Models and Examples

Example 1: A Conservation Law; Transport Equation

Consider a 1-lane road, where the density $n = n(t,x)$ depends upon position x, time t. Let

$$f(t,x) \;:=\; \text{rate (in cars per unit time) at which cars are passing point } x \text{ at time } t.$$

Fix (t,x). Then

$$\begin{aligned} f(t,x)\Delta t &\;\approx\; \#\{\text{cars passing } x \text{ during time interval } [t, t+\Delta t]\}, \\ n(t,x)\Delta x &\;\approx\; \#\{\text{cars on stretch } [x, x+\Delta x] \text{ at time } t\}. \end{aligned}$$

So,

$$[n(t+\Delta t, x) - n(t,x)]\Delta x \approx -[f(t, x+\Delta x) - f(t,x)]\Delta t\ .$$

In the limit, get

$$\frac{\partial n}{\partial t} + \frac{\partial f}{\partial x} \;=\; 0\ ,$$

a **conservation law**. Suppose $f = f(n)$ (i.e., **flux** f depends only upon car density). Then,

$$\frac{\partial f(n)}{\partial x} \;=\; f'(n)\,\frac{\partial n}{\partial x}\ ,$$

and our conservation law becomes the **transport equation**

$$\frac{\partial n}{\partial t} + f'(n)\,\frac{\partial n}{\partial x} \;=\; 0\ . \tag{1}$$

Aspects of (1):

- 1st-order
- *linear* if $f'(n) \equiv c$ (constant); *nonlinear* if truly dependent upon n
- time-varying
- a *differential equation* because involves derivatives of unknown density $n(t,x)$
- *partial* because $n = n(t,x)$ relies on more than one independent variable
- can think of this PDE on a bounded or unbounded domain; need initial density profile $n(0,x)$ to solve IC, perhaps conditions at "boundary" of road

■

Example 2: Conservation Law in 3D/Heat Equation

Consider an open, simply-connected region Ω in $\mathbb{R}^3$ and an open ball $B \subset \Omega$ with boundary ∂B. Let

$\rho(t, \mathbf{x})$: density of some substance (amount per unit volume)

$\Rightarrow \quad \iiint_B \rho(t, \mathbf{x})\, dV$ gives total amount inside B

$\phi(t, \mathbf{x})$: flux vector

direction corresponds to flow at position $\mathbf{x}$, time t

magnitude has units amount per unit area per unit time

$\Rightarrow$ For $\mathbf{x} \in \partial\Omega$, $\phi(t, \mathbf{x}) \cdot \mathbf{n}$ gives outward ($\mathbf{n}$ oriented appropriately) flow rate at $\mathbf{x}$

$f(t, x)$: creation (when positive) rate, in amount per unit volume per unit time

$\Rightarrow \quad \iiint_B f(t, \mathbf{x})\, dV$ gives total amount created inside B

Assuming conservation, we have

$$\frac{d}{dt} \iiint_B \rho(t, \mathbf{x})\, dV \;=\; \iiint_B f(t, \mathbf{x})\, dV - \oiint_{\partial B} \phi(t, \mathbf{x}) \cdot \mathbf{n}\, d\sigma \,. \tag{2}$$

By the Divergence Theorem,

$$\oiint_{\partial B} \phi \cdot \mathbf{n}\, d\sigma \;=\; \iiint_B \nabla \cdot \phi\, dV \,.$$

Passing the time derivative through the integral on the left-hand side of (**??**), and accumulating everything into one triple integral, we have

$$\iiint_B (\rho_t + \nabla \cdot \phi - f)\, dV \;=\; 0 \,.$$

Since this *global* relation holds for all balls B inside Ω, we conclude the local (PDE) relationship

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \phi \;=\; f \,.$$

Many substances have a flux that is proportional to density, *flowing down the gradient*, as in

$$\phi(t, \mathbf{x}) = -c(\mathbf{x}) \nabla \rho(t, \mathbf{x}) \,.$$

(The coefficient c is often assumed to be constant.) Employing this assumption we get the **diffusion equation** (inhomogeneous if $f \not\equiv 0$)

$$\frac{\partial \rho}{\partial t} - \nabla \cdot [c(\mathbf{x}) \nabla \rho] \;=\; f \,. \tag{3}$$

Heat is one substance that is closely modeled by (3). Specifically, if $u(t, \mathbf{x})$ is the temperature in some medium, and γ is the (constant) **thermal diffusivity**, then we get the **heat equation**

$$\frac{\partial u}{\partial t} - \gamma \Delta u \;=\; f \,, \tag{4}$$

Aspects:

- linear, 2nd-order, homogeneous or not depending upon if $f \equiv 0$
- equilibrium version has $\partial u/\partial t = 0$ (u constant in time)
  - equilibrium leads to Poisson eqn.: $-\Delta u = f/\gamma$.
  - equilibrium and homogeneous leads to Laplace eqn.: $\Delta u = 0$.
- 4 independent variables, $\mathbf{x} = (x, y, z)$, t (so time-varying)
  - 1st-order in t, so will require 1 IC (general principle)
  - 2nd-order in x, y, z, so will require 2 BCs for each (general principle)
- lower-dimensional versions of heat equation

■

Example 3: Laplace's Equation in 2D

Consider a 2D N-by-N mesh with horizontal/vertical spacing h. At grid point $(x_i, y_j) = (ih, jh)$ stands a person with (scalar) opinion p_{ij}, one that may be shared only with immediate neighbors.

Wishing to *minimize conflict*, each person is willing to take an opinion which is the average of his neighbors:

$$p_{ij} \;=\; \frac{1}{4}(p_{i+1,j} + p_{i-1,j} + p_{i,j+1} + p_{i,j-1}) \;,$$

or

$$p(x,y) \;=\; \frac{1}{4}[p(x+h,y) + p(x-h,y) + p(x,y+h) + p(x,y-h)] \;,$$

which may be rearranged as

$$\frac{p(x-h,y) - 2p(x,y) + p(x+h,y)}{h^2} + \frac{p(x,y-h) - 2p(x,y) + p(x,y+h)}{h^2} \;=\; 0 \;.$$

Letting $h \to 0$, we get **Laplace's equation**

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} \;=\; 0 \;.$$

Aspects:

- 2nd-order, linear, homogeneous
- equilibrium (i.e., not varying with time)
- Difference equation cannot apply on boundary of mesh. Need **boundary conditions** to solve (both difference equation and PDE).
- 'del' and Δ notation for Laplace's equation
- 3D version
- Poisson's equation (nonhomogeneous version of Laplace's)

■

Example 4: Korteweg-de Vries equation

The KdV equation is

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 .$$

Ask

- linear or NL?
- order?

Note: some authors may call this homogeneous, but homogeneity is important mainly when equation is *linear*.

■

Example 5: Wave Equation in 1D

Let

$u = u(t,x)$: displacement from equilibrium at position x, time t
$\rho = \rho(t,x)$: density of string (mass/length)
$\rho_0 = \rho_0(x)$: density at x when string in equilibrium
$T = T(t,x)$: tension (force) right of x exerts on left (assume directed along tangent)
$\theta = \theta(t,x)$: angle with horizontal

Assumptions:

(i) displacements $u(t,x)$ are small

(ii) no displacement other than transversal

For each a, b, horizontal tension components are equal:

$$T(t,b)\cos(\theta(t,b)) - T(t,a)\cos(\theta(t,a)) = 0 \quad \Rightarrow \quad T(t,x)\cos(\theta(t,x)) = \tau(t) \qquad \text{(independent of } x\text{)}$$

Conservation of mass:

Let a, b be arbitrary along x-axis between endpoints

$$\text{mass between } x = a, x = b \;=\; \int_a^b \rho(t,x)\sqrt{1 + u_x(t,x)^2}\,dx \;=\; \int_a^b \rho_0(x)\,dx$$

For last equality have used

- assumption (ii) above
- conservation of mass

a, b arbitrary $\quad\Rightarrow\quad \rho(t,x)\sqrt{1+u_x(t,x)^2} \;=\; \rho_0(x)$

Newton's 2nd law:

- "time rate of change of momentum equals sum of external forces"
- applied to vertical direction:

$$\text{Total momentum} \;=\; \int_a^b u_t(t,x)\rho(t,x)\sqrt{1+u_x(t,x)^2}\,dx \;=\; \int_a^b u_t(t,x)\rho_0(x)\,dx,$$

which implies

$$\begin{aligned}
&\frac{d}{dt}(\text{total momentum between } a, b)\\
&= T(t,b)\sin(\theta(t,b)) - T(t,a)\sin(\theta(t,a)) - g\int_a^b \rho_0(x)\,dx \qquad (g \text{ is gravity})\\
&= T(t,b)\cos(\theta(t,b))\tan(\theta(t,b)) - T(t,a)\cos(\theta(t,a))\tan(\theta(t,a)) - g\int_a^b \rho_0(x)\,dx\\
&= \tau(t)[u_x(t,b) - u_x(t,a)] - g\int_a^b \rho_0(x)\,dx \qquad (\text{Note: } u_x(t,x) = \tan(\theta(t,x))\\
&= \int_a^b [u_{xx}(t,x)\tau(t) - g\rho_0(x)]\,dx\,.
\end{aligned}$$

Under suitable assumptions, pass derivative (in t) through momentum integral (in x):

$$\frac{d}{dt}\int_a^b u_t(t,x)\rho_0(x)\,dx = \int_a^b u_{tt}(t,x)\rho_0(x)\,dx$$

Global Law:

We have

$$\int_a^b u_{tt}(t,x)\rho_0(x)\,dx \;=\; \int_a^b [u_{xx}(t,x)\tau(t) - g\rho_0(x)]\,dx \qquad \text{(global law)}$$

$$a, b \text{ arbitrary} \quad\Rightarrow\quad \rho_0(x)(u_{tt}+g) \;=\; \tau(t)u_{xx} \qquad \text{(local law).}$$

Possible additional assumption: gravity is not a factor. Then we have

$$\rho_0(x)u_{tt} \;=\; \tau(t)u_{xx}\,.$$

Employing "small displacements" assumption (ii) above, which suggests $\tau(t) \approx \tau_0$ (a constant), we get the **one-dimensional (in space) wave equation**

$$u_{tt} \;=\; c_0^2(x)u_{xx}\,, \tag{5}$$

where $c_0^2(x) := \tau_0/\rho_0(x) > 0$.

