator from the heat equation:

$$L: C_0^2[0,1] \to C[0,1], \quad Lu = -(k(x)u'(x))', \quad k(x) > 0.$$

Then
$$(Lu, v) = \int_0^1 -(k(x)u'(x))'v(x)\,dx$$

$$\text{IBP} = \underbrace{\left[-k(x)u'(x)v(x)\right]_0^1}_{=0 \text{ since } v(0)=v(1)=0} + \int_0^1 k(x)u'(x)v'(x)\,dx$$

(Note how $k(x)$ moves from u to v...)

$$\text{IBP} = \underbrace{\left[k(x)u(x)v'(x)\right]_0^1}_{=0 \text{ since } u(0)=u(1)=0} - \int_0^1 u(x)(k(x)v'(x))'\,dx$$

$$= (u, Lv).$$

Def A linear operator $L: \mathcal{U} \to \mathcal{V}$ has an eigenvalue λ with corresponding eigenvector/eigenfunction/eigenmode $\psi \in \mathcal{U}$ provided: $\psi \neq 0$ and

$$L\psi = \lambda\psi.$$

(We don't allow $\psi = 0$ because then $L\psi = 0 = \lambda\psi$ for *all* λ, so every point λ would look like an eigenvalue!)

Lecture 10: Eigenvalues and Eigenfunctions

Last time: $L\psi = \lambda\psi$, $\psi \neq 0$

$\Rightarrow \lambda$ is an eigenvalue of L corresponding to the eigenfunction ψ.

Recall that matrices $A \in \mathbb{R}^{N\times N}$ can have <u>complex</u> eigenvalues and eigenvectors. For example, the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

has eigenvalue $\lambda = i$, eigenvector $V = \begin{bmatrix} 1 \\ i \end{bmatrix}$

and eigenvalue $\lambda = -i$, eigenvector $V = \begin{bmatrix} 1 \\ -i \end{bmatrix}$.

To handle complex-valued functions and complex scalars, we must (temporarily) work with a slight generalization of our usual inner product.

<u>Def</u> A function $(\cdot,\cdot): V \times V \to \mathbb{C}$ is an inner product over the complex scalars if:

i) $(u+v, w) = (u,w) + (v,w)$, $(\alpha u, w) = \alpha(u,w)$ } linearity

ii) $(u,w) = \overline{(w,u)}$ } <u>conjugate</u> symmetry

iii) $(u,u) \geq 0$ and $(u,u) = 0 \Leftrightarrow u = 0$. } Positivity

Note that i) and ii) imply that

$$(u, \beta w) = \overline{(\beta w, u)} = \bar{\beta}\,\overline{(w,u)} = \bar{\beta}\,(u,w)$$

So the scalar in the second argument picks up a conjugate bar when you pull it out of the inner product.

Thus a complex inner product is called a "sesquilinear form".

(one-and-a-half times – like sesquicentenial = 150 years ...)

We will only use complex inner products in the next two proofs.

Theorem Let λ be an eigenvalue of a symmetric linear operator. Then λ is real.

Proof Let λ be an eigenvalue of the linear operator L, and let ψ be an eigenfunction of L corresponding to λ. Since we can multiply ψ by any nonzero scalar and we still have an eigenfunction, we can assume that $(\psi, \psi) = \|\psi\|^2 = 1$.

($L\psi = \lambda\psi \Rightarrow L(\alpha\psi) = \lambda(\alpha\psi)$ for any $\alpha \neq 0$).

Then
$$\begin{aligned}
\lambda = \lambda(\psi,\psi) &= (\lambda\psi, \psi) \\
&= (L\psi, \psi) && \text{since } L\psi = \lambda\psi \\
&= (\psi, L\psi) && \text{since } L \text{ is symmetric} \\
&= (\psi, \lambda\psi) && \text{since } L\psi = \lambda\psi \\
&= \overline{(\lambda\psi, \psi)} && \text{conjugate symmetry of i.p.} \\
&= \overline{\lambda}\,\overline{(\psi,\psi)} && \text{linearity of i.p. in first argument} \\
&= \overline{\lambda}\,(\psi,\psi) && \text{conjugate symmetry} \\
&= \overline{\lambda}. && \|\psi\|^2 = (\psi,\psi) = 1
\end{aligned}$$

Thus $\lambda = \overline{\lambda}$, which implies that $\lambda \in \mathbb{R}$. ∎

Did you follow that last step?

If $\lambda = \alpha + i\beta$, $\alpha, \beta \in \mathbb{R}$

then $\bar{\lambda} = \alpha - i\beta$ (definition of complex conjugation)

So if $\lambda = \bar{\lambda}$, then $i\beta = -i\beta$,

which is only possible if $\beta = 0$.

The last theorem has a beautiful companion.

Theorem Let L be a symmetric linear operator with eigenvalues λ and γ, $\lambda \neq \gamma$, with corresponding eigenfunctions ψ and ϕ. Then ψ and ϕ are orthogonal: $(\psi, \phi) = 0$.

Proof The proof is a simple chain of arguments:

$$\begin{aligned}
\lambda(\psi, \phi) = (\lambda\psi, \phi) &= (L\psi, \phi) && L\psi = \lambda\psi \\
&= (\psi, L\phi) && \text{symmetry of } L \\
&= (\psi, \gamma\phi) && L\phi = \gamma\phi \\
&= \overline{(\gamma\phi, \psi)} && \text{by conjugate symmetry of inner product} \\
&= \bar{\gamma}\,\overline{(\phi, \psi)} && \text{by linearity of the inner product} \\
&= \gamma\,\overline{(\phi, \psi)} && \text{since } \gamma \text{ is real by the last theorem.} \\
&= \gamma(\psi, \phi) && \text{by conjugate symmetry of the inner product}
\end{aligned}$$

Hence $(\lambda - \gamma)(\psi, \phi) = 0$.

Since $\lambda \neq \gamma$, this implies $(\psi, \phi) = 0$:

the eigenfunctions of symmetric linear operators are orthogonal. ∎

Lecture 11: Eigenfunctions of the Laplacian; Spectral Method

LAST TIME
- Symmetric linear operators have real eigenvalues.
- Eigenfunctions associated with distinct eigenvalues are orthogonal.

Henceforth we shall not need complex-valued inner products; real inner products shall suffice.

We now seek to compute eigenvalues and eigenfunctions of the Laplacian operator

$$L: C_0^2[0,1] \longrightarrow C[0,1], \quad Lu = -u''.$$

(λ, ψ) is an eigenvalue-eigenfunction pair ("eigenpair") when $L\psi = \lambda\psi$.

This translates to an (ordinary) differential equation boundary value problem: Find λ, ψ such that

$$-\psi''(x) = \lambda\psi(x), \qquad \psi(0) = \psi(1) = 0.$$

Differential equations of the form $-\psi''(x) = \lambda\psi(x)$ have the general solution

$$\psi(x) = A\sin(\sqrt{\lambda}\,x) + B\cos(\sqrt{\lambda}\,x)$$

for constants A and B. (Which functions are a negative multiple of their second derivative? sine & cosine)

Can we find A, B, λ to satisfy the boundary conditions $\psi(0) = \psi(1) = 0$?

left boundary condition:

$$0 = \psi(0) = A\sin(\sqrt{\lambda}\cdot 0) + B\cos(\sqrt{\lambda}\cdot 0)$$
$$= A\cdot 0 + B\cdot 1 = B$$

$$\Longrightarrow \boxed{B = 0.}$$

right boundary condition: $0 = \psi(1) = A\sin(\sqrt{\lambda}\cdot 1) = A\sin(\sqrt{\lambda})$.

This implies either $A=0$ or $\sin(\sqrt{\lambda})=0$.

If $A=0$, then $\psi(x)=0 \quad \forall x\in[0,1]$:
we do not allow the zero function to be an eigenfunction.

Thus we need $\sin(\sqrt{\lambda})=0$

$\Longrightarrow \sqrt{\lambda}$ is an integer multiple of π:

$$\sqrt{\lambda_n} = n\pi \Longrightarrow \lambda_n = n^2\pi^2$$

$$\psi_n(x) = A\sin(n\pi x).$$

- We do not allow $n=0$ (then $\psi(x)=0$: not allowed)
- If n is negative, we get the same eigenvalue and eigenfunction:

$$\left.\begin{aligned}(-n)^2\pi^2 &= n^2\pi^2\\ \sin(-n\pi x) &= -\sin(n\pi x)\end{aligned}\right\}\ \text{same eigenvalue and eigenfunction.}$$

Thus the eigenvalues and eigenfunctions of the Laplacian are:

$$\boxed{\begin{aligned}\lambda_n &= n^2\pi^2 \qquad n=1,2,3,\dots\\ \psi_n(x) &= \sqrt{2}\sin(n\pi x)\end{aligned}}$$

We pick $A=\sqrt{2}$ so that $(\psi_n,\psi_n) = \|\psi_n\|^2 = 1$;
this is merely a convenient choice.

Note: different boundary conditions can give very different eigenvalues and eigenfunctions.

See demo lapeig_dd.m.

Section 6 THE SPECTRAL METHOD

If we want to solve the matrix equation $Ax=b$ for the unknown x, we use GAUSSIAN ELIMINATION. But there is no easy way to generalize GAUSSIAN ELIMINATION TO general operator equations $Lu=f$. Here we derive another strategy: "the spectral method", that uses eigenvalues and eigenfunctions. (The term "spectral" comes from the term "spectrum" — a generalization of the set of eigenvalues, closely connected to the term "spectrum" in chemical analysis.....)

KEY IDEA FIND THE FUNCTION $u_N \in \text{span}\{\psi_1,\dots,\psi_N\}$ THAT MINIMIZES THE ERROR $\|f-Lu_N\|$:

$$\|f-Lu_N\| = \min_{\hat{u}\in \text{span}\{\psi_1,\dots,\psi_N\}} \|f-L\hat{u}\|.$$

We assume L is a symmetric linear operator with eigenvalues $\lambda_1,\lambda_2,\dots$ and corresponding eigenfunctions $\psi_1,\psi_2,\dots$, and that $\psi_1,\psi_2,\dots$ are orthogonal. (Since L is symmetric, the only concern arises if $\lambda_j=\lambda_k$ for some $j\neq k$. In this case, we can still pick ψ_j and ψ_k to be orthogonal eigenfunctions in the cases we encounter in this class.)

The Best Approximation Theorem says that

LU_N is the best approximation to f

from $\text{Span}\{L\psi_1, \dots, L\psi_N\}$

$= \text{Span}\{\lambda_1\psi_1, \dots, \lambda_N\psi_N\} = \text{Span}\{\psi_1, \dots, \psi_N\}$

(provided all $\lambda_j \neq 0$.)

if and only if

$$(f - LU_N, v) = 0 \quad \text{for all } v \in \text{Span}\{L\psi_1, \dots, L\psi_N\} = \text{Span}\{\psi_1, \dots, \psi_N\}$$

Assume L has no zero eigenvalues for the rest of this discussion.

We will enforce $(f - LU_N, v) = 0$ for all $v \in \text{span}\{\psi_1, \dots, \psi_N\}$

Write $U_N(x) = c_1\psi_1(x) + \cdots + c_N\psi_N(x)$

$(f - LU_N, v) = 0$ for all $v \in \text{span}\{\psi_1, \dots, \psi_N\}$

if and only if $(f - LU_N, \psi_j) = 0$ for $j = 1, \dots, N$.

(Make sure you understand why this is so....)

$$0 = (f - LU_N, \psi_j) = \Big(f - L\underbrace{\sum_{k=1}^{N} c_k\psi_k}_{U_N}, \psi_j\Big)$$

$$= (f, \psi_j) - \Big(\sum_{k=1}^{N} c_k L\psi_k, \psi_j\Big)$$

$$= (f, \psi_j) - \sum_{k=1}^{N} c_k (L\psi_k, \psi_j) \quad \text{linearity of the inner product}$$

$$= (f, \psi_j) - \sum_{k=1}^{N} c_k (\lambda_k\psi_k, \psi_j) \quad (L\psi_k = \lambda_k\psi_k)$$

$$= (f, \psi_j) - \sum_{k=1}^{N} c_k\lambda_k \underbrace{(\psi_k, \psi_j)}_{= 0 \text{ if } k \neq j}$$

orthogonality of eigenfunctions.

Thus $0 = (f - Lu_N, \psi_j) = (f, \psi_j) - c_j \lambda_j (\psi_j, \psi_j)$.

This implies that

$$\boxed{c_j = \frac{1}{\lambda_j} \cdot \frac{(f, \psi_j)}{(\psi_j, \psi_j)}}$$ (recall we assume $\lambda_j \neq 0$)

We arrive at the spectral method solution:

$$\boxed{u_N(x) = \sum_{j=1}^{N} \frac{1}{\lambda_j} \frac{(f, \psi_j)}{(\psi_j, \psi_j)} \psi_j(x)}$$

Compare this to $f_N(x) = \sum_{j=1}^{N} \frac{(f, \psi_j)}{(\psi_j, \psi_j)} \psi_j(x)$, $= Lu_N$

the best approximation to f from $\text{span}\{\psi_1, \ldots, \psi_N\}$.

We expect this will tend in the limit to

$$u(x) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \frac{(f, \psi_j)}{(\psi_j, \psi_j)} \psi_j(x)$$

a series formula for the solution of $Lu = f$.

«The spectral method and symmetric matrices.»

Let $A \in \mathbb{R}^{N\times N}$ be a symmetric matrix, no zero eigenvalues.
We want to confirm that the spectral method leads to the correct solution of $Ax=b$.

Suppose $Av_j = \lambda_j v_j$ for eigenvectors $v_1, \dots, v_N$

such that $v_j^T v_k = \begin{cases} 1 & j=k \quad (\text{i.e., } \|v_j\|=1 \\ 0 & j\neq k \end{cases}$

$$\left\{ Av_1 = \lambda_1 v_1 \quad Av_2 = \lambda_2 v_2 \quad \cdots \quad Av_N = \lambda_N v_N \right\}$$

STACK THESE N EQUATIONS AS COLUMNS OF A MATRIX

$$\left[Av_1 \,\middle|\, Av_2 \,\middle|\, \cdots \,\middle|\, Av_N \right] = \left[\lambda_1 v_1 \,\middle|\, \lambda_2 v_2 \,\middle|\, \cdots \,\middle|\, \lambda_N v_N \right]$$

FACTOR THESE MATRICES:

$$A \left[v_1 \,\middle|\, v_2 \,\middle|\, \cdots \,\middle|\, v_N \right] = \left[v_1 \,\middle|\, v_2 \,\middle|\, \cdots \,\middle|\, v_N \right] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix} \text{ (diagonal)}$$

↑ Post-multiplying by a diagonal matrix scales columns.

Write this as

$$AV = V\Lambda$$

Orthogonality of eigenvectors $\Rightarrow V^T V = I$

So $V^{-1} = V^T$, hence $VV^T = I$. too.

Thus $A = V \Lambda V^T$ (diagonalization of A)

We can invert: $A^{-1} = V \Lambda^{-1} V^T$

Confirm: $A A^{-1} = (V \Lambda V^T)(V \Lambda^{-1} V^T)$

$$= V \Lambda \underbrace{(V^T V)}_{I} \Lambda^{-1} V^T$$

$$= V \underbrace{(\Lambda \Lambda^{-1})}_{=I} V^T = \underbrace{V V^T}_{I} = I \quad \checkmark$$

Thus $Ax = b$ is solved by $x = A^{-1} b$

$$X = A^{-1} b = V \Lambda^{-1} V^T b$$

$$= \left[v_1 \,\middle|\, \cdots \,\middle|\, v_N \right] \begin{bmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_N \end{bmatrix} \left[\begin{array}{c} v_1^T \\ \hline \vdots \\ \hline v_N^T \end{array} \right] b$$

$$= \left[v_1 \,\middle|\, \cdots \,\middle|\, v_N \right] \begin{bmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_N \end{bmatrix} \begin{bmatrix} v_1^T b \\ \hline \vdots \\ \hline v_N^T b \end{bmatrix}$$

$$= \left[v_1 \,\middle|\, \cdots \,\middle|\, v_N \right] \begin{bmatrix} \frac{v_1^T b}{\lambda_1} \\ \hline \vdots \\ \hline \frac{v_N^T b}{\lambda_N} \end{bmatrix}$$

$$= \sum_{j=1}^{N} \frac{v_j^T b}{\lambda_j} v_j = \sum_{j=1}^{N} \frac{(b, v_j)}{\lambda_j} v_j$$

$$= \sum_{j=1}^{N} \frac{1}{\lambda_j} \frac{(b, v_j)}{(v_j, v_j)} v_j$$

(Since $(v_j, v_j) = \|v_j\|^2 = 1$)

This is precisely the spectral method solution!

Next pages: worked examples of the spectral method...

Two Examples of Solving $-u''(x) = f(x)$

We wish to solve $-u''(x) = f(x)$ with homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$ for three different choices of f. The key idea here is that the smoothness of f will vary in these three examples, which will be reflected in the decay rates of the inner products (ψ_n, f).

Let $L : C_D^2[0,1] \to C[0,1]$ be given by $Lu = -u''$.

In what follows, denote the nth eigenvalue and eigenfunction of L by

$$\lambda_n = n^2\pi^2, \qquad \psi_n(x) = \sqrt{2}\sin(n\pi x).$$

Notice that the leading $\sqrt{2}$ factor in ψ_n ensures that $(\psi_n, \psi_n) = 1$, i.e., it makes ψ_n a unit vector.

1. Solve $-u''(x) = 1$, $u(0) = u(1) = 0$.

   We can find the exact solution by just integrating twice and using the boundary parameters to determine the constants of integration. This gives $u(x) = (x - x^2)/2$.

   However (anticipating time dependent problems to come, like $u_t = u_{xx}$), we will solve this by eigenfunction expansion. First we will consider expansions of f (i.e., the limits of best approximations):

   $$1 = f(x) = \sum_{n=1}^{\infty} \frac{(\psi_n, f)}{(\psi_n, \psi_n)}\psi_n(x)$$

   and the solution

   $$u(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n}\frac{(\psi_n, f)}{(\psi_n, \psi_n)}\psi_n(x).$$

   For both we must compute

   $$(\psi_n, f) = (\sqrt{2}\sin(n\pi x), 1) = \int_0^1 \sqrt{2}\sin(n\pi x)\cdot 1\, dx = \frac{\sqrt{2}(1-(-1)^n)}{n\pi}.$$

   The first few values are:

   $$(\psi_1, f) = \frac{2\sqrt{2}}{\pi}, \quad (\psi_2, f) = 0, \quad (\psi_3, f) = \frac{2\sqrt{2}}{3\pi}, \quad (\psi_4, f) = 0, \quad (\psi_5, f) = \frac{2\sqrt{2}}{5\pi}.$$

   Thus we might want to write

   $$\begin{aligned} 1 = f(x) = \sum_{n=1}^{\infty} \frac{(\psi_n, f)}{(\psi_n, \psi_n)}\psi_n(x) &= \sum_{n=1}^{\infty} \frac{\sqrt{2}(1-(-1)^n)}{n\pi}(\sqrt{2}\sin(n\pi x)) \\ &= \sum_{n=1}^{\infty} \frac{2(1-(-1)^n)}{n\pi}\sin(n\pi x). \end{aligned}$$

The right-hand side looks like an absurd way to write $f(x) = 1$, but you can see that it seems to work. Run `lapex1.m`.

We can also write the solution as

$$\frac{x - x^2}{2} = u(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(\psi_n, f)}{(\psi_n, \psi_n)} \psi_n(x) = \sum_{n=1}^{\infty} \frac{1}{n^2\pi^2} \frac{\sqrt{2}(1-(-1)^n)}{n\pi} (\sqrt{2}\sin(n\pi x))$$
$$= \sum_{n=1}^{\infty} \frac{2(1-(-1)^n)}{n^3\pi^3} \sin(n\pi x).$$

Notice in `lapex1.m` that the series for u converges so much faster than the series for f ! Why? First, note that u satisfies the boundary conditions, but f doesn't (and there is no need for it to do so) - hence the eigenfunctions (which must obey the boundary conditions) can approximate u better than f. Secondly, notice that dividing by $\lambda_n = n^2\pi^2$ makes the coefficients in the series decay like $1/n^3$ in the series for u, instead of $1/n$ as in the series for f. Faster decay of the coefficients means faster convergence.

2. Solve $-u''(x) = f(x)$, $u(0) = u(1) = 0$, where

$$f(x) = \begin{cases} x, & x \in [0, 1/2]; \\ 1 - x, & x \in [1/2, 1]. \end{cases}$$

Notice that this function f is in $C[0, 1]$ but not $C^1[0, 1]$, because the derivative of f is discontinuous at $1/2$. However, $f(0) = f(1) = 0$, so you might have some hope that the eigenfunctions will do a better job of approximating f, since $\psi_n(0) = \psi_n(1) = 0$.