■

Example 6: General n^{th}-order PDE

General n^{th}-order PDE in 2 independent variables (t, x) has form

$$F\left(\frac{\partial^n u}{\partial t^n}, \frac{\partial^n u}{\partial t^{n-1}\partial x}, \ldots, \frac{\partial^n u}{\partial x^n}, \frac{\partial^{n-1} u}{\partial t^{n-1}}, \frac{\partial^{n-1} u}{\partial t^{n-2}\partial x}, \ldots, \frac{\partial^{n-1} u}{\partial x^{n-1}}, \ldots, \frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}, u\right) = 0 .$$

If linear in the highest-order derivatives (i.e., coefficients of said derivatives rely only on ind. vars. and lower-order derivatives of u), this PDE is said to be **quasi-linear**.

If the PDE is (fully) **linear** (specializing to the 2nd-order case), it takes the form

$$a_{2,0}(t,x)\frac{\partial^2 u}{\partial t^2} + a_{1,1}(t,x)\frac{\partial^2 u}{\partial t \partial x} + a_{0,2}(t,x)\frac{\partial^2 u}{\partial x^2} + a_{1,0}(t,x)\frac{\partial u}{\partial t} + a_{0,1}(t,x)\frac{\partial u}{\partial x} + a_{0,0}(t,x)u = f(t,x) ,$$

or, written in **operator form**,

$$\begin{aligned} L[u](t,x) &= f(t,x) , \\ L &= a_{2,0}\frac{\partial^2}{\partial t^2} + a_{1,1}\frac{\partial^2}{\partial t \partial x} + a_{0,2}\frac{\partial^2}{\partial x^2} + a_{1,0}\frac{\partial}{\partial t} + a_{0,1}\frac{\partial}{\partial x} + a_{0,0} , \end{aligned} \tag{6}$$

where L, called a **linear (2nd-order) differential operator**, acts on suitably *smooth* functions $u(t, x)$. The PDE (6) is **homogeneous** if $f \equiv 0$, **nonhomogeneous** otherwise.

■

Important distinctions in DEs:

- order
- ordinary vs. partial
- time-dependent vs. equilibrium/steady-state
- one vs. multiple spatial dimensions
- linear vs. NL
  - linear may always be written in operator form $L[u] = f$
  - properties which make L a *linear* operator
  - superposition
- homogeneous vs. inhomogeneous: for homogeneous (linear) problems $L[u] = 0$:
  - solutions form subspace of some vector space consisting of functions

  - this subspace called **kernel of** L (like nullspace of a matrix)

- single vs. system
  - System Example: Navier-Stokes equations (1.4), p. 3
  - systems not explored in text

- real vs. complex
  - Example ((1.9) in text) of complex: Schrödinger's equation

$$i\hbar\frac{\partial u}{\partial t} = -\frac{\partial^2 u}{\partial x^2} + V(x)u$$

Nondimensionalization (Incomplete; Skipped)

We seek to

- reduce number of parameters which determine solution of problem
- properly *scale* variables so that
  - relative magnitude of various terms is revealed
  - perturbation methods become available

Cf. Lin and Segel, *Mathematics Applied to Deterministic Problems in the Natural Sciences*.

Example 7:

Consider our general 1D wave equation, including gravitational effects, on a bounded interval

$$\left.\begin{array}{l} \rho_0(x)(u_{tt}+g) = \tau(t)u_{xx}\,, \qquad x\in[0,\ell],\ \ t>0\,, \\ \text{subject to}\quad u(t,0)=0,\ \ u(t,\ell)=0,\ \ u(0,x)=f(x),\ \ u_t(0,x)=g(x)\,. \end{array}\right\} \tag{7}$$

Assume $\rho_0(\cdot)$, $\tau(\cdot)$ are constant.

Goal: *Assess negligibility of gravity term*.

Procedure:

1. List all parameters and variables, with their dimensions.
   Here,

name	units	name	units
u	length	ρ	$(\text{mass})(\text{length})^{-1}$
τ	$(\text{mass})(\text{length})(\text{time})^{-2}$	g	$(\text{length})(\text{time})^{-2}$
x	length	t	time
ℓ	length	y_0	length (maximum initial displacement of string)

2. Find *dimensionless* combinations of parameters/variables

   Here, we have

$$\frac{u}{y_0} \le 1 \quad \Rightarrow \quad \text{set } \bar{u} := \frac{u}{y_0}\,, \qquad \text{and} \qquad \frac{x}{\ell} \le 1 \quad \Rightarrow \quad \text{set } \bar{x} := \frac{x}{\ell}\,.$$

   Look for combination of form $\tau^a g^b \rho^c \ell^d t$, yielding units

$$(\text{time})^{1-2a-2b}(\text{mass})^{a+d}(\text{length})^{a+b+d-c} \quad \Rightarrow \quad \left\{ \begin{array}{l} a+c=0 \\ 2a+2b=1 \\ a+b+d-c=0 \end{array} \right\} \quad \Rightarrow \quad \left\{ \begin{array}{l} c=-a \\ b=-a+\frac{1}{2} \\ d=-a-\frac{1}{2} \end{array} \right\}.$$

   That is, $\tau^a g^{-a+1/2} \rho^{-a} \ell^{-a-1/2} t$ is dimensionless for each a. We might take $a = 1/2$, and set

$$\bar{t} \;:=\; \frac{t}{\ell}\sqrt{\frac{\tau}{\rho}}\,.$$

...

■

Meaning of 'solution'

Background concepts:

- domain D in 'space' of independent variables
- open set
- class C^n of functions on an open set D
- connected set
- simply connected set

We say a function $\tilde{u}(t, \mathbf{x})$ is a (**classical**, or **strong**, though I think there is some discrepancy in meaning for the two terms) **solution** to the linear PDE

$$L[u] \;=\; f(t, \mathbf{x})$$

on D, an open region of $(t, \mathbf{x})$-space, if $\tilde{u}$ has sufficiently many derivatives in D that $L[\tilde{u}]$ makes sense pointwise in D, and the equation is satisfied at each point of D when $u = \tilde{u}$. One often expects to find a solution u which has continuous partial derivatives up to the order of the differential operator L. Though this is not always the case, it makes sense to expect most classical solutions to be found in

$$C^n(D) := \left\{ u \colon D \to \mathbb{R} \text{ (or } \mathbb{C}) \,\middle|\, \text{all derivatives of } u \text{ of order } \leq n \text{ exist and are continuous throughout } D \right\} .$$

Example 8: 1D Wave Equation on Bounded Interval

Consider the 1D wave equation

$$u_{tt} = c^2 u_{xx} , \qquad x \in [0, \ell], \quad t > 0 . \tag{8}$$

The following are classical solutions:

$$u_n(t,x) = \cos\left(\frac{n\pi ct}{\ell}\right)\sin\left(\frac{n\pi ct}{\ell}\right), \quad n = 1, 2, 3, \ldots ,$$
$$v(t,x) = \alpha x + \beta t + \gamma , \quad \alpha, \beta, \gamma \in \mathbb{R}$$
etc.

Remarks:

- Note that each is in $C^2(D)$—in fact, in $C^\infty(D)$—where D is the region of the (t,x)-plane given by $\left\{(t,x) \,\middle|\, 0 < x < \ell, \quad t > 0\right\}$.
- While there are many solutions of (8), they become less numerous as we impose conditions. The u_n also satisfy both the BCs (of **Dirichlet type**)

$$u(t,0) = 0, \quad u(t,\ell) = 0, \qquad t > 0,$$

but not v. Adding ICs

$$u(0,x) = f(x) , \quad u_t(0,x) = g(x) , \qquad x \in [0,\ell],$$

to the problem is enough to yield uniqueness of solution.

It is even possible, for a given choice of BCs, ICs, that the unique solution is no longer a classical one. For example, when $\ell = 1$ and the ICs are for a *plucked string*

$$u(0,x) = \left\{ \begin{array}{ll} x , & 0 \leq x < 1/2 \\ 1 - x , & 1/2 \leq x \leq 1 \end{array} \right\} , \quad u_t(0,x) = 0 , \qquad x \in [0,1].$$

■

Linearity

Definition 1. A *mapping* (*transformation*, *operator*, *function*) $L\colon \mathcal{V} \to \mathcal{W}$ from a vector space $\mathcal{V}$ into a vector space $\mathcal{W}$ is said to be **linear** if, for every $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and each pair of scalars α, β,

$$L(\alpha\mathbf{u} + \beta\mathbf{v}) \;=\; \alpha L\mathbf{u} + \beta L\mathbf{v}\,.$$

1st Order Problems/Method of Characteristics

Example 9: A quirk in PDEs (purpose(?) of Section 2.1 in Olver)

Consider the DE $u_t = 0$ as an

- ODE (really $u' = 0$)
  Answer?: $u(t) = \xi$ (i.e., value independent of t). Not always!
  In fact, it depends upon the domain in which problem is posed. If posed on single connected interval, answer is correct. If posed on two disjoint intervals, u may equal two different constants, one in each interval
- PDE in two independent variables (t, x)
  Answer?: $u(t, x) = f(x)$. Again, not always!
  Consider the domain of definition to be tx-plane minus the negative x-axis. Then PDE has (classical) solution

$$u(t,x) = \begin{cases} 0, & x > 0 \\ -x^2, & x \le 0,\ t < 0, \\ x^2, & x \le 0,\ t > 0. \end{cases}$$

■

Result 2 (Exercise 2.1.9)**.** Suppose $D \subset \mathbb{R}^2$ is an open set having the property that its intersection with any line x = constant is either empty or a connected interval. For any classical solution $u(t, x)$ to

$$\frac{\partial u}{\partial t} \;=\; 0\,, \qquad (t,x) \in D,$$

$u(t, x) = f(x)$ (i.e., independent of t).