Can you find the solution exactly? (Integrate twice on each half of the domain; you will have four constants (two on each domain). Use the boundary conditions to set two of the integration constants; enforce continuity of $u(1/2)$ and $u'(1/2)$ to find the other two.

One can now compute

$$(\psi_n, f) = \int_0^1 \psi_n(x) f(x)\, dx = \frac{8\sqrt{2}\cos(n\pi/4)\sin^3(n\pi/4)}{n^2\pi^2}.$$

The first few values are:

$$(\psi_1, f) = \frac{2\sqrt{2}}{\pi^2}, \quad (\psi_2, f) = 0, \quad (\psi_3, f) = -\frac{2\sqrt{2}}{9\pi^2}, \quad (\psi_4, f) = 0, \quad (\psi_5, f) = \frac{2\sqrt{2}}{25\pi^2}.$$

Now the solution is

$$u(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(\psi_n, f)}{(\psi_n, \psi_n)} \psi_n(x) = \sum_{n=1}^{\infty} \frac{16\cos(n\pi/4)\sin^3(n\pi/4)}{n^4\pi^4} \sin(n\pi x),$$

which converges even faster than in the last case, because the series for f converged even faster. Notice (by running `lapex2.m`) that f' is not continuous as $x = 1/2$, but the solution u indeed appears smooth.

Lecture 13: Convergence of Spectral Method

In our examples, it appears that $u_N \to u$ very quickly as N increases.

What can we say about the convergence of $u_N \to u$?

This topic is largely beyond the scope of this class, but we can give a few coarse indications.

First, it is absurd to say that

$$f_N(x) = \sum_{j=1}^{N} \frac{(f, \psi_j)}{(\psi_j, \psi_j)} \psi_j(x)$$

converges to $f(x)$ for all $x \in [0,1]$, in general.

We have seen how we can approximate $f(x) = 1$ with f_N, but always $\psi_j(0) = \psi_j(1) = 0$ means

$$f_N(0) = f_N(1) = \sum_{j=1}^{N} \frac{(f, \psi_j)}{(\psi_j, \psi_j)} \underbrace{\psi_j(x = 0 \text{ or } x = 1)}_{0} = 0.$$

So $f_N(0) \not\to f(0)$

$f_N(1) \not\to f(1)$.

When we say $f_N \to f$, we mean

$$\| f_N - f \| \longrightarrow 0 \quad \text{as} \quad N \to \infty.$$

In other words, the x values for which $f_N(x) \not\to f(x)$ is very small. (One can make this rigorous with measure theory / functional analysis.)

Another perspective: how rapidly do the values of $\dfrac{(f,\psi_n)}{(\psi_n,\psi_n)} \longrightarrow 0$ as $n\to\infty$?

Assume f is sufficiently differentiable for the calculation below to make sense. Then

$$\int_0^1 f(x)\sin(n\pi x)\,dx$$

$$\text{IBP} = \left[-\frac{f(x)\cos(n\pi x)}{n\pi}\right]_0^1 + \int_0^1 \frac{f'(x)\cos(n\pi x)}{n\pi}\,dx$$

$$= \left(\frac{f(0)-(-1)^n f(1)}{n\pi}\right) + \int_0^1 \frac{f'(x)\cos(n\pi x)}{n\pi}\,dx$$

$$\text{IBP} = \left(\frac{f(0)-(-1)^n f(1)}{n\pi}\right) + \underbrace{\left[\frac{f'(x)\sin(n\pi x)}{n^2\pi^2}\right]_0^1}_{0} - \int_0^1 \frac{f''(x)\sin(n\pi x)}{n^2\pi^2}\,dx$$

$$= \left(\frac{f(0)-(-1)^n f(1)}{n\pi}\right) + \int_0^1 \frac{f''(x)\sin(n\pi x)}{n^2\pi^2}\,dx.$$

Hence

$$(f,\psi_n) = \int_0^1 f(x)\left(\sqrt{2}\sin(n\pi x)\right)dx$$

$$= \left(\frac{\sqrt{2}}{n\pi}\left(f(0)-(-1)^n f(1)\right)\right) + \underbrace{\int_0^1 \frac{f''(x)}{n^2\pi^2}\sqrt{2}\sin(n\pi x)\,dx}_{\text{CAUCHY-SCHWARZ}}$$

C-S $\Rightarrow$

$$\left|\int_0^1 \frac{f''(x)}{n^2\pi^2}\sqrt{2}\sin(n\pi x)\,dx\right| = \left|\left(\frac{f''}{n^2\pi^2},\psi_n\right)\right|$$

$$\le \left\|\frac{f''}{n^2\pi^2}\right\| \underbrace{\|\psi_n\|}_{=1} = \frac{1}{n^2\pi^2}\|f''\|.$$

Thus we have this very interesting bound:

$$\boxed{|(f,\psi_n)| \le \frac{\sqrt{2}}{n\pi}\left|f(0)-(-1)^n f(1)\right| + \frac{\|f''\|}{n^2\pi^2}} \qquad (*)$$

See the example from the last lecture: $f(x)=1$.

$$(f,\psi_n) = \frac{\sqrt{2}}{n\pi}\left(1-(-1)^n\right)$$

precisely keeping with the bound $(*)$ ($f''\equiv 0$ in this case.)

So $|(f,\psi_n)| \to 0$ as $n\to\infty$.

Now if $f(0)=f(1)=0$ (so the function f satisfies the boundary conditions of the problem) then

$$f(0)=f(1)=0 \implies f(0)-(-1)^n f(1)=0$$

$$\Rightarrow |(f,\psi_n)| \le \frac{\|f''\|}{n^2\pi^2}$$

So the coefficients decay even faster with n.

If we keep integrating by parts, and $f^{(k)}$ is sufficiently differentiable and continues to satisfy the boundary conditions, we get faster decay still.

In particular, the spectral method gives coefficients

$$\frac{1}{\lambda_n}\frac{(f,\psi_n)}{(\psi_n,\psi_n)} = \frac{1}{n^2\pi^2}\frac{(f,\psi_n)}{(\psi_n,\psi_n)}$$

So the coefficients for u decay very quickly
$\Rightarrow$ FAST CONVERGENCE OF THE SPECTRAL METHOD!

① How should we handle inhomogeneous boundary conditions?

$$-u''(x) = f(x), \quad u(0) = \alpha, \quad u(1) = \beta.$$

Note that we can't incorporate inhomogeneous b.c.'s into the operator. For example,

$$\{u \in C^2[0,1] : u(0) = \alpha, \ u(1) = \beta\}$$

fails the conditions for being a subspace. (Why? Try it.)
Similarly, it makes no sense to construct eigenfunctions that satisfy these boundary conditions... (Multiples of such "eigenfunctions" would fail to be eigenfunctions.)

Instead, here is a better approach. Construct the solution in two parts:

$$u(x) = v(x) + w(x)$$

v will satisfy	w will satisfy
$-v'' = f$	$-w'' = 0$
but have	but have
$v(0) = v(1) = 0$	$w(0) = \alpha, \ w(1) = \beta.$

First, consider $w(x)$: we need $w''(x) = 0$

$$\Rightarrow \quad w(x) = c + dx$$

We need $\alpha = w(0) = c + d \cdot 0 = c \Rightarrow c = \alpha$

$$\beta = w(1) = \alpha + d \cdot 1 \Rightarrow d = \beta - \alpha.$$

So $\boxed{w(x) = \alpha + (\beta - \alpha)x}$.

Now we need $-v''(x) = f(x)$ with $v(0) = v(1) = 0$.

Construct v using the spectral method for homogeneous boundary conditions:

$$v(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)$$

when $\lambda_n = n^2\pi^2$, $\psi_n(x) = \sqrt{2}\sin(n\pi x)$

as usual.

Solution to $-u''(x) = f(x)$, $u(0) = \alpha$, $u(1) = \beta$

is then

$$\boxed{u(x) = \alpha + (\beta - \alpha)x + \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)}$$

(2) What if the boundary conditions change type? Then we need to compute new eigenvalues and eigenfunctions.

Example $-u''(x) = f(x)$, $u(0) = 0$ ← Dirichlet condition on the left.

$u'(1) = 0$ ← Neumann condition on right.

Define $C_m^2[0,1] = \{ u \in C^2[0,1] : u(0) = 0, \; u'(1) = 0 \}$

("m" for "mixed")

$$\left\{ \begin{array}{l} L : C_m^2[0,1] \longrightarrow C[0,1] \\ Lu = -u''. \\ \text{Solve } Lu = f. \end{array} \right\} \Leftrightarrow \begin{array}{l} -u''(x) = f(x) \\ u(0) = 0 \\ u'(1) = 0. \end{array}$$

- Is L symmetric?

Let $u, v \in C^2_m[0,1]$.

$$(Lu, v) = \int_0^1 -u''(x)\, v(x)\, dx$$

$$(IBP) = \underbrace{\Big[-u'(x)\, v(x)\Big]_0^1}_{\substack{u'(1)=0,\ v(0)=0 \\ \Rightarrow \text{boundary term is zero}}} + \int_0^1 u'(x)\, v'(x)\, dx$$

$$= \int_0^1 u'(x)\, v'(x)\, dx$$

$$(IBP) = \underbrace{\Big[u(x)\, v'(x)\Big]_0^1}_{\substack{v'(1)=0,\ u(0)=0 \\ \text{boundary term is zero}}} - \int_0^1 u(x)\, v''(x)\, dx$$

$$= \int_0^1 u(x)\,(-v''(x))\, dx = (u, Lv)$$

Thus L is symmetric.

- Compute eigenvalues and eigenfunctions of L.

Solve $L\psi = \lambda\psi \iff -\psi''(x) = \lambda\psi(x)$ $\begin{cases} \psi(0)=0 \\ \psi'(1)=0 \end{cases}$

General solution of $-\psi'' = \lambda\psi$:

$$\psi(x) = A\sin(\sqrt{\lambda}\,x) + B\cos(\sqrt{\lambda}\,x)$$

$$0 = \psi(0) = A\sin(0) + B\cos(0) = B \Rightarrow \boxed{B=0.}$$

$$\Rightarrow \quad \psi(x) = A\sin(\sqrt{\lambda}x)$$

$$\psi'(x) = A\sqrt{\lambda}\cos(\sqrt{\lambda}x)$$

$$0 = \psi'(1) = A\sqrt{\lambda}\cos(\sqrt{\lambda}).$$

$$\Rightarrow \text{either } \lambda = 0 \quad \text{or} \quad \cos(\sqrt{\lambda}) = 0.$$

If $\lambda = 0$, $\psi(x) = A\sin(0\cdot x) = 0$

This cannot be an eigenfunction.