Remarks:

- In the case of Result **??**, have *standing waves*.
- For what follows, problems are stated on entire region of tx-plane in which $t \geq 0$ (a region like that described in Result **??**.

Consider a linear transport equation

$$u_t + c(t,x)u_x \;=\; 0\,, \quad x \in \mathbb{R}, \;\; t > 0. \tag{9}$$

Note that

- solution $u(t,x)$ has graph which is a surface in $\mathbb{R}^3$. For each point (t,x) (in the domain), there corresponds a point $(t,x,u(t,x))$ on the surface.
- A curve $x = x(t)$ of points in tx-plane will have a corresponding collection of points $(t,x(t),u(t,x(t)))$—a parametrized curve in $\mathbb{R}^3$ parametrized by t—on surface.
- For domain points along such a fixed curve, chain rule gives that

$$\frac{d}{dt}u(t,x(t)) \;=\; \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\,\frac{dx}{dt} \;=\; u_t + \left(\frac{dx}{dt}\right)u_x\,.$$

If we can solve the ODE

$$\frac{dx}{dt} \;=\; c(t,x)$$

for $x(t)$, then the solution $u(t,x)$ of (9) will satisfy

$$\frac{d}{dt}u(t,x(t)) \;=\; 0 \qquad \Rightarrow \qquad u(t,x(t)) \;=\; \text{constant}$$

when restricted to domain points along the **characteristic curve** $x(t)$. According to *counting principle*, a unique solution should require an initial condition, say, at $t_0 = 0$:

$$u(0,x) \;=\; \varphi(x), \quad x \in \mathbb{R}.$$

If such a φ is specified, finding the value of $u(t,x)$ anywhere in tx-plane amounts to tracing along the characteristic curve to the x-axis (assuming it gets there) and evaluating φ at the point of intersection.

Example 10: Problem: $u_t + cu_x \;=\; 0$ (c a constant), $u(0,x) = \sin x$

Here

$$\frac{dx}{dt} \;=\; c \qquad \Rightarrow \qquad x(t) = ct + \xi\,,$$

so the characteristic curves are *lines* with slope c: $x = ct + \xi$, one for each choice of ξ. For each (t,x) in the plane, the characteristic line has x-intercept $\xi = x - ct$, yielding solution

$$u(t,x) \;=\; \sin(x - ct)\,.$$

■

Example 11: Problem: $u_t + 3x^{2/3}u_x = 0, \quad u(0,x) = \varphi(x)$

Characteristics solve

$$\frac{dx}{dt} = 3x^{2/3} \qquad \Rightarrow \qquad x = (t+\xi)^3 .$$

So, given any valid (t,x) determine the characteristic (by finding the appropriate value of ξ) on which it lies:

$$\xi = x^{1/3} - t ,$$

and the initial value $x_0 = x(t_0)$ (assume $t_0 = 0$) of that characteristic:

$$x_0 = (0+\xi)^3 = (x^{1/3} - t)^3 .$$

Then

$$u(t,x) = \varphi((x^{1/3} - t)^3) .$$

Employ MATHEMATICA or SAGE to view characteristics, the surface $z = u(t,x)$, and a movie of the traveling waves (moving with nonuniform speed); see `characteristics.nb` or `characteristics2.nb`.

■

The method of characteristics may be applied to the nonhomogeneous transport equation

$$u_t + c(t,x)u_x = b(t,x) , \quad x \in \mathbb{R}, \quad t > 0, \qquad \text{subject to} \qquad u(0,x) = \varphi(x), \quad x \in \mathbb{R} . \tag{10}$$

Again, the characteristics are curves $x = x(t)$ in the tx-plane satisfying

$$\frac{dx}{dt} = c(t,x) ,$$

and along a characteristic $x(t)$,

$$\frac{d}{dt}u(t,x(t)) = u_t + c(t,x(t))u_x = b(t,x(t)) .$$

Thus, the solution u (when it exists) is not constant along characteristics for the nonhomogeneous PDE (10), but satisfies

$$u(t,x(t)) = \varphi(x(t_0)) + \int_0^t b(\tau, x(\tau))\, d\tau .$$

Example 12: Nonhomogeneous Transport Equation with Extra Term

To solve

$$u_t + u_x + u = e^{x+2t}, \quad x \in \mathbb{R}, \quad t > 0, \qquad \text{subject to} \qquad u(0,x) = \varphi(x), \quad x \in \mathbb{R}$$

using the characteristics $x(t) = t + \xi$, we note that

$$\frac{dy}{dt} + y \;=\; e^{x(t)+2t}\;,$$

where $y = u(t, x(t))$. Employing the integrating factor $\mu = e^t$, we get

$$e^t y(t) - y(0) \;=\; \int_0^t e^{x(\tau)+3\tau}\, d\tau \;=\; e^{\xi} \int_0^t e^{4\tau}\, d\tau \;=\; \frac{1}{4} e^{\xi} \left(e^{4t} - 1\right).$$

But $\xi = x - t$, and $y(0) = u(0, x(0)) = u(0, \xi) = \varphi(x - t)$. Thus,

$$u(t, x) \;=\; \varphi(x - t)e^{-t} + \frac{1}{4}\left(e^{x+2t} - e^{x-2t}\right).$$

■

Initial and Boundary Conditions

From ODEs, recall that a differential equation alone may have infinitely many solutions. The **general solution** of an n^{th} order (linear) DE generally involves n arbitrary constants. To nail down a unique solution requires n additional (initial) conditions. Peter Olver offers the following general guideline concerning the general solution of an n^{th} order PDE:

Counting principle (rough guide):

> The solution of an n^{th} order (linear) PDE in m independent variables depends on n arbitrary functions of $m - 1$ variables.

To nail down a *unique* solution of $L[u] = f$ requires additional conditions. There are several possibilities:

- **Initial Conditions (ICs)**
  - type studied in MATH 231
  - needed only for dynamical problems
  - Specify values of u and its derivatives $u', u'', \ldots, u^{(n-1)}$ at a single point t_0. The number of ICs corresponds to the highest order of time derivative in PDE.
- **Boundary conditions (BCs)** (most commonly in the case $n = 2$): Specify values of u and/or its derivatives at boundary points.

  Mention different senses of word "bounded"
  - bounded domain

– bounded solution

When domain of definition is bounded, the BCs may be of:

1. **Dirichlet type**: values of u on boundary are specified.
   Example 13: Heat Problem on a Uniform Bar

   $$u_t = u_{xx}, \quad a < x < b, \quad t > 0,$$

   subject to IC $u(0, x) = \varphi(x)$ and BCs

   $$u(t, a) = f(t), \quad u(t, b) = g(t).$$

   ■

   If both f, g are identically zero, call these *homogeneous Dirichlet BCs*

2. **Neumann type**: values of normal derivative of u are specified.
   Example 14: 1D Heat Problem with Neumann BCs

   $$u_t = u_{xx}, \quad a < x < b, \quad t > 0,$$

   subject to IC $u(0, x) = \varphi(x)$ and BCs

   $$u_x(t, a) = f(t), \quad u_x(t, b) = g(t).$$

   ■

   When f, g are 0, have *homogeneous Neumann BCs* which, for heat problem, correspond to endpoints being *insulated*. In 2-D, corresponding problem might look like this:
   Example 15: 2D Heat Problem with Homogeneous Neumann BCs

   $$u_t = \Delta u, \quad \text{in } R := \left\{(x, y) \,\middle|\, x^2 + y^2 < 1\right\}, \quad t > 0,$$

   subject to IC $u(0, x, y) = \varphi(x, y)$ and BCs

   $$\frac{\partial u}{\partial \mathbf{n}} := \nabla u \cdot \mathbf{n} = 0 \qquad \text{on } \partial R.$$

   ■

3. **Mixed type**
4. **Robin type**
   Will not see often in this course.
   Example 16:

$$u_t = u_{xx}, \quad a < x < b, \quad t > 0,$$

subject to IC $u(0, x) = \varphi(x)$ and BCs

$$u_x(t, a) + ku(t, a) = f(t), \quad u_x(t, b) + hu(t, b) = g(t).$$

■

Such a condition with k, $h > 0$ models heat exchange resulting from the ends of bar being placed in a reservoirs at temperatures $f(t)$, $g(t)$ respectively. See also **Newton's law of cooling**.

When domain of definition is unbounded:

Often we'll require solution to remain bounded within D (i.e., $|u(t, x)| \leq M$)

Consequences of Linearity

From ODEs, recall:

- **Superposition**. If L is a linear (ordinary) differential operator, and u_1, u_2 solve the ODEs

$$L[u] = f_1 \qquad \text{and} \qquad L[u] = f_2 \,,$$

then $\tilde{u} := u_1 + u_2$ solves the ODE $L[u] = f_1 + f_2$.

- Homogeneous linear n^{th}-order ODEs have a general solution

$$y_H = c_1 y_1(t) + \cdots + c_n y_n(t) \tag{11}$$

that involves n arbitrary constants $c_1, \ldots, c_n$.

  - Linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ are represented by matrices, with *input* $\mapsto$ *output* map given by $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. To each m-by-n matrix $\mathbf{A}$ is associated a **nullspace**, the set

$$\left\{ \mathbf{x} \in \mathbb{R}^n \,\middle|\, \mathbf{A}\mathbf{x} = \mathbf{0} \right\} \,,$$

along with a number known as its *nullity*, $\text{nullity}\,(\mathbf{A}) = \dim(\text{null}\,(\mathbf{A}))$. Call $\text{null}\,(\mathbf{A})$ the **kernel** of the associated linear transformation. A **basis** of this kernel will consist of $\ell = \text{nullity}\,(\mathbf{A})$ independent vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_\ell\}$; the general form of an element of the nullspace/kernel has the form

$$c_1\mathbf{u}_1 + \cdots + c_\ell\mathbf{u}_\ell$$

($\ell = \text{nullity}\,(\mathbf{A})$ degrees of freedom).

- For a linear ordinary differential operator L, call the *space* of solutions to $L[u] = 0$ the kernel of L, $\ker(L)$. A general element in $\ker(L)$ has the form (11).
- If $\mathbf{x}_p$ is any solution of the matrix problem $\mathbf{Ax} = \mathbf{b}$ (a nonhomogeneous problem) and $\mathbf{x}_H$ is in the nullspace (kernel) of $\mathbf{A}$, then $\mathbf{x}_p + \mathbf{x}_H$ is also a solution of $\mathbf{Ax} = \mathbf{b}$.

  Likewise, if y_p is a solution of the nonhomogeneous linear ODE $L[u] = f$, and $y_H \in \ker(L)$, then $y_p + y_H$ is a solution of $L[u] = f$ as well. Generally speaking, if $L[u] = f$ has *any* solution, then it has *infinitely many*.

Fourier Transform (Incomplete; Skipped)

Definition 3. Suppose f is a piecewise continuous (perhaps complex-valued) function on $\mathbb{R}$, with $\lim_{x\to\pm\infty} f(x) = 0$. We define the **Fourier transform** $\hat{f}(k) = \mathcal{F}([]\, f(x)]$ by

$$\hat{f}(k) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx}\, dx,$$

whenever this integral exists for each real k. In such instances, the **inverse Fourier transform** is given by

$$\mathcal{F}^{-1}([]\, \hat{f}(k)] := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{ikx}\, dk.$$

Remarks:

- The Fourier transform is a linear operator on a space of functions. See Exercise 7.1.9 concerning the inverse Fourier transform.
- Relationship to Fourier series:
- We would like to say $f(x) = \mathcal{F}^{-1}([]\, \hat{f}(k)]$. In fact, we have

  Theorem 4. Suppose $f \in C^1(\mathbb{R})$, and $f(x) \to 0$ as $|x| \to \infty$ quickly enough so that the integral defining its Fourier transform $\hat{f}(k)$ converges absolutely for each $k \in \mathbb{R}$. Then at each $x \in \mathbb{R}$ the inverse Fourier transform $\mathcal{F}^{-1}([]\, \hat{f}(k)]$ converges to

  $$\frac{1}{2}[f(x^-) + f(x^+)]\,,$$

  which equals $f(x)$ whenever x is a point of continuity for f.

- Symmetry in definitions of Fourier and inverse Fourier transforms

- no uniformity amongst authors on actual *definition* of Fourier transform
- If $\hat{f}(k) = \mathcal{F}([]\, f(x)]$, then $\hat{\hat{f}}(k) = \mathcal{F}([]\, \hat{f}(x)] = f(-k)$.
-

Cauchy Problem for 1D Wave Equation

We consider the Cauchy problem

$$u_{tt} = c^2 u_{xx}, \qquad \text{subject to} \qquad u(0,x) = \phi(x), \quad u_t(0,x) = \psi(x). \tag{12}$$

There is another form of the wave equation (called *canonical form*), obtained via the change of variables

$$\xi = x + ct, \qquad \tau = x - ct,$$

(to characteristic coordinates) that is more readily-solved. Applying this transformation,

$$\begin{aligned}
u_x &= u_\xi \frac{\partial \xi}{\partial x} + u_\tau \frac{\partial \tau}{\partial x} \\
&= u_\xi + u_\tau. \\
\Rightarrow \quad u_{xx} &= \frac{\partial}{\partial x} u_\xi + \frac{\partial}{\partial x} u_\tau \\
&= \left(\frac{\partial}{\partial \xi} u_\xi\right)\frac{\partial \xi}{\partial x} + \left(\frac{\partial}{\partial \tau} u_\xi\right)\frac{\partial \tau}{\partial x} + \left(\frac{\partial}{\partial \xi} u_\tau\right)\frac{\partial \xi}{\partial x} + \left(\frac{\partial}{\partial \tau} u_\tau\right)\frac{\partial \tau}{\partial x} \\
&= u_{\xi\xi} + 2u_{\xi\tau} + u_{\tau\tau},
\end{aligned}$$

and, similarly,

$$\begin{aligned}
u_t &= cu_\xi - cu_\tau \\
\Rightarrow \quad u_{tt} &= c\left(u_{\xi\xi}\frac{\partial \xi}{\partial t} + u_{\tau\xi}\frac{\partial \tau}{\partial t}\right) - c\left(u_{\xi\tau}\frac{\partial \xi}{\partial t} + u_{\tau\tau}\frac{\partial \tau}{\partial t}\right) \\
&= c(cu_{\xi\xi} - cu_{\tau\xi}) - c(cu_{\xi\tau} - cu_{\tau\tau}) \\
&= c^2(u_{\xi\xi} - 2u_{\xi\tau} + u_{\tau\tau}).
\end{aligned}$$

Thus, the PDE $u_{tt} - c^2 u_{xx} = 0$ becomes

$$-4c^2 u_{\xi\tau} = 0, \qquad \text{or simply} \qquad u_{\xi\tau} = 0.$$

The reason for desiring this form is that we can simply integrate twice:

$$\begin{aligned}
u_\xi &= \int u_{\xi\tau}\, d\tau = \int 0\, d\tau = f(\xi). \qquad (f \text{ an arbitrary function}) \\
\Rightarrow \quad u &= \int f(\xi)\, d\xi \\
&= F(\xi) + G(\tau) \\
&= F(x+ct) + G(x-ct) \qquad \text{(back in original coordinates)},
\end{aligned}$$

where F and G are arbitrary functions. Note that

$$u_t(x,t) = cF'(x+ct) - cG'(x-ct),$$

so

$$\begin{aligned}
&\phi(x) = u(0,x) = F(x) + G(x) \\
&\psi(x) = u_t(0,x) = cF'(x) - cG'(x)
\end{aligned} \quad \Rightarrow \quad \begin{cases} c\phi'(x) = cF'(x) + cG'(x) \\ \psi(x) = cF'(x) - cG'(x) \end{cases} \quad \text{(assuming } \phi \text{ is differentiable)}$$

$$\Rightarrow \quad \begin{cases} F'(x) = \dfrac{1}{2c}[c\phi'(x) + \psi(x)] \\ G'(x) = \dfrac{1}{2c}[c\phi'(x) - \psi(x)] \end{cases}$$

$$\Rightarrow \quad \begin{cases} F(x) = F(0) + \dfrac{1}{2c}\displaystyle\int_0^x [c\phi'(z) + \psi(z)]\,dz \\ G(x) = G(0) + \dfrac{1}{2c}\displaystyle\int_0^x [c\phi'(z) - \psi(z)]\,dz \end{cases}$$

So, we have

$$\begin{aligned}
u(t,x) &= F(x+ct) + G(x-ct) \\
&= F(0) + G(0) + \frac{1}{2c}\left\{\int_0^{x+ct} [c\phi'(z) + \psi(z)]\,dz + \int_0^{x-ct} [c\phi'(z) - \psi(z)]\,dz\right\} \\
&= \phi(0) + \frac{1}{2c}\left\{\int_0^{x+ct} [c\phi'(z) + \psi(z)]\,dz + \int_{x-ct}^{0} [\psi(z) - c\phi'(z)]\,dz\right\} \\
&= \phi(0) + \frac{1}{2c}\int_{x-ct}^{x+ct} \psi(z)\,dz + \frac{1}{2}\left[\int_0^{x+ct} \phi'(z)\,dz - \int_{x-ct}^{0} \phi'(z)\,dz\right] \\
&= \phi(0) + \frac{1}{2c}\int_{x-ct}^{x+ct} \psi(z)\,dz + \frac{1}{2}\Big[\phi(x+ct) - \phi(0) - \phi(0) + \phi(x-ct)\Big] \\
&= \frac{1}{2}\Big[\phi(x+ct) + \phi(x-ct)\Big] + \frac{1}{2c}\int_{x-ct}^{x+ct} \psi(z)\,dz.
\end{aligned}$$

This is d'Alembert's formula for the solution of the Cauchy problem for the wave equation.

Note the form of the solution: left and right-travelling waves, similar to solutions of the transport equation. Note also that the wave operator is the composition of two transport equation operators in opposite directions:

$$\frac{\partial^2}{\partial t^2} - c^2\,\frac{\partial^2}{\partial x^2} \;=\; \left(\frac{\partial}{\partial t} + c\,\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} - c\,\frac{\partial}{\partial x}\right).$$

Transport Equation: Finite Difference Solutions

Discuss finite difference approximations to derivatives:

- 1st derivative: forward/backward differences, centered difference

- 2nd derivative: centered difference

Example 17: An Ordinary BVP

Consider the (ordinary, not partial) BVP

$$y'' + y = 0, \quad x > 0 \qquad \text{subject to} \qquad y(0) = y(\pi) = 1.$$

Note: $y = \sin x$ is a soln.