Thus $\cos(\sqrt{\lambda}) = 0$. This implies that

$$\sqrt{\lambda} = \frac{\pi}{2}, \ \frac{3\pi}{2}, \ \frac{5\pi}{2}, \ldots = \frac{(2n-1)\pi}{2}, \quad n = 1, 2, \ldots$$

$$\Rightarrow \boxed{\begin{aligned} &\lambda_n = \left(\tfrac{2n-1}{2}\right)^2 \pi^2 \qquad n = 1, 2, \ldots \\ &\psi_n(x) = \sqrt{2}\sin(\sqrt{\lambda_n}\,x) = \sqrt{2}\sin\left(\tfrac{(2n-1)}{2}\pi x\right) \end{aligned}}$$

Now solve $-u''(x) = f(x)$, $u(0) = u'(1) = 0$ via the spectral method with the new eigenvalues and eigenfunctions:

$$u(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)$$

$$= \sum_{n=1}^{\infty} \frac{2^2}{(2n-1)^2\pi^2} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \left(\sqrt{2}\sin\left(\tfrac{(2n-1)}{2}\pi x\right)\right).$$

Note: different b.c.'s can give very different eigenvalues and eigenfunctions. See Lecture 17 for $u'(0) = u'(1) = 0$; further examples are on Problem Set 3.

(3) How can we handle variable coefficients?

Recall the steady-state heat equation

$$-(k(x)u'(x))' = f(x) \qquad u(0)=0,\ u(1)=0$$

where $k(x) > 0$ can vary with space (to describe an inhomogeneous bar in the heat model — material properties vary with space).

To solve this equation via the spectral method, we need to find eigenvalues and eigenfunctions of

$$L : C_0^2[0,1] \to C[0,1]$$
$$Lu = -(k(x)u'(x))'$$

L is a symmetric operator. To compute its eigenvalues and eigenfunctions, we must solve

$$L\psi = \lambda\psi, \quad \text{i.e.,}$$
$$-(k(x)\psi'(x))' = \lambda\psi(x), \quad \psi(0)=\psi(1)=0.$$

In general, this is a difficult differential equation to solve. The $k(x)\equiv 1$ case is a special (important) case. In general, for most $k(x)$ we cannot write down eigenvalues & eigenfunctions — or we can find eigenfunctions, but we must solve a transcendental equation to find eigenvalues.

This suggests that the Spectral method is limited to "nice" problems like $-u'' = f$.

To illustrate how it could work in more general circumstances, we provide a script "eigendemo3.m" on the class website, which uses Chebfun / Chebops function to numerically approximate the eigenvalues and eigenfunctions $\psi_1, \ldots, \psi_N$ up to some limit N. This illustrates how the Spectral Method would work, even though we don't have an explicit form for ~~the~~ λ_n and ψ_n.

Next, we consider a very different approach to approximating the solution — the finite element method.

Lecture 15: Weak Form of the Differential Equation and Galerkin's Method

Section 7 Approximate solution of differential equations via the finite element Method.

Begin with the general form of the steady-state heat equation with Dirichlet boundary conditions:

$$-(K(x)u'(x))' = f(x), \quad u(0)=u(1)=0.$$

For most nontrivial choices of the material parameter $K(x)>0$, we cannot exactly compute the eigenvalues and eigenfunctions, so the spectral method is limited. We instead seek approximate solutions via a different approach.

Weak Form of the Diff Eq.

Let $v \in C_0^2[0,1]$. We call this generic v a "test function".

If $-(k(x)u'(x))' = f(x), \quad u(0)=u(1)=0,$

then

$$-(K(x)u'(x))'v(x) = f(x)v(x)$$

and

$$\int_0^1 -(K(x)u'(x))'v(x)\,dx = \int_0^1 f(x)v(x)\,dx$$

Integrate the LHS by parts to obtain

$$\underbrace{\left[-k(x)\,u'(x)\,v(x)\right]_0^1}_{=0} + \int_0^1 k(x)\,u'(x)\,v'(x)\,dx = \int_0^1 f(x)\,v(x)\,dx$$

$= 0$ since $v \in C_0^2[0,1] = \{u \in C^2[0,1] : u(0) = u(1) = 0\}$

$\Rightarrow v(0) = v(1) = 0$

So

$$\underbrace{\int_0^1 k(x)\,u'(x)\,v'(x)\,dx}_{\text{Write this as } a(u,v)} = \underbrace{\int_0^1 f(x)\,v(x)\,dx}_{\text{write this as } (f,v)}$$

$a(u,v)$ — "energy inner product"

$=$

(f,v) ↑ Standard inner product

Def The energy inner product associated with $-(k(x)\,u'(x))' = f(x)$ is given by

$$a(u,v) = \int_0^1 k(x)\,u'(x)\,v'(x)\,dx.$$

One can show that $a(u,v)$ defines a valid inner product on $C_0^2[0,1]$.

In particular, the Dirichlet boundary conditions ensure that

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

only when $u = 0$. (Think about why this is the case.)

Thus we have the <u>weak form</u> of the differential equation:

> If $-(k(x)u'(x))' = f(x)$, $u(0)=u(1)=0$, then
>
> $$a(u,v) = (f,v) \quad \text{for all } v \in C_0^2[0,1].$$

Why "weak"? Note that $a(u,v)=(f,v)$ only involves first derivatives — in other PDE settings, we can look for solutions that do not have enough derivatives to be "strong" solutions of the original PDE, but they still satisfy the weak form.

In this case, we can show that if $u \in C_0^2[0,1]$ satisfies the <u>weak form</u>, it must also satisfy the <u>strong form</u>.

Suppose $a(u,v) = (f,v)$ for all $v \in C_0^2[0,1]$.

Then $$\int_0^1 k(x)u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx$$

$$\text{IBP} \Rightarrow \underbrace{\left[-(k(x)u'(x))v(x)\right]_0^1}_{=0 \text{ since } v \in C_0^2[0,1]} - \int_0^1 (k(x)u'(x))'v(x)\,dx = \int_0^1 f(x)v(x)\,dx$$

$$\Rightarrow \int_0^1 -(k(x)u'(x))'v(x)\,dx = \int_0^1 f(x)v(x)\,dx$$

$$\Rightarrow \int_0^1 \left(\left(-k(x)u'(x) \right)' - f(x) \right) v(x)\, dx = 0$$

for all $v \in C_0^2[0,1]$.

So the potential "mismatch"

$$-\left(k(x)u'(x)\right)' - f(x)$$

is orthogonal to all $v \in C_0^2[0,1]$.

In particular, if we can take

$$v(x) = -\left(k(x)\,u'(x)\right)' - f(x)$$

then

$$0 = \int_0^1 \left(\left(-k(x)u'(x) \right)' - f(x) \right) v(x)\, dx$$

$$= \int_0^1 \left(\left(-k(x)u'(x) \right)' - f(x) \right) \left(\left(-k(x)u'(x) \right)' - f(x) \right) dx$$

$$= \left\| \left(-k(x)u'(x) \right)' - f(x) \right\|^2$$

So the mismatch is zero!

Now, we can't be certain that $\left(-k(x)u'(x)\right)' - f(x) \in C_0^2[0,1]$

but we can approximate it to arbitrary accuracy from $C_0^2[0,1]$ functions. (Think about this...)

Thus, a solution $u \in C_0^2[0,1]$ to the weak form is also a solution to the strong form

STRONG SOLUTION $\Longleftrightarrow$ WEAK SOLUTION

GALERKIN APPROXIMATION

We want to use the weak form to derive a computational scheme for approximating the solution $u \in C_0^2[0,1]$.

Since $C_0^2[0,1]$ is an infinite dimensional space, there is no efficient way to find $u \in C_0^2[0,1]$ such that

$$a(u,v) = (f,v) \quad \text{for} \quad \underline{\underline{\text{all}}} \quad v \in C_0^2[0,1].$$

GALERKIN APPROXIMATION imposes the weak form only on a finite dimensional subspace V_N of $C_0^2[0,1]$.

In response, we shall only look for approximations u_N (to u) that come from $C_0^2[0,1]$ as well.

> GALERKIN PROBLEM. Find $u_N \in V_N \subseteq C_0^2[0,1]$ such that
>
> $$a(u_N, v) = (f,v) \quad \text{for all } v \in V_N.$$

Suppose $V_N = \text{span}\{\phi_1, \dots, \phi_N\}$, where $\phi_1, \dots, \phi_N$ are some linearly independent functions in $C_0^2[0,1]$. Then write

$$v = d_1\phi_1 + \cdots + d_N\phi_N.$$

Note: $a(u_N, v) = (f,v)$ implies

$$\sum_{j=1}^{N} d_j\, a(u_N, \phi_j) = \sum_{j=1}^{N} d_j\, (f, \phi_j)$$

Thus $a(u_N, v) = (f,v)$ for <u>all</u> $v \in V_N$ if and only if

$$a(u_N, \phi_j) = (f, \phi_j) \quad \text{for } j = 1, \dots, N.$$

Now write $u_N(x) = c_1\phi_1(x) + \cdots + c_N\phi_N(x)$.

We seek $c_1, \ldots, c_N$.

$$a(u_N, \phi_j) = (f, \phi_j) \Rightarrow \boxed{\sum_{k=1}^{N} c_k\, a(\phi_k, \phi_j) = (f, \phi_j)}$$

Taking this equation for each of $j = 1, \ldots, N$ gives the matrix equation

$$\underbrace{\begin{bmatrix} a(\phi_1,\phi_1) & \cdots & a(\phi_N,\phi_1) \\ \vdots & & \vdots \\ a(\phi_1,\phi_N) & \cdots & a(\phi_N,\phi_N) \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix}}_{c} = \underbrace{\begin{bmatrix} (f,\phi_1) \\ \vdots \\ (f,\phi_N) \end{bmatrix}}_{f}$$

K: "stiffness matrix"

f: "load vector"

Note: $K \in \mathbb{R}^{N\times N}$ is a symmetric matrix.

- Solve $Kc = f$ for $c \in \mathbb{R}^N$
- Construct $u_N(x) = c_1\phi_1(x) + \cdots + c_N\phi_N(x)$.