Introduce mesh on $[0, \pi]$: $x_j = j\Delta x,\ j = 0, 1, ..., N+1$.
Notation for approximate soln. at mesh points: $y_i\ y(x_i)$
Finite difference eqn (using 2nd-order approx for y') for the y_{ij}:

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{2h} + y_i \ = \ 0 \qquad \Rightarrow \qquad y_{i-1} + 2(h-1)y_i + y_{i+1} = 0, \quad i = 2, ..., N-1.$$

Get similar equations employing BCs at $i = 1$, $i = N$
Write as matrix problem.
Solve matrix problem—see file `odeBVPfinDiffs.m`
Note:

- soln looks like zero function; indeed, zero is a solution of BVP
- problem has many solutions, so cannot fault method

■

Example 18: Finite Differences on the Transport Equation

from Section 3.1 in G. Sewell, *The Numerical Solution of Ordinary and Partial DEs*, 2nd Ed.

Consider elementary transport problem ($c > 0$, constant)

$$u_t + cu_x = 0, \quad x \in \mathbb{R}, \ \ t > 0, \qquad \text{subject to} \qquad u(0, x) = \phi(x).$$

Know (from theory) soln. is right-traveling wave with form identical to $\phi(x)$.

Attempt to solve using finite difference schemes:

Argue that it makes sense to emply forward difference approx. to u_t

Several approaches involving u_x:

1. Use centered difference formula

$$\frac{u_{i+1,j} - u_{ij}}{\Delta t} \ = \ -c\,\frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{2\delta x} \qquad \Rightarrow \qquad u_{i+1,j} \ = \ u_{ij} - \alpha(u_{i,j+1} - u_{i,j-1}),$$

where $\alpha = c\Delta t/(2\Delta x)$.

For the case $c = 1$, see the file `finDiffsTranport1.m`. Notice

- does not "travel"
- develops "artifacts" immediately (*unstable*)

Q: What is wrong?
Consider same problem posed on x-interval $[0, 1]$

- now need BC
  Argue BC at $x = 0$ is relevant, but one at $x = 1$ is irrelevant (true because soln is right-traveling wave)
  Say: value of u at $x = 0$ is *upwind* of values at other x-values
- Display the **stencil**. Point out that, without BC at $x = 1$, cannot solve for u_{ij} at final j's when $i > 0$.

2. Use backward difference formula
   Method, called *upwind scheme*, is motivated by previous attempt
   The difference eqn:

$$\frac{u_{i+1,j} - u_{ij}}{\Delta t} = -c\,\frac{u_{ij} - u_{i,j-1}}{\Delta x} \qquad \Rightarrow \qquad u_{i+1,j} = \beta u_{i,j-1} + (1-\beta)u_{ij}, \quad \text{where } \beta = c\,\frac{\Delta t}{\Delta x}.$$

   Display stencil.
   For case $c = 1$, see file `finDiffsTranport2.m`.
   Try out several choices of Δt, leaving Δx fixed.
   Note that

   - We get similar "artifacts" as method above when $\Delta t > \Delta x (= 0.2)$.
     More generally, this happens when $\Delta t > \Delta x/c$, for if we write $r := \Delta x/\Delta t$ with $r < c$, then the point $(0, x - ct)$ is outside the triangle with vertices $(0, x - rt)$, $(0, x)$ and (t, x). Essentially, the part of IC allowed to contribute to soln at (t, x) consists of points which are outside the **domain of dependence**, and do not enough time to reach (t, x).

     > [Sewell] The true soln depends on the value of *phi*(x) at the point $x - ct \ldots$ If $x - ct$ lies outside the interval $[x - rt, x]$, $\ldots$ then the approx. solns cannot possibly be convergint to the true soln, in general, because we can change the IV at $x - ct$, changing the true soln but not the limit of the approx. solns.

   - We get expected soln when $\Delta t = \Delta x/c$
   - We get "diffused" traveling wave soln (somewhat expected) if $\Delta t < \Delta x/c$.
     Two interesting quotes from Sewell:

     > If, on the other hand, $c < r$, the student may draw the incorrect conclusion that the method is still unstable, because the IVs may be changed at points inside the domain of dependence of the approx. soln, but not at $x - ct$, thereby changing the approx. solns and not the true soln. The error in this reasoning is that changing the approx. solns does not necessarily change their limit.

and

> It is interesting to note that using an upwind difference approx. to u_x is equivalent to adding an "artificial diffusion" term, $|c|\Delta x/(2u_{xx}$, to the transport problem ... In other words, upwinding has the same smoothing effect as adding a small amount of diffusion to the convection model.

■

Separation of Variables: Part I

Review of linear 1st order systems of ODEs $\mathbf{y}' = \mathbf{A}\mathbf{y}$:

- Assume solution of form $\mathbf{y}(t) = e^{\lambda t}\mathbf{v}$. Deduce that nontrivial solutions of this form arise if and only if $(\lambda, \mathbf{v})$ is an eigenpair of $\mathbf{A}$.
- If $\mathbf{A}$ has a "full set of eigenvectors", $\mathbf{v}_1, \ldots, \mathbf{v}_n$ associated with eigenvalues $\lambda_1, \ldots, \lambda_n$, general solution is

$$\mathbf{y}(t) \;=\; c_1 e^{\lambda_1 t}\mathbf{v}_1 + c_2 e^{\lambda_2 t}\mathbf{v}_2 + \cdots + c_n e^{\lambda_n t}\mathbf{v}_n \;.$$

Theorem 5 (Spectral Theorem for Real Matrices)**.** If $\mathbf{A}$ is a symmetric real matrix, then it has a full set of eigenvectors. Moreover, the eigenvalues of $\mathbf{A}$ are all real numbers, and eigenvectors associated with distinct eigenvalues are orthogonal—that is, $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ whenever $\lambda_i \neq \lambda_j$.

- Why is orthogonality nice?
  Suppose have basis $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbb{R}^n$, and wish to write some vector $\mathbf{u}$ as linear combination of vectors in S

$$\mathbf{u} \;=\; c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n \;. \tag{13}$$

$$\mathbf{u} \;=\; c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n \;.$$

  The usual thing: solve the matrix problem

$$\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \;=\; \mathbf{u} \;.$$

  If S is orthogonal basis, can get c_j, $j = 1, \ldots, n$, by dotting both sides of (13) with $\mathbf{v}_j$:

$$\mathbf{u}\cdot\mathbf{v}_j \;=\; (c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n)\cdot\mathbf{v}_j \;=\; \cdots \;=\; c_j(\mathbf{v}_j\cdot\mathbf{v}_j) \qquad \Rightarrow \qquad c_j \;=\; \frac{\mathbf{u}\cdot\mathbf{v}_j}{\mathbf{v}_j\cdot\mathbf{v}_j} \;.$$

Vector Space $L^2(a,b)$

- Elements ("vectors") are "functions" defined on a domain $[a,b]$ for which $\int_a^b |f(x)|^2\,dx < \infty$
- Define inner product beteen $f, g \in L^2(a,b)$ to be $\langle f, g\rangle := \int_a^b f(x)\overline{g(x)}\,dx$.
  - Meaning of $\bar{z}$, where $z = \alpha + \beta i \in \mathbb{C}$ with $\alpha, \beta \in \mathbb{R}$
  - Compare with usual dot product in $\mathbb{C}^n$

    $$\mathbf{v}\cdot\mathbf{w} := \sum_{j=1}^{n} v_j\overline{w_j}\,, \qquad \text{where} \qquad \mathbf{v} = (v_1,\ldots,v_n),\ \ \mathbf{w} = (w_1,\ldots,w_n) \in \mathbb{C}^n\,.$$

    Elements in $\mathbb{C}^n$ have n components; those in $L^2(a,b)$ have infinitely many.
  - Complex numbers $z = \alpha + \beta i$
    - plotted in plane
    - modulus $|z|$: intuitively ought to equal $\sqrt{\alpha^2+\beta^2}$
      Note that $z\bar{z} = (\alpha+\beta i)(\alpha-\beta i) = \alpha^2+\beta^2 =: |z|^2$
  - Call f, g **orthogonal**, writing $f \perp g$, if $\langle f, g\rangle = 0$.
  - We define length (or **norm**) in $\mathbb{R}^n$ via dot product $\|\mathbf{v}\| = \sqrt{\sum_{j=1}^{n} v_j^2} = \sqrt{\mathbf{v}\cdot\mathbf{v}}$.

    Similarly, define norm in $L^2(a,b)$: $\|f\|_2 := \sqrt{\langle f, f\rangle}$.
    Thus, membership in $L^2(a,b)$ means $\int_a^b |f(x)|^2\,dx = \int_a^b f(x)\overline{f(x)}\,dx = \|f\|_2^2 < \infty$.
    A **norm** on a vector space $\mathcal{V}$ is a function which satisfies the following properties:
    1. **Nonnegativity**: $\|v\| \geq 0$ for all $v \in \mathcal{V}$.
    2. **Positive Definiteness**: $\|v\| = 0$ if and only if $v = 0$.
    3. **Homogeneity**: $\|\alpha v\| = |\alpha|\,\|v\|$ for all $v \in \mathcal{V}$ and all scalars α.
    4. **Triangle Inequality**: $\|u+v\| \leq \|u\| + \|v\|$ for all $u, v \in \mathcal{V}$.

    Strictly speaking, $\|\cdot\|_2$ is not a norm, lacking positive definiteness. (Give example of a function f which is nonzero only on a set of measure zero, so $\|f\|_2 = 0$.) If we consider such functions as indistinguishable from zero, $\|\cdot\|_2$ becomes a norm.
  - note simplification (for inner products in both $\mathbb{C}^n$, $L^2(a,b)$) when objects are real
- For each nonnegative integer n, the *space* $C^n(a,b)$ of n-times continuously differentiable functions defined for $a \leq x \leq b$ is a subspace of $L^2(a,b)$.

  Another popular norm used for continuous fns: $\|\cdot\|_\infty$.
  - its definition
  - visual depiction

- truly a norm (satisfies all 4 properties)
- undesirable characteristics:
  1. not applicable to as many fns as $\|\cdot\|_2$
  2. does not arise from any inner product

Recasting linear ODEs in operator form with spatial operator

Examples:

1. transport equation: $\partial u/\partial t = A[u]$, where $A[u] = -c(du/dx)$
2. 1D heat equation
   2D heat equation
3. 3D wave equation

Example 19: IBVP for 1D Heat Equation

Consider the Dirichlet heat problem on a bounded interval

$$u_t = u_{xx}, \quad 0 < x < 1, \quad t > 0, \qquad \text{subject to} \qquad u(0,x) = f(x), \quad u(t,0) = u(t,1) = 0.$$

As with the 1st-order linear system of ODEs $\mathbf{y}' = \mathbf{A}\mathbf{y}$, propose (*separable*) solutions of the form

$$u(t,x) \;=\; e^{\lambda t} v(x)$$

and deduce that, if nontrivial solutions of this form exist, then $v(\cdot)$ satisfies

$$v'' = \lambda v, \qquad \text{subject to} \qquad v(0) = 0 = v(1).$$

Show that $\lambda \le 0$, via the argument:

Starting with $v'' = \lambda v$, multiply through by $\overline{v}$ and integrate:

$$v''\overline{v} = \lambda v\overline{v} \qquad \Rightarrow \qquad \int_0^1 v''(x)\overline{v(x)}\,dx \;=\; \lambda \int_0^1 v(x)\overline{v(x)}\,dx$$

$$\Rightarrow \qquad \lambda \;=\; \frac{\int_0^1 v''\overline{v}\,dx}{\int_0^1 v\overline{v}\,dx} \;=\; \frac{v'(x)\overline{v(x)}\Big|_0^1 - \int_0^1 v'(x)\overline{v'(x)}\,dx}{\|v\|_2^2}$$

$$= \; -\frac{\int_0^1 v'(x)\overline{v'(x)}\,dx}{\|v\|_2^2} \;=\; -\frac{\|v'\|_2^2}{\|v\|_2^2} \;\le\; 0\,.$$

(The ratio in the expression for λ is called a **Rayleigh quotient**.) By this result, can write $\lambda = -\omega^2$ for $0 \le \omega < \infty$. So, our problem in v becomes

$$v'' + \omega^2 v = 0, \qquad \text{subject to} \qquad v(0) = 0 = v(1), \tag{14}$$

which (prior to applying the BCs) has solution

$$v(x) \;=\; a\cos(\omega x) + b\sin(\omega x)\;.$$

The BC $v(0) = 0$ implies $a = 0$.
The BC $v(1) = 0$ implies $\omega = n\pi$, $n = 1, 2, \ldots$.
Thus, the only instances in which (14) has a nontrivial solution are those in which

$$\omega = \omega_n = n\pi, \;\; n = 1, 2, \ldots, \qquad \text{in which case} \qquad v = v_n(x) = \sin(n\pi x)\;.$$

Therefore, our heat problem has infinitely many separable solutions

$$u_n(t, x) \;=\; e^{-n^2\pi^2 t}\sin(n\pi x), \quad n = 1, 2, \ldots.$$

The general solution (still not having applied the IC) is

$$u(t, x) \;=\; \sum_{n=1}^{\infty} c_n u_n(t, x) \;=\; \sum_{n=1}^{\infty} c_n e^{-n^2\pi^2 t}\sin(n\pi x)\;.$$

As is the case with IVPs for ODEs, the IC should allow us to determine a unique solution—that is, find correct values for the c_n. We have

$$f(x) \;=\; u(0, x) \;=\; \sum_{n=1}^{\infty} c_n \sin(n\pi x)\;.$$

It turns out (see HW) that, in the inner product of $L^2(0, 1)$, $\langle \sin(m\pi\cdot), \sin(n\pi\cdot)\rangle = 0$ whenever $m \ne n$. Employing this orthogonality, we take the inner product with $\sin(k\pi\cdot)$:

$$\langle f, \sin(k\pi\cdot)\rangle \;=\; \left\langle \sum_{n=1}^{\infty} c_n \sin(n\pi\cdot), \sin(k\pi\cdot)\right\rangle \;=\; \cdots \;=\; c_k \|\sin(k\pi\cdot)\|_2^2\;,$$

and thus

$$c_k \;=\; \frac{\langle f, \sin(k\pi\cdot)\rangle}{\|\sin(k\pi\cdot)\|_2^2} \;=\; 2\langle f, \sin(k\pi\cdot)\rangle \qquad \text{(see HW)}\;.$$

So, we get solution

$$u(t, x) \;=\; \sum_{n=1}^{\infty} 2\langle f, \sin(n\pi\cdot)\rangle\, e^{-n^2\pi^2 t}\sin(n\pi x)\;.$$

■

Some remarks:

- Overview of what we did:
  - We viewed the PDE as $\partial u/\partial t = A[u]$, where A is a *spatial* differential operator.
  - We assumed separable solutions of form $e^{\lambda t}v(x)$ (more general form would be $p(t)q(x)$).
  - Separability led to a BVP

$$A[v] = \lambda v, \qquad \text{subject to BCs},$$

    one which had (countably) infinitely-many eigenvalues, all real and non-positive.
  - We superimposed these solutions of the PDE/BCs

$$u(t,x) \;=\; \sum_{n=1}^{\infty} c_n u_n(t,x)\,, \tag{15}$$

    where the $u_n(t,x)$ are the separable solutions arising from the BVP above. Impicitly, we assumed that the $u_n(t,x)$ are sufficiently *rich* (in the sense of forming a basis for all solutions of the BVP) that we may consider (15) to be the **general solution** of the heat problem (modulo IC).
  - We used orthogonality of the eigenfunctions to obtain appropriate coefficients c_n in (15) to satisfy the IC.

- Along with a change in the operator, changes to the BCs and/or (spatial) domain on which the problem is stated have an effect on the eigenpairs. For example, if the above problem is stated for $0 < x < \ell$, then the eigenpairs are

$$\lambda_n = -\frac{n^2\pi^2}{\ell^2}, \qquad v_n(x) = \sin\left(\frac{n\pi x}{\ell}\right), \qquad n = 1, 2, \ldots,$$

  with the (still mutually orthogonal) eigenfunctions having squared $L^2(0,\ell)$-norm

$$\|v_n(\cdot)\|_2^2 \;=\; \left\|\sin\left(\frac{n\pi\cdot}{\ell}\right)\right\|_2^2 \;=\; \int_0^\ell sin^2\left(\frac{n\pi x}{\ell}\right) dx \;=\; \frac{\ell}{2}\,.$$

- We will handle the Cauchy (pure IVP) heat problem differently

- Some big questions:
  1. Are the eigenfunctions going to be orthogonal routinely?
  2. Is it generally true that the separable eigensolutions are rich enough to form a general solution?
  3. Is the term-by-term differentiation of this series—necessary in demonstrating it satisfies the PDE—valid?
  4. In what sense does the series (15) converge?

Fourier Series

Let us solve another heat problem, this time with Neumann BCs.

Example 20: 1D Heat Problem with Homogeneous Neumann Conditions

Consider the Dirichlet heat problem (with $\gamma > 0$) on a bounded interval

$$u_t = \gamma u_{xx}, \quad 0 < x < \ell, \quad t > 0, \qquad \text{subject to} \qquad u(0,x) = f(x), \quad u_x(t,0) = u_x(t,\ell) = 0.$$

Carry out a similar analysis. Assume separable solutions $u(t,x) = e^{\lambda t} v(x)$ to arrive at the BVP

$$v'' - \frac{\lambda}{\gamma} v \;=\; 0, \qquad \text{subject to} \qquad v'(0) = 0 = v'(\ell)\,.$$

Use the Rayleigh quotient again to show that $\lambda \le 0$, so can write $\lambda/\gamma = -\omega^2$, with $0 \le \omega < \infty$. The resulting BVP

$$v'' + \omega^2 v \;=\; 0, \qquad \text{subject to} \qquad v'(0) \;=\; 0 \;=\; v'(\ell), \tag{16}$$

has general solution (modulo the BCs)

$$\begin{aligned} v(x) \;&=\; A\cos(\omega x) + B\sin(\omega x) \qquad &\Rightarrow \qquad v'(x) \;&=\; -A\omega\sin(\omega x) + B\omega\cos(\omega x) \\ \text{BC} \quad v'(0) \;&=\; 0 \qquad &\Rightarrow \qquad \text{either} \quad \omega = 0 \;\text{ or }\; B = 0\,. \end{aligned}$$

Since $\omega = 0$ yields a nontrivial (constant) solution, $\lambda_0 = 0$ is truly an eigenvalue with corresponding eigenfunction $v_0(x) = 1$ (because it satisfies the ODE and *both* BCs). Focusing now on the case when $\omega \neq 0$, we have

$$\begin{aligned} v(x) \;&=\; A\cos(\omega x) \qquad &\Rightarrow \qquad v'(x) \;&=\; -A\omega\sin(\omega x) \\ \text{BC} \quad v'(\ell) \;&=\; 0 \qquad &\Rightarrow \qquad \omega_n \;&=\; \frac{n\pi}{\ell}, \quad n = 0, 1, 2, \ldots. \end{aligned}$$

(Note: The case $n = 0$ is redundant.) Thus, we have eigenvalues

$$\lambda_n \;=\; -\frac{\gamma n^2\pi^2}{\ell^2}, \qquad \text{with corresp. eigenfns} \qquad v_n(x) \;=\; \cos\left(\frac{n\pi x}{\ell}\right), \qquad n = 0, 1, 2, \ldots.$$