EXAMPLE

Solve $-u''(x) = f(x) \quad u(0)=u(1)=0 \quad$ (i.e., $k(x)\equiv 1$)

via Galerkin's method using basis functions

$$\phi_j(x) = \psi_j(x) = \sqrt{2}\sin(j\pi x) \qquad j=1,\dots,N,$$

i.e., try using the Galerkin method with eigenfunctions as basis vectors defining V_N.

Then

$$a(\phi_k,\phi_j) = \int_0^1 \phi_j'(x)\,\phi_k'(x)\,dx$$

$$= 2\int_0^1 \cos(j\pi x)(j\pi)\cos(k\pi x)(k\pi)\,dx$$

$$= 2jk\pi^2 \underbrace{\int_0^1 \cos(j\pi x)\cos(k\pi x)\,dx}$$

$$= \begin{cases} \frac{1}{2} & \text{if } j=k \\ 0 & \text{if } j\neq k \end{cases}$$

$$\Rightarrow K = \begin{pmatrix} \pi^2 & & 0 \\ & 4\pi^2 & \\ 0 & & \ddots \\ & & & N^2\pi^2 \end{pmatrix} \Rightarrow K^{-1} = \begin{pmatrix} \frac{1}{\pi^2} & & 0 \\ & \frac{1}{4\pi^2} & \\ 0 & & \ddots \\ & & & \frac{1}{N^2\pi^2} \end{pmatrix}$$

$$Kc = f \Rightarrow c = \begin{bmatrix} (f,\psi_1)/(\pi^2) \\ (f,\psi_2)/(4\pi^2) \\ \vdots \\ (f,\psi_N)/(N^2\pi^2) \end{bmatrix}$$

Thus, for this choice of V_N,

$$U_N = c_1 \phi_1(x) + \cdots + c_N \phi_N(x)$$

$$= \frac{(f_1, \psi_1)}{\pi^2} \psi_1(x) + \cdots + \frac{(f_1, \psi_N)}{N^2 \pi^2} \psi_N(x)$$

$$= \sum_{j=1}^{N} \frac{1}{j^2 \pi^2} \frac{(f_1, \psi_j)}{(\psi_j, \psi_j)} \psi_j(x)$$

[Note: we have normalized so $(\psi_j, \psi_j) = 1$]

$$= \text{PARTIAL SUM OF SPECTRAL METHOD SOLUTION.}$$

This should give us some comfort: If we use Galerkin's method with exact eigenfunctions as basis vectors, we get the same solution as produced by the spectral method.

Of course, we are interested in cases where we do not know the eigenfunctions, and hence need to use more primitive choices for V_N and the basis vectors $\phi_1, \ldots, \phi_N$.

Lecture 16: Galerkin's Method with Hat Functions: The Finite Element Method

Last time we derived the weak form of the differential equation, and the Galerkin approximation, for $-u''(x) = f(x)$, $u(0) = u(1) = 0$.

Weak Form Find $u \in C_D^2[0,1]$ such that

$$a(u,v) = (f,v) \quad \text{for all } v \in C_D^2[0,1]$$

Galerkin Problem Find $u \in V_N$ such that

$$a(u,v) = (f,v) \quad \text{for all } v \in V_N$$

for $V_N \subseteq C_D^2[0,1]$.

Note: the Galerkin problem leads to the <u>best approximation</u> u_N to u, measured in the energy norm/inner product.

Energy inner product:

$$a(u,v) = \int_0^1 u(x)\, v(x)\, dx \qquad u, v \in C_D^2[0,1]$$

Energy norm:

$$\|u\|_E = \sqrt{a(u,u)}.$$

CONSIDER THE ERROR $u - u_N$.

DOES ANY OTHER $\hat{u} \in V_N$ GIVE A SMALLER ERROR?
LET $v \in V_N$ BE ANY TEST FUNCTION.

$$a(u - u_N, v) = a(u, v) - a(u_N, v)$$

$a(u, v) = (f, v)$ from weak form
$(v \in V_N \Rightarrow v \in C_0^2[0,1])$

$a(u_N, v) = (f, v)$ from Galerkin problem.

Hence the error $u - u_N$ is orthogonal to any $v \in V_N$: (in the energy inner product)

$$a(u - u_N, v) = (f, v) - (f, v) = 0.$$

Thus $u - u_N$ is orthogonal to V_N IN THE ENERGY INNER PRODUCT. By our PREVIOUS CHARACTERIZATION OF BEST APPROXIMATIONS: ORTHOGONALITY OF THE ERROR $\Rightarrow$ BEST APPROXIMATION.

THEOREM The approximation $u_N \in V_N$ that satisfies the Galerkin problem is the best approximation to the exact solution $u \in C_0^2[0,1]$ from V_N in the energy norm:

$$\|u - u_N\|_E = \min_{\hat{u} \in V_N} \|u - \hat{u}\|_E.$$

WE WILL NOW CONSIDER

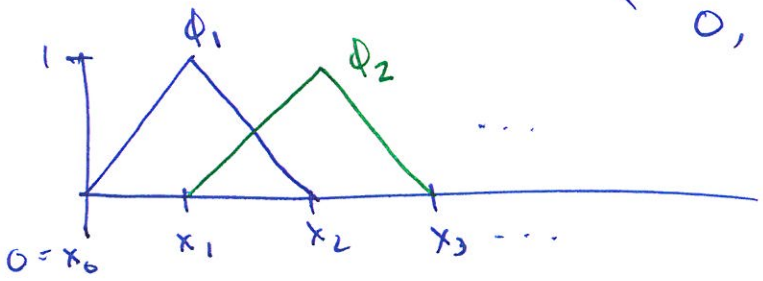
$$V_N = \text{span}\{\phi_1, \ldots, \phi_N\}$$

where $\phi_j(x)$ is the j^{th} hat function on the grid

$$x_j = jh, \quad h = \frac{1}{N+1}, \quad j = 0, \ldots, N+1$$

given by

$$\phi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{h}, & x \in [x_{j-1}, x_j]; \\ \dfrac{x_{j+1} - x}{h}, & x \in [x_j, x_{j+1}]; \\ 0, & \text{otherwise.} \end{cases}$$



Are you worried that these basis functions do not satisfy $\phi_j \in C^2_0[0,1]$? Don't fret: the discontinuities in ϕ_j' at x_{j-1}, x_j, x_{j+1} can be massaged away with more theory.

[GALERKIN + HAT FUNCTIONS $\Longrightarrow$ "FINITE ELEMENTS"]

BECAUSE THE HAT FUNCTIONS ARE MOSTLY ZERO ON $x \in [0,1]$; THEY ARE ONLY NONZERO ("SUPPORTED") ON $[x_{j-1}, x_{j+1}]$. WE SAY THAT HAT FUNCTIONS HAVE "SMALL SUPPORT."

TO COMPUTE THE FINITE ELEMENT SOLUTION u_N, WE MUST FORM THE STIFFNESS MATRIX K,

$$K(j,k) = a(\phi_j, \phi_k).$$

Note:

$$\phi_j'(x) = \begin{cases} 1/h, & x \in (x_{j-1}, x_j); \\ -1/h, & x \in (x_j, x_{j+1}); \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$K(j,j) = \int_0^1 \left(\phi_j'(x)\right)^2 dx = \int_{x_{j-1}}^{x_j} \left(\frac{1}{h}\right)^2 dx + \int_{x_j}^{x_{j+1}} \left(\frac{-1}{h}\right)^2 dx$$

$$= \left[\frac{x}{h^2}\right]_{x_{j-1}}^{x_j} + \left[\frac{x}{h^2}\right]_{x_j}^{x_{j+1}} = \frac{h}{h^2} + \frac{h}{h^2} = \frac{2}{h}.$$

$$K(j, j+1) = \int_0^1 \phi_j'(x)\, \phi_{j+1}'(x)\, dx$$

$$= \int_{x_j}^{x_{j+1}} \left(-\frac{1}{h}\right)\left(\frac{1}{h}\right) dx = \left[\frac{-x}{h^2}\right]_{x_j}^{x_{j+1}} = -\frac{1}{h}.$$

↑

Note: $\phi_j'(x)\, \phi_{j+1}'(x) \neq 0$ only when $x \in (x_j, x_{j+1})$.

If $|j-k| > 1$, then $\phi_j'(x)\, \phi_k'(x) = 0 \;\; \forall\, x \in [0,1]$, so

$$K(j,k) = 0.$$

In summary:

$$K(j,k) = \begin{cases} 2/h, & j = k; \\ -1/h, & |j-k| = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$K = \frac{1}{h}\begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{N\times N} \qquad (\text{"stiffness matrix"})$$

for finite elements.

To find u_N,

- Construct K
- Compute the load vector components via

$$f(j) = \int_0^1 f(x)\,\phi_j(x)\,dx = \int_{x_{j-1}}^{x_{j+1}} f(x)\,\phi_j(x)\,dx.$$

- Solve $Kc = f$ for c

  (e.g., $c = K \backslash f$ in MATLAB)

- Build u_N from the coefficients in c:

$$u_N(x) = c_1\,\phi_1(x) + \cdots + c_N\,\phi_N(x).$$

Interesting note: the c_k value gives $u_N(x_k)$, the approximate solution evaluated at the grid point x_k, $k = 1, \ldots, N$, since

$$u_N(x_k) = \sum_{j=1}^{N} c_j \underbrace{\phi_j(x_k)}_{= \begin{cases} 1 \text{ if } j=k \\ 0 \text{ if } j\neq k \end{cases}} = c_k \cdot 1 = c_k.$$

Lecture 17: Neumann Boundary Conditions: Spectral Method

Goal: Solve $-u''(x) = f(x) \quad x \in [0,1]$

$u'(0) = u'(1) = 0$

$Lu = f, \quad L: C_N^2[0,1] \to C[0,1], \quad Lu = -u''$

$C_N^2[0,1] = \{ u \in C^2[0,1] : u'(0) = u'(1) = 0 \}$

Step 1 Check if L is symmetric.

Let $u, v \in C_N^2[0,1]$.

$$(Lu, v) = \int_0^1 -u''(x) v(x)\, dx = \underbrace{\left[-u'(x) v(x)\right]_0^1}_{0 \text{ since } u'(0)=u'(1)=0} + \int_0^1 u'(x) v'(x)\, dx$$

$$= \underbrace{\left[u(x) v'(x)\right]_0^1}_{0 \text{ since } v'(0)=v'(1)=0.} - \int_0^1 u(x) v''(x)\, dx = (u, Lv)$$

Thus L is symmetric $\Rightarrow$ Eigenvalues of L are real

Eigenfunctions are orthogonal.

Step 2 Compute eigenvalues and eigenfunctions.