The corresponding separable solutions are

$$u_n(t,x) \;=\; e^{-\gamma n^2\pi^2 t/\ell^2}\cos\left(\frac{n\pi x}{\ell}\right)\,,$$

which we superimpose (assuming completeness) to get general solution (modulo IC)

$$u(t,x) \;=\; \sum_{n=0}^{\infty} c_n u_n(t,x) \;=\; \sum_{n=0}^{\infty} c_n e^{-\gamma n^2\pi^2 t/\ell^2}\cos\left(\frac{n\pi x}{\ell}\right)\,.$$

As in the previous example, the eigenfns of the spatial differential operator with BCs for this problem are mutually orthogonal in the $L^2(0,\ell)$-inner product. Moreover,

$$\left\|\cos\left(\frac{n\pi\cdot}{\ell}\right)\right\|_2^2 = \begin{cases} \ell, & \text{if } n = 0, \\ \dfrac{\ell}{2}, & \text{if } n = 1, 2, \ldots. \end{cases}$$

Using this, if we define

$$a_k := \frac{2}{\ell}\left\langle f, \cos\left(\frac{k\pi\cdot}{\ell}\right)\right\rangle = \frac{2}{\ell}\int_0^\ell f(x)\cos\left(\frac{k\pi x}{\ell}\right)dx\,, \tag{17}$$

then (we assert without proper justification) these a_n satisfy

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{\ell}\right) \tag{18}$$

(series converges in the L^2-norm sense), and we propose the following as a solution of the homogeneous Neumann IBVP:

$$u(t,x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n e^{-\gamma n^2\pi^2 t/\ell^2}\cos\left(\frac{n\pi x}{\ell}\right).$$

■

Remarks:

- Equations (17), (18) together are called the **Fourier cosine series** (FCS) of f. If these eigenfns form a *basis* for $L^2(0,\ell)$—and they do!—then (18) holds in $[0,\ell]$ for every $f \in L^2(0,\ell)$ given that the coefficients are the ones from (17).
- The equals sign in (18) must be understood in the L^2-sense. (Olver writes $\sim$ in place of of =.) That is, the series converges in norm to f, or

  $$\lim_{n\to\infty}\left\| f - \frac{a_0}{2} - \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi\cdot}{\ell}\right)\right\|_2 = 0\,.$$

  Visually, this means (in a technical sense) something about the disappearance of *space* between f and its truncated FCS as more terms are kept, quite different from pointwise convergence.
- Use the Octave program `fcs_approx.m` to demonstrate this convergence for cosine series. View f and its truncated series on $(0,\ell)$, $(-\ell,\ell)$, and $(-2\ell,2\ell)$. Conclude the FCS converges to the even (2ℓ)-periodic extension of f.
- Have already hinted that each $f \in L^2(0,\ell)$ has a **Fourier sine series**

  $$f(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{\ell}\right), \qquad \text{with} \qquad b_k := \frac{2}{\ell}\int_0^\ell f(x)\sin\left(\frac{k\pi x}{\ell}\right)dx\,.$$

Again this is so because these eigenfns form a basis for $L^2(0,\ell)$ so long as the equals sign is understood in the sense of L^2 convergence on $(0,\ell)$. The series may be more broadly understood as converging (in L^2-sense) to the odd (2ℓ)-periodic extension of f.

- In the text, Olver solves the famous "Fourier (heat equation on a 1D) ring" problem

$$u_t = u_{xx}, \quad -\ell < x \le \ell, \quad t > 0,$$

subject to

$$u(t,-\ell) = u(t,\ell), \quad u_x(t,-\ell) = u_x(t,\ell), \quad u(0,x) = f(x)\,.$$

(Olver does the case $\ell = \pi$.) This heat problem gives rise to yet another boundary value (eigenvalue) problem

$$v'' = \lambda v, \qquad \text{subject to} \qquad v(-\ell) = v(\ell) \quad \text{and} \quad v'(-\ell) = v'(\ell)\,.$$

Once again, eigenvalues $\lambda = -\omega^2 \le 0$ are real and non-positive. As in the previous example

$$\lambda_0 = 0 \quad \text{is an eigenvalue w/ corresp. eigenfn} \quad v_0(x) = 1.$$

The other eigenvalues are $\lambda_n = -n^2\pi^2/\ell^2$, $n = 1, 2, \ldots$, this time having *two* corresponding *independent* eigenfns

$$\cos\left(\frac{n\pi x}{\ell}\right) \qquad \text{and} \qquad \sin\left(\frac{n\pi x}{\ell}\right).$$

Result 6. The functions $1, \cos(\pi x/\ell), \sin(\pi x/\ell), \cos(2\pi x/\ell), \sin(2\pi x/\ell), \cos(3\pi x/\ell), \sin(3\pi x/\ell), \ldots$ are an orthogonal basis for $L^2(-\ell,\ell)$ and have corresponding squared norms

$$\|1\|_2^2 = 2\ell\,, \qquad \left\|\cos\left(\frac{n\pi x}{\ell}\right)\right\|_2^2 = \ell\,, \qquad \text{and} \qquad \left\|\sin\left(\frac{n\pi x}{\ell}\right)\right\|_2^2 = \ell\,.$$

Because of this result, each $f \in L^2(-\ell,\ell)$ has a **classical Fourier series** expansion

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{n\pi x}{\ell}\right) + b_n \sin\left(\frac{n\pi x}{\ell}\right)\right], \tag{19}$$

where the coefficients are given by

$$a_n := \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\left(\frac{n\pi x}{\ell}\right)dx \qquad \text{and} \qquad b_n := \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\left(\frac{n\pi x}{\ell}\right)dx\,. \tag{20}$$

Once again, the equality (19) is to be understood, in general, as holding in the L^2-sense. The series, in fact, converges (in this sense) to the (2ℓ)-periodic extension of f.

The corresponding solution to the Fourier ring problem is

$$u(t,x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} e^{-n^2\pi^2 t/\ell^2}\left[a_n \cos\left(\frac{n\pi x}{\ell}\right) + b_n \sin\left(\frac{n\pi x}{\ell}\right)\right],$$

where the a_n, b_n are given by (20).

Demonstrate for various f defined on $[0, \ell]$ the

- odd (2ℓ)-periodic extension
- even (2ℓ)-periodic extension

and for various $f \in L^2(-\ell, \ell)$ the (2ℓ)-periodic extension, labeled $\tilde{f}$ in the text.

Show convergence of

- Fourier sine/cosine series (use `fss_approx.m`, `fcs_approx.m`), and
- classical Fourier series (use `cfs_approx.m`).

(Note: All three of these routines make calls to `fourierCoeff.m`.) Specifically, run commands like

```
> function y = f(x)
>  y = 4 - (x - .5).^2;
> end

> plot(xs, f(xs), xs, fss_approx(@f, xs, k, 2)), axis([-4 4 -4 4]), pause, end
```

Convergence of classical Fourier series

Definitions:

- A function f is said to be **piecewise continuous** on a bounded interval $[a, b]$ if it is continuous throughout $[a, b]$ except possibly at finitely many points x_j, $j = 1, \ldots, N$, and at each of these points the right and left-hand limits (as appropriate)

  $$f(x_j^-) := \lim_{x \to x_j^-} f(x) \qquad \text{and} \qquad f(x_j^+) := \lim_{x \to x_j^+} f(x)$$

  exist and are finite. Said another way, f cannot have *infinite discontinuities*, and only finitely many discontinuities of the other 2 types.

  Display some examples and nonexamples.

- A function f is said to be **piecewise** C^1 (**piecewise** C^n) on a bounded interval $[a, b]$ if both f and f' ($f, f', f'', \ldots, f^{(n)}$) are piecewise continuous on $[a, b]$.

  So, at every point (including discontinuities) f has well-defined left and right tangent lines (infinite slopes are excluded).

- A function f is piecewise continuous (piecewise C^1) on $\mathbb{R}$ if it is piecewise continuous (piecewise C^1) on every bounded interval $[a, b]$ within $\mathbb{R}$.

Theorem 7 (Pointwise Convergence of Classical Fourier Series)**.** If $\tilde{f}(x)$ is piecewise C^1 and is (2ℓ)-periodic, then its (classical) Fourier series converges pointwise (not just in the L^2-sense, though it does that, too). For $x \in \mathbb{R}$, the value to which the series converges is

$$\frac{1}{2}\left[\tilde{f}(x^+) + \tilde{f}(x^-)\right],$$

which is simply $\tilde{f}(x)$ for those points x at which $\tilde{f}$ is continuous.

Theorem 7 speaks to the pointwise convergence of

- classical Fourier series
- for a (2ℓ)-periodic function.

What if these criteria are not met?

- Case: f is C^1 on $\mathbb{R}$, but not (2ℓ)-periodic
  f has a unique (2ℓ)-periodic extension $\tilde{f}$ that equals f on $-\ell < x \leq \ell$. This extension is piecewise C^1 since f is. The classical FS is the same for both f and $\tilde{f}$ (considered as functions in $L^2(-\ell, \ell)$), so Theorem 7 is in play.
- Case: f is C^1 on $\mathbb{R}$, and we have its FSS
  Take $\tilde{f}$ to be the odd (2ℓ)-periodic extension of f. The key is to notice the classical FS of $\tilde{f}$ (considered as a function in $L^2(-\ell, \ell)$) and the FS series of f (considered as a function in $L^2(0, \ell)$) are identical.

Draw some example functions f and the corresponding pointwise limit of its Fourier series.

Theorem 8 (Differentiation of Classical Fourier Series)**.** Suppose the (2ℓ)-periodic extension of f is piecewise C^2. Then its Fourier series may be differentiated term-by-term. In particular,

$$f'(x) = \sum_{n=1}^{\infty} \frac{n\pi}{\ell}\left[b_n \cos\left(\frac{n\pi x}{\ell}\right) - a_n \sin\left(\frac{n\pi x}{\ell}\right)\right].$$

Discuss the meaning of "equals" in the previous equation.