Solve $-\psi''(x) = \lambda \psi(x), \quad \psi \in C_N^2[0,1]$

$\Rightarrow \psi(x) = A \sin(\sqrt{\lambda} x) + B \cos(\sqrt{\lambda} x).$

$\psi'(x) = \sqrt{\lambda} A \cos(\sqrt{\lambda} x) - \sqrt{\lambda} B \sin(\sqrt{\lambda} x)$

$\psi'(0) = 0 \Rightarrow \sqrt{\lambda} A \cos(0) - \sqrt{\lambda} B \sin(0) = 0$

$\Rightarrow \sqrt{\lambda} A = 0 \Rightarrow \begin{cases} \lambda = 0 \\ \text{or } A = 0. \end{cases}$

$A=0$, $\psi'(1)=0 \Rightarrow \quad -\sqrt{\lambda}\, B \sin(\sqrt{\lambda}\cdot 1) = 0$

$$\Rightarrow \begin{cases} \lambda = 0 \\ \text{or} \\ B = 0 \\ \text{or} \\ \sin(\sqrt{\lambda}) = 0 . \end{cases}$$

If $\lambda = 0$, then $\psi(x) = B\cos(0x) = \underline{1}$

This is a nontrivial function, so $\boxed{\lambda_0 = 0, \ \psi_0(x) = 1}$ is an eigenfunction.

If $B = 0$, $\psi(x) = 0$

This is a trivial solution, not an eigenfunction.

If $\lambda \neq 0$, $\sin(\sqrt{\lambda}) = 0$, we have

$$\sqrt{\lambda} = n\pi \quad \text{for} \quad n = 1, 2, 3, \ldots$$

This gives eigenvalues and eigenfunctions $\psi(x) = B\cos(\sqrt{\lambda}x)$

$$\lambda_n = n^2\pi^2, \qquad \psi_n(x) = \sqrt{2}\cos(n\pi x)$$

↑

This is a convenient normalization,

$\Rightarrow (\psi_n, \psi_n) = 1$.

<u>STEP 3</u> Solve $Lu = f$ via the spectral method:

This suggests

$$u = \sum_{n=0}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)$$

↑

sum should run over all eigenvalues/eigenfunctions.

<u>BUT</u> This could give $\frac{1}{0}$ when $n = 0$....

CONSIDER AN EXAMPLE.

Solve $-u''(x) = f(x)$, $u'(0) = u'(1) = 0$

with $f(x) = x$.

$$-u''(x) = x \Rightarrow u''(x) = -x$$

$$\Rightarrow u'(x) = -\frac{x^2}{2} + c$$

$$\Rightarrow u(x) = -\frac{x^3}{6} + cx + d.$$

Find c, d to satisfy the boundary conditions.

$$0 = u'(0) = -\frac{0^2}{2} + c = c \Rightarrow c = 0.$$

$$0 = u'(1) = -\frac{1^2}{2} + c = -\frac{1}{2} \Rightarrow \text{Contradiction.}$$

No function $u(x)$ with $u'(0) = u'(1) = 0$ SATISFIES $-u''(x) = x$.

Now try $f(x) = x - 1/2$.

$$-u''(x) = x - \frac{1}{2} \Rightarrow u'' = -x + \frac{1}{2}$$

$$\Rightarrow u' = -\frac{x^2}{2} + \frac{x}{2} + c$$

$$\Rightarrow u = -\frac{x^3}{6} + \frac{x^2}{4} + cx + d$$

$$0 = u'(0) = -\frac{0^2}{2} + \frac{0}{2} + c \Rightarrow c = 0$$

$$0 = u'(1) = -\frac{1^2}{2} + \frac{1}{2} + 0 = 0 \Rightarrow \text{SATISFIED FOR ANY CHOICE OF } d.$$

Thus $-u''(x) = x - \frac{1}{2}$ has infinitely many solutions that satisfy $-u''(x) = x - \frac{1}{2}$ and $u'(0) = u'(1) = 0$:

$$u(x) = -\frac{x^3}{6} + \frac{x^2}{4} + d \qquad \text{for all constants } d.$$

Return to the spectral method.

What if the $n=0$ term $\frac{1}{\lambda_0} \frac{(f, \psi_0)}{(\psi_0, \psi_0)} \psi_0(x)$ gives a $\frac{0}{0}$ form, i.e., what if $(f, \psi_0) = 0$?

If we ignore the $n=0$ term, we could have

$$\hat{u}(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)$$

↑ (Notation reflects that we don't know (yet) if this is a solution of the DIFF eq.)

Then
$$L\hat{u} = L \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)$$

$$= \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} L\psi_n(x), \qquad L\psi_n = \lambda_n \psi_n$$

$$= \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \lambda_n \psi_n(x)$$

$$= \sum_{n=1}^{\infty} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x)$$

$$= \sum_{n=0}^{\infty} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x) \qquad \text{Since } (f, \psi_0) = 0 \text{ we can include this in the sum}$$

$$= f \qquad \text{(eigenfunctions give a basis for } C[0,1]\text{)}$$

So $\hat{u} = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n$

is a solution if $(f, \psi_0) = 0$.

In this case,

$$u(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} \psi_n(x) + d \cdot \psi_0(x)$$

is also a solution, for any constant d!

$$Lu(x) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \frac{(f, \psi_n)}{(\psi_n, \psi_n)} L\psi_n(x) + d \underbrace{L\psi_0(x)}_{=0}$$

Since $L\psi_0 = 0\psi_0$, with $\lambda_0 = 0$.

$$= f(x).$$

Notice: $(x - \frac{1}{2}, \psi_0) = \int_0^1 (x - \frac{1}{2}) \cdot 1 \, dx$

$$= \int_0^1 x - \frac{1}{2} \, dx = 0$$

This explains why we could solve

$$-u''(x) = x - \frac{1}{2}, \quad u'(0) = u'(1) = 0.$$

Notice $(x, \psi_0) = \int_0^1 x \, dx = \frac{1}{2} \neq 0$

This explains why we could <u>not</u> solve $-u''(x) = x$ with $u'(0) = u'(1) = 0$.

The situation perfectly parallels the case for matrix equations. If

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Then $A\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}2 & -1 & -1\\ -1 & 1 & 0\\ -1 & 0 & 1\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$

A has a zero eigenvalue with eigenvector $V_0 = \begin{bmatrix}1\\1\\1\end{bmatrix}$

$Ax = b$ has:

- no solution if $b^T V_0 \neq 0$
- Infinitely many solutions of the form $X + d\cdot\begin{bmatrix}1\\1\\1\end{bmatrix}$ if $b^T V_0 = 0$.

Lecture 18: Neumann boundary conditions: finite element method

Now we solve (or attempt to solve....)

$$-u''(x) = f(x) \qquad u'(0) = u'(1) = 0$$

using the finite element / Galerkin method.

STEP 1 DERIVE THE WEAK FORM

Let $u \in C_N^2[0,1] = \{u \in C^2[0,1] : u'(0) = u'(1) = 0\}$

Let $v \in C^2[0,1]$ be a "test function"

(↑ NOTE: $C^2[0,1]$, not $C_N^2[0,1]$)

$$Lu = f \Rightarrow (Lu, v) = (f, v)$$

$$\Rightarrow \int_0^1 -u''(x)\, v(x)\, dx = \int_0^1 f(x) v(x)\, dx$$

$$\text{I.B.P.} \Rightarrow \underbrace{\left[-u'(x) v(x)\right]_{x=0}^{x=1}}_{=0 \text{ since } u'(0)=u'(1)=0} + \int_0^1 u'(x) v'(x)\, dx = \int_0^1 f(x) v(x)\, dx$$

NO BOUNDARY CONDITIONS ARE NEEDED FOR v !

$$\Rightarrow \int_0^1 u'(x) v'(x)\, dx = \int_0^1 f(x) v(x)\, dx$$

$$\Rightarrow \boxed{a(u,v) = (f, v) \quad \text{for all } v \in C^2[0,1]}$$

WEAK FORM OF THE DIFF EQ.

SINCE WE DO NOT IMPOSE THE BOUNDARY CONDITIONS ON THE TEST FUNCTIONS, NEUMANN B.C.'S ARE CALLED "NATURAL BOUNDARY CONDITIONS"

IN CONTRAST TO DIRICHLET B.C.'S, WHICH ARE CALLED "ESSENTIAL BOUNDARY CONDITIONS."

STEP 2 GALERKIN APPROXIMATION.

IMPOSE THE WEAK FORM ONLY ON A FINITE DIMENSIONAL SUBSPACE $V_N \subset C^2[0,1]$.

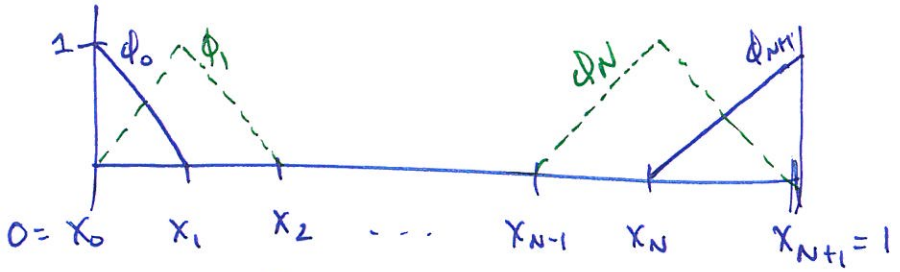
GALERKIN PROBLEM:

> Find $u_N \in V_N$ such that
> $$a(u_N, v) = (f, v) \quad \text{for all } v \in V_N.$$

STEP 3 LET V_N BE THE SET OF HAT FUNCTIONS

$$V_N = \text{span}\{\phi_0, \phi_1, \ldots, \phi_N, \phi_{N+1}\}$$

NOTE THAT WE NOW INCLUDE THESE "HALF-HAT" FUNCTIONS ON THE BOUNDARY



$$\left(x_j = jh, \quad h = \frac{1}{N+1} \right)$$

[IF WE DON'T INCLUDE ϕ_0, ϕ_{N+1}, WE WOULD BE IMPLICITLY FORCING DIRICHLET CONDITIONS, SINCE

$$\phi_1(0) = \cdots = \phi_N(0) = 0$$
$$\phi_1(1) = \cdots = \phi_N(1) = 0.$$]

STEP 4 SET UP STIFFNESS MATRIX, LOAD VECTOR, TRY TO SOLVE.

WRITE $u_N(x) = c_0\phi_0(x) + \cdots + c_{N+1}\phi_{N+1}(x)$.

GALERKIN PROBLEM ON V_N

$$\Leftrightarrow \quad a(u_N, \phi_j) = (f, \phi_j) \qquad j = 0, \ldots, N+1$$

$$\Leftrightarrow \quad \begin{bmatrix} a(\phi_0,\phi_0) & \cdots & a(\phi_0,\phi_{N+1}) \\ \vdots & & \\ a(\phi_{N+1},\phi_0) & \cdots & a(\phi_{N+1},\phi_{N+1}) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_N \\ c_{N+1} \end{bmatrix} = \begin{bmatrix} (f,\phi_0) \\ (f,\phi_1) \\ \vdots \\ (f,\phi_N) \\ (f,\phi_{N+1}) \end{bmatrix}$$

Stiffness matrix — load vector

$$K c = f.$$

Note: For $j,k \in \{1,\dots,N\}$, $a(\phi_j,\phi_k)$ is the same as for the Dirichlet case:

$$a(\phi_j,\phi_k) = \begin{cases} \frac{2}{h} & j=k \in \{1,\dots,N\} \\ -\frac{1}{h} & |j-k|=1,\ j,k\in\{1,\dots,N\} \\ 0 & \text{otherwise},\ j,k\in\{1,\dots,N\}. \end{cases}$$

What about ϕ_0, ϕ_{N+1}?

$$a(\phi_0,\phi_0) = \int_0^h (\phi_0'(x))^2\,dx = \int_0^h \left(-\tfrac{1}{h}\right)^2 dx$$

$$= \frac{1}{h}$$

Similarly

$$a(\phi_{N+1},\phi_{N+1}) = \frac{1}{h}.$$

The other terms are the same:

$$a(\phi_0,\phi_1) = a(\phi_1,\phi_0) = a(\phi_N,\phi_{N+1}) = a(\phi_{N+1},\phi_N) = -\frac{1}{h}$$

$$a(\phi_0,\phi_k) = 0 \quad \text{if} \quad k>1$$

$$a(\phi_{N+1},\phi_k) = 0 \quad \text{if} \quad k<N.$$

Work out K for $N=2$.

$$K = \frac{1}{h}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

Notice $K\begin{bmatrix}1\\1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\0\\0\\0\end{bmatrix} \Rightarrow \quad V_0 = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}$

K has a zero eigenvalue: It is not invertible!

What function in $C[0,1]$ does this eigenvector $\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$ correspond to?

$$1\cdot\phi_0(x) + 1\cdot\phi_1(x) + 1\cdot\phi_2(x) + \cdots + 1\cdot\phi_{N+1}(x)$$
$$= \sum_{j=0}^{N+1} \phi_j(x) = 1 \quad \text{for } x \in [0,1].$$

This is exactly the eigenfunction $\psi_0(x)$ from the last lecture!

Note that $Kc=f$ will have a solution only when $f^T V_0 = 0$

$$f^T V_0 = \sum_{j=0}^{N+1} (f_i \phi_j)\cdot 1 = \left(f_i \sum_{j=0}^{N+1} \phi_j\right) = (f_i\, 1)$$
$$= (f_i \psi_0)$$

Thus $Kc = f$ has a solution in the exact same case that $-u'' = f$, $u'(0) = u'(1) = 0$ has a solution!

Similarly, if $f^T V_0 = 0$, then $Kc = f$ will have <u>infinitely many</u> solutions of the form

$$c + d \cdot V_0 \qquad \text{for } d \text{ any constant}$$

which adds $d \cdot V_0 \sim d \cdot \Psi_0$ to any solution.

See fem_neumann.m on the website.

Note: the "natural" boundary conditions are <u>not</u> imposed exactly, they emerge "naturally" — and so are only approximated: as $N \to \infty$, the approximation improves, and $U_N'(0), U_N'(1) \approx 0$.

∎

CMDA 4604: Intermediate Topics in Mathematical Modeling
Lecture 19: Interpolation and Quadrature

In this lecture we make a brief diversion into the areas of *interpolation* and *quadrature.*

Given a function $f \in C[a,b]$, we say that a polynomial p *interpolates* f at the point $\widehat{x} \in [a,b]$ if

$$f(\widehat{x}) = p(\widehat{x}).$$

In the context of today's lecture, we aim to use interpolation as a way to construct good polynomial approximations to f. The next lecture we will put today's results into the broader context of the course: we will show how the approximate solutions constructed by the finite element method can be related to interpolating polynomials, and so the accuracy of interpolating polynomials will lead to error bounds for the finite element method.

We have three goals today: (1) construct interpolating polynomials p_n; (2) bound the error $f(x) - p_n(x)$; (3) integrate interpolating polynomials to approximate the integral of f

1. Constructing Interpolating Polynomials.

We seek to solve the following problem:

Polynomial interpolation problem. Given a function $f \in C[a,b]$ and points $x_0, \ldots, x_n \in [a,b]$, construct a polynomial p_n of degree not exceeding n such that

$$p_n(x_j) = f(x_j), \qquad j = 0, \ldots, n.$$

In any numerical analysis course, one learns that a unique solution p_n to this problem always exists, and be constructed in various ways. Here we shall just detail one elegant way to develop the interpolant, called the *Lagrange form.*

The idea behind the Lagrange form is simple. Consider the functions

$$L_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}.$$

Note that each L_k is a degree-n polynomial. (For each value of k, the product contains n terms of the form $(x - x_j)/(x_k - x_j)$. Each of these terms is a degree-1 polynomial. The product of n degree-1 polynomials is a degree-n polynomial.) Moreover, these polynomials have a very special property: by construction, L_k takes the value 1 at x_k and has roots at each of the points x_j, $j \neq k$:

$$L_k(x_j) = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases}$$

These $n+1$ polynomials $L_0, \ldots, L_n$ form a basis for the $n+1$ dimensional vector space of polynomials having degree n or less.

These *Lagrange basis functions* make it trivial to construct the solution p_n to the polynomial interpolation function:

$$p_n(x) = \prod_{k=0}^{n} f(x_k) L_k(x).$$

Since p_n is the sum of degree-n polynomials, it too is a degree-n polynomial. The property that $L_k(x_i) = 0$ for $i \neq k$ ensures that at the interpolation point x_j,

$$p_n(x_j) = \prod_{k=0}^{n} f(x_k) L_k(x_j) = f(x_j) L_j(x_j) = f(x_j).$$

Thus the polynomial p_n passes through f at the designated points. But how accurately does p_n approximate f at the other points in the interval $[a, b]$ where we have not specified the interpolation condition?

2. Interpolation Error Analysis.

We now seek to characterize the maximum error

$$\max_{x \in [a,b]} |f(x) - p_n(x)|.$$

The characterization of this error is one of the most fundamental results in numerical analysis.

Theorem (Interpolation Error Bound). Suppose $f \in C^{n+1}[a, b]$ and let $p_n \in \mathcal{P}_n$ denote the polynomial that interpolates f at the points $x_0, \ldots, x_n \in [a, b]$ for $j = 0, \ldots, n$. The for every $x \in [a, b]$ there exists $\xi \in [a, b]$ such that

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (x - x_j).$$

This result yields a bound for the worst error over the interval $[a, b]$:

$$\max_{x \in [a,b]} |f(x) - p_n(x)| \leq \left(\max_{\xi \in [a,b]} \frac{|f^{(n+1)}(\xi)|}{(n+1)!} \right) \left(\max_{x \in [a,b]} \prod_{j=0}^{n} |x - x_j| \right). \tag{1}$$

Proof. Consider some arbitrary point $\widehat{x} \in [a, b]$. We seek a descriptive expression for the error $f(\widehat{x}) - p_n(\widehat{x})$. If $\widehat{x} = x_j$ for some $j \in \{0, \ldots, n\}$, then $f(\widehat{x}) - p_n(\widehat{x}) = 0$ and there is nothing to prove. Thus, suppose for the rest of the proof that $\widehat{x}$ is not one of the interpolation points.

To describe $f(\widehat{x}) - p_n(\widehat{x})$, we shall build the polynomial of degree $n + 1$ that interpolates f at $x_0, \ldots, x_n$, and also $\widehat{x}$. Of course, this polynomial will give zero error at $\widehat{x}$, since it interpolates f there. From this polynomial we can extract a formula for $f(\widehat{x}) - p_n(\widehat{x})$ by measuring how much the degree $n + 1$ interpolant improves upon the degree-n interpolant p_n at $\widehat{x}$.

Since we wish to understand the relationship of the degree $n + 1$ interpolant to p_n, we shall write that degree $n + 1$ interpolant in a manner that explicitly incorporates p_n. Given this setting, use of the Newton form of the interpolant is natural; i.e., we write the degree $n + 1$ polynomial as

$$p_n(x) + \gamma \prod_{j=0}^{n} (x - x_j)$$

for some constant γ chosen to make the interpolant exact at $\widehat{x}$. For convenience, we write

$$w(x) \equiv \prod_{j=0}^{n} (x - x_j)$$

and then denote the error of this degree $n+1$ interpolant by

$$\phi(x) \equiv f(x) - (p_n(x) + \gamma w(x)).$$

To make the polynomial $p_n(x) + \gamma w(x)$ interpolate f at $\widehat{x}$, we shall pick γ such that $\phi(\widehat{x}) = 0$. The fact that $\widehat{x} \notin \{x_j\}_{j=0}^n$ ensures that $w(\widehat{x}) \neq 0$, and so we can force $\phi(\widehat{x}) = 0$ by setting

$$\gamma = \frac{f(\widehat{x}) - p_n(\widehat{x})}{w(\widehat{x})}.$$

Furthermore, since $f(x_j) = p_n(x_j)$ and $w(x_j) = 0$ at all the $n+1$ interpolation points $x_0, \ldots, x_n$, we also have $\phi(x_j) = f(x_j) - p_n(x_j) - \gamma w(x_j) = 0$. Thus, ϕ is a function with at least $n+2$ zeros in the interval $[a, b]$. Rolle's Theorem[1] tells us that between every two consecutive zeros of ϕ, there is some zero of ϕ'. Since ϕ has at least $n+2$ zeros in $[a, b]$, ϕ' has at least $n+1$ zeros in this same interval. We can repeat this argument with ϕ' to see that ϕ'' must have at least n zeros in $[a, b]$. Continuing in this manner with higher derivatives, we eventually conclude that $\phi^{(n+1)}$ must have at least one zero in $[a, b]$; we denote this zero as ξ, so that $\phi^{(n+1)}(\xi) = 0$.

We now want a more concrete expression for $\phi^{(n+1)}$. Note that

$$\phi^{(n+1)}(x) = f^{(n+1)}(x) - p_n^{(n+1)}(x) - \gamma w^{(n+1)}(x).$$

Since p_n is a polynomial of degree n or less, $p_n^{(n+1)} \equiv 0$. Now observe that w is a polynomial of degree $n+1$. We could write out all the coefficients of this polynomial explicitly, but that is a bit tedious, and we do not need all of them. Simply observe that we can write $w(x) = x^{n+1} + q(x)$, for some $q \in \mathcal{P}_n$, and this polynomial q will vanish when we take $n+1$ derivatives:

$$w^{(n+1)}(x) = \left(\frac{d^{n+1}}{dx^{n+1}} x^{n+1} \right) + q^{(n+1)}(x) = (n+1)! + 0.$$

Assembling the pieces, $\phi^{(n+1)}(x) = f^{(n+1)}(x) - \gamma\,(n+1)!$. Since $\phi^{(n+1)}(\xi) = 0$, we conclude that

$$\gamma = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Substituting this expression into $0 = \phi(\widehat{x}) = f(\widehat{x}) - p_n(\widehat{x}) - \lambda w(\widehat{x})$, we obtain

$$f(\widehat{x}) - p_n(\widehat{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (\widehat{x} - x_j). \qquad \blacksquare$$

[1] Recall the Mean Value Theorem from calculus: Given $d > c$, suppose $f \in C[c, d]$ is differentiable on (c, d). Then there exists some $\eta \in (c, d)$ such that $(f(d) - f(c))/(d - c) = f'(\eta)$. Rolle's Theorem is a special case: If $f(d) = f(c)$, then there is some point $\eta \in (c, d)$ such that $f'(\eta) = 0$.

This error bound has strong parallels to the remainder term in Taylor's formula. Recall that for sufficiently smooth h, the Taylor expansion of f about the point x_0 is given by

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \cdots + \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + \frac{f^{(k+1)}(\xi)}{(k+1)!}(x - x_0)^k.$$

Ignoring the remainder term at the end, note that the Taylor expansion gives a polynomial model of f, but one based on local information about f and its derivatives, as opposed to the polynomial interpolant, which is based on global information, but only about f, not its derivatives. Rearranging this expression, we have

$$f(x) - \left(f(x_0) + (x - x_0)f'(x_0) + \cdots + \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) \right) = \frac{f^{(k+1)}(\xi)}{(k+1)!}(x - x_0)^k,$$

a perfect analogue of the interpolation error formula we have just proved.

3. Interpolatory Quadrature Formulas.

The finite element method requires computations like

$$(f, \phi_k) = \int_0^1 f(x)\phi_k(x) \; \mathrm{d}x$$

to construct the load vector. It may be inconvenient (or even impossible for some f) to compute this inner product. For such cases we wish to approximate the integral.

We shall consider the generic problem of approximating

$$I(f) = \int_a^b f(x)\,\mathrm{d}x.$$

Polynomial interpolation provides a simple way to approximate the integral:

- Construct the polynomial interpolant p_n to f at designated points;
- Approximate $\int_a^b f(x)\,\mathrm{d}x$ by $\int_a^b p_n(x)\,\mathrm{d}x$.

If we construct p_n using the Lagrange form described above, this procedure becomes very simple:

- Construct the interpolating polynomial

$$p_n(x) = \sum_{j=0}^{n} f(x_j)L_j(x);$$

- Integrate the interpolating polynomial to obtain $I_n(f)$, approximating the exact integral $I(f)$:

$$\begin{aligned} I_n(f) &= \int_a^b p_n(x)\,\mathrm{d}x \\ &= \int_a^b \sum_{j=0}^{n} f(x_j)L_j(x)\,\mathrm{d}x \\ &= \sum_{j=0}^{n} f(x_j) \int_a^b L_j(x)\,\mathrm{d}x. \end{aligned}$$

Notice that the integrals that remain depend on the Lagrange basis functions L_j *but not on* f. We will call these integrals the *weights* of the quadrature rule:

$$w_j = \int_a^b L_j(x)\,\mathrm{d}x.$$

Then the quadrature rule takes the simple form

$$I_n(f) = \sum_{j=0}^{n} w_j f(x_j).$$

The points $x_0, \ldots, x_n$ are called the *nodes* of the quadrature rule. When you choose evenly spaced points over $[a, b]$, you recover familiar rules that you have already encountered in calculus:

- $n = 0$ ($x_0 = (a+b)/2$, $L_0(x) = 1$) gives
$$\int_a^b f(x)\,\mathrm{d}x \approx (b-a) f(\tfrac{1}{2}(a+b));$$
- $n = 1$ ($x_0 = a$, $x_1 = b$) gives the trapezoid rule:
$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\Big(f(a) + f(b)\Big);$$
- $n = 2$ ($x_0 = a$, $x_1 = (a+b)/2$, $x_2 = b$) gives Simpson's rule:
$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\Big(f(a) + 4f(\tfrac{1}{2}(a+b)) + f(b)\Big).$$

The first rule approximates f with an interpolating constant; the trapezoid rule approximates f with an interpolating linear polynomial; Simpson's rule approximates f with an interpolating quadratic.

How does one quantify the error $I(f) - I_n(f)$? Simply integrate the error formula for polynomial interpolation! One must then calculate:

$$\begin{aligned} I(f) - I_n(f) &= \int_a^b f(x) - p_n(x)\,\mathrm{d}x \\ &= \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{j=0}^{n} (x - x_j)\,\mathrm{d}x. \end{aligned}$$

The error analysis for the trapezoid rule (where $x_0 = a$ and $x_1 = b$) follows from application of the mean value theorem for integrals:

$$\begin{aligned} \int_a^b f(x)\,\mathrm{d}x - \int_a^b p_1(x)\,\mathrm{d}x &= \int_a^b \tfrac{1}{2} f''(\xi(x))(x-a)(x-b)\,\mathrm{d}x \\ &= \tfrac{1}{2} f''(\eta) \int_a^b (x-a)(x-b)\,\mathrm{d}x \\ &= \tfrac{1}{2} f''(\eta)(\tfrac{1}{6}a^3 - \tfrac{1}{2}a^2 b + \tfrac{1}{2}ab^2 - \tfrac{1}{6}b^3) \\ &= -\tfrac{1}{12} f''(\eta)(b-a)^3 \end{aligned}$$

for some $\eta \in [a,b]$. As we expect, if $f(x)$ is a linear polynomial, then $f''(x) = 0$ for all x, and hence the trapezoid rule will be exact.

The analysis for Simpson's rule is a bit more complicated. One can actually show something stronger than what you might expect from integrating the polynomial interpolation error:

$$\int_a^b f(x)\,\mathrm{d}x - \int_a^b p_2(x)\,\mathrm{d}x = -\frac{1}{90}\,\frac{(b-a)^5}{2^5}\,f^{(4)}(\eta)$$

for some $\eta \in [a,b]$. Notice that this bound involves $f^{(4)}$, rather than the expected $f^{(3)}$: *Simpson's rule will be exact for cubic polynomials, not just quadratics!*

If you want greater accuracy than these bounds suggest, you could simply increase the degree n, and there are some settings in which this makes great sense: but one must be careful about how to select the nodes $x_0, \ldots, x_n$, and uniformly spaced points are not the best choice.

Composite rules. As an alternative to integrating a high-degree polynomial, one can pursue a simpler approach that is often very effective, especially for problems that are not particularly smooth (e.g., our hat functions): Break the interval $[a,b]$ into subintervals, and apply the trapezoid rule or Simpson's rule on each subinterval. Applying the trapezoid rule gives

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x \approx \sum_{j=1}^{n} \frac{(x_j - x_{j-1})}{2}\Big(f(x_{j-1} + f(x_j)\Big).$$

The standard implementation assumes that f is evaluated at uniformly spaced points between a and b, $x_j = a + jh$ for $j = 0, \ldots, n$ and $h = (b-a)/n$, giving the following famous formulation:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{2}\Big(f(a) + 2\sum_{j=1}^{n-1} f(a+jh) + f(b)\Big).$$

(Of course, one can readily adjust this rule to cope with irregularly spaced points.) The error in the composite trapezoid rule can be derived by summing up the error in each application of the trapezoid rule:

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x - \frac{h}{2}\Big(f(a) + 2\sum_{j=1}^{n-1} f(a+jh) + f(b)\Big) &= \sum_{j=1}^{n}\Big(-\tfrac{1}{12} f''(\eta_j)(x_j - x_{j-1})^3\Big) \\ &= -\frac{h^3}{12}\sum_{j=1}^{n} f''(\eta_j)\end{aligned}$$

for $\eta_j \in [x_{j-1}, x_j]$. We can simplify these f'' terms by noting that $\frac{1}{n}(\sum_{j=1}^{n} f''(\eta_j))$ is the average of n values of f'' evaluated at points in the interval $[a,b]$. Naturally, this average cannot exceed the maximum or minimum value that f'' assumes on $[a,b]$, so there exist points $\xi_1, \xi_2 \in [a,b]$ such that

$$f''(\xi_1) \le \frac{1}{n}\sum_{j=1}^{n} f''(\eta_j) \le f''(\xi_2).$$

Thus the intermediate value theorem guarantees the existence of some $\eta \in [a,b]$ such that

$$f''(\eta) = \frac{1}{n}\sum_{j=1}^{n} f''(\eta_j).$$

The composite trapezoid error bound thus simplifies to

$$\int_a^b f(x)\,\mathrm{d}x - \frac{h}{2}\Big(f(a) + 2\sum_{j=1}^{n-1} f(a+jh) + f(b)\Big) = -\frac{h^2}{12}(b-a)f''(\eta).$$

Similar analysis can be performed to derive the composite Simpson's rule. We now must ensure that n is even. since each interval on which we apply the standard Simpson's rule has width $2h$. Simple algebra leads to the formula

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}\Big(f(a) + 4\sum_{j=1}^{n/2} f(a+(2j-1)h) + 2\sum_{j=1}^{n/2-1} f(a+2jh) + f(b)\Big).$$

Derivation of the error formula for the composite Simpson's rule follows the same strategy as the analysis of the composite trapezoid rule. One obtains

$$\int_a^b f(x)\,\mathrm{d}x - \frac{h}{3}\Big(f(a) + 4\sum_{j=1}^{n/2} f(a+(2j-1)h) + 2\sum_{j=1}^{n/2-1} f(a+2jh) + f(b)\Big) = -\frac{h^4}{180}(b-a)f^{(4)}(\eta)$$

for some $\eta \in [a,b]$.