Inner Product Spaces

Definition 9. An **inner product space** is a vector space $\mathcal{V}$ equipped with an inner product $\langle \cdot, \cdot \rangle$. What makes it an inner product is not the symbol used, but rather the properties it has—namely, it must be the case that

(i) $\langle \mathbf{u}, \mathbf{v} \rangle$ is a scalar[1] for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

(ii) $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ for all $\mathbf{v} \in \mathcal{V}$, with equality if and only if $\mathbf{v} = \mathbf{0}$.

(iii) $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$ for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$. (Complex conjugation is unnecessary when $\mathcal{V}$ is a *real* vector space.)

(iv) $\langle a\mathbf{u}, \mathbf{v} \rangle = a \langle \mathbf{u}, \mathbf{v} \rangle$ for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and all scalars a.

(v) $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$.

Such a $\mathcal{V}$ has a ready-made norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$, turning it into a **normed vector space**.

Remarks:

- The Euclidean spaces $\mathbb{R}^n$, $\mathbb{C}^n$ are inner product spaces when equipped with their usual inner products:
$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=1}^{n} u_j v_j \quad \text{in } \mathbb{R}^n, \qquad \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=1}^{n} u_j \overline{v_j} \quad \text{in } \mathbb{C}^n.$$
Indeed, the Euclidean spaces were the models for the theory of inner product space.
- Olver speaks of **Hilbert space**, calling $L^2(a, b)$ one. All Hilbert spaces are *a priori* inner product spaces. They have the additional property of *completeness* (Cauchy sequences converge to a value inside the space).

In any inner product space, we have the following familiar theorem:

Theorem 10 (Pythagorean Theorem)**.** Suppose $\mathcal{V}$ is an normed inner product space (using the induced norm). If $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ are orthogonal, then $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

[1] The word *scalar*, here, refers to a complex number. The exception is when our vector space is *real*, in which case a scalar must be a *real* number.

Proof. We have

$$\|\mathbf{u}+\mathbf{v}\|^2 \;=\; \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}\rangle \;=\; \cdots \;=\; \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 \,.$$

□

Projections (somewhat different than handled in class)

In homework (Problem ⋆11) we saw that a typical term in a Fourier series for f—

$$b_n \sin\left(\frac{n\pi x}{\ell}\right) \;=\; \frac{\left\langle f, \sin\left(\frac{n\pi\cdot}{\ell}\right)\right\rangle}{\left\|\sin\left(\frac{n\pi\cdot}{\ell}\right)\right\|_2^2} \sin\left(\frac{n\pi x}{\ell}\right) \;=\; \mathrm{proj}_g f \,,$$

—is just the projection of f onto the function $g(\cdot) = \sin(n\pi\,\cdot/\ell)$. In a **normed vector space** $\mathcal{V}$, a projection $\mathrm{proj}_{\mathbf{u}}\mathbf{w}$ of one vector onto another is simply the vector in the subspace of $\mathcal{V}$ spanned by $\mathbf{u}$ that is closest to $\mathbf{w}$; in symbols,

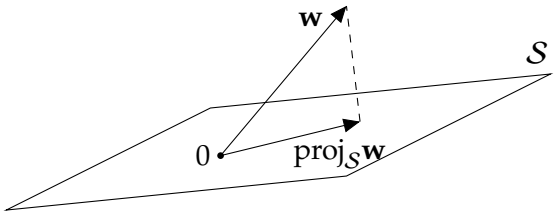
$$\|\mathbf{w} - \mathrm{proj}_{\mathbf{u}}\mathbf{w}\| \;\le\; \|\mathbf{w} - c\mathbf{u}\| \tag{21}$$

for all scalars c. When our norm arises from an inner product (i.e., $\|\mathbf{w}\| = \sqrt{\langle \mathbf{w}, \mathbf{w}\rangle}$), we have this formula for the projection

$$\mathrm{proj}_{\mathbf{u}}\mathbf{w} \;=\; \frac{\langle \mathbf{w}, \mathbf{u}\rangle}{\|\mathbf{u}\|^2}\,\mathbf{u} \,,$$

and so equality in (21) when $c = \langle \mathbf{w}, \mathbf{u}\rangle / \|\mathbf{u}\|^2$.

Things get a little harder when we wish to find the projection of a vector onto a subspace whose dimension is greater than 1. Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be a basis for an n-dimensional subspace $\mathcal{S}$ of $\mathcal{V}$. Analogous to our understanding of the projection of $\mathbf{w}$ onto another vector $\mathbf{u}$, we think of $\mathrm{proj}_{\mathcal{S}}\mathbf{w}$ as the vector in $\mathcal{S}$ closest to $\mathbf{w}$, or



$$\|\mathbf{w} - \mathrm{proj}_{\mathcal{S}}\mathbf{w}\| \;\le\; \|\mathbf{w} - \mathbf{v}\| \,, \qquad \text{for all vectors } \; \mathbf{v} \in \mathcal{S} \,.$$

Since each $\mathbf{v} \in \mathcal{S}$ is a linear combination of the basis vectors of S, we may write

$$\left\|\mathbf{w} - \mathrm{proj}_{\mathcal{S}}\mathbf{w}\right\| \;\le\; \left\|\mathbf{w} - \sum_{j=1}^{n} d_j \mathbf{u}_j\right\| \,, \qquad \text{for all choices of scalars } \; d_1, \ldots, d_n \,.$$

We infer from Problem ⋆11 that simply adding the projections of $\mathbf{w}$ onto the individual basis vectors

$$\mathrm{proj}_{\mathbf{u}_1}\mathbf{w} + \mathrm{proj}_{\mathbf{u}_2}\mathbf{w} + \cdots + \mathrm{proj}_{\mathbf{u}_n}\mathbf{w}$$

does not, in general, yield $\mathrm{proj}_{\mathcal{S}}\mathbf{w}$. Nevertheless, the same problem might make us guess this result:

Result 11. Suppose $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is an *orthogonal* basis of $\mathcal{S}$, a subspace of an inner product (normed) vector space $\mathcal{V}$. Then

$$\text{proj}_{\mathcal{S}}\mathbf{w} \;=\; \text{proj}_{\mathbf{u}_1}\mathbf{w} + \text{proj}_{\mathbf{u}_2}\mathbf{w} + \cdots + \text{proj}_{\mathbf{u}_n}\mathbf{w}$$

Proof. (*real* vector space case.) What we must prove is really that $\left\|\mathbf{w} - \sum_{j=1}^{n} \text{proj}_{\mathbf{u}_j}\mathbf{w}\right\| \;\leq\; \left\|\mathbf{w} - \sum_{j=1}^{n} d_j\mathbf{u}_j\right\|$ or, equivalently

$$\left\|\mathbf{w} - \sum_{j=1}^{n} \text{proj}_{\mathbf{u}_j}\mathbf{w}\right\|^2 \;\leq\; \left\|\mathbf{w} - \sum_{j=1}^{n} d_j\mathbf{u}_j\right\|^2 , \qquad \text{for all choices of scalars } \; d_1, \ldots, d_n \; .$$

Throughout the proof we will take $c_j := \left\langle \mathbf{w}, \mathbf{u}_j \right\rangle / \|\mathbf{u}_j\|^2$, which is the scalar multiple of $\mathbf{u}_j$ which gives $\text{proj}_{\mathbf{u}_j}\mathbf{w}$. (So, $\text{proj}_{\mathbf{u}_j}\mathbf{w} = c_j\mathbf{u}_j$.) Starting with the right-hand side, we have

$$\begin{aligned}
\|\mathbf{w} - \sum_j d_j\mathbf{u}_j\|^2 &= \left\langle \mathbf{w} - \sum_j d_j\mathbf{u}_j, \mathbf{w} - \sum_k d_k\mathbf{u}_k \right\rangle \;=\; \cdots \\
&= \|\mathbf{w}\|^2 - 2\sum_j d_j \left\langle \mathbf{w}, \mathbf{u}_j \right\rangle + \sum_j d_j^2\|\mathbf{u}_j\|^2 \\
&= \|\mathbf{w}\|^2 - 2\sum_j d_j c_j\|\mathbf{u}_j\|^2 + \sum_j d_j^2\|\mathbf{u}_j\|^2 \\
&= \|\mathbf{w}\|^2 - 2\sum_j d_j c_j\|\mathbf{u}_j\|^2 + \sum_j d_j^2\|\mathbf{u}_j\|^2 + \sum_j c_j^2\|\mathbf{u}_j\|^2 - \sum_j c_j^2\|\mathbf{u}_j\|^2 \\
&= \|\mathbf{w}\|^2 - \sum_j c_j^2\|\mathbf{u}_j\|^2 + \sum_j (d_j^2 - 2d_jc_j + c_j^2)\|\mathbf{u}_j\|^2 \\
&= \|\mathbf{w}\|^2 - \sum_j c_j^2\|\mathbf{u}_j\|^2 + \sum_j (d_j^2 - c_j)^2\|\mathbf{u}_j\|^2 \\
&\geq \|\mathbf{w}\|^2 - \sum_j c_j^2\|\mathbf{u}_j\|^2
\end{aligned}$$

To finish, we note first that $\left(w - \sum_j c_j\mathbf{u}_j\right)$ and $\left(\sum_j c_j\mathbf{u}_j\right)$ are orthogonal:

$$\left\langle \mathbf{w} - \sum_k c_k\mathbf{u}_k, \sum_j c_j\mathbf{u}_j \right\rangle \;=\; \sum_j c_j \left\langle \mathbf{w}, \mathbf{u}_j \right\rangle - \sum_k \sum_j c_k c_j \left\langle \mathbf{u}_k, \mathbf{u}_j \right\rangle \;=\; \sum_j c_j^2\|\mathbf{u}_j\|^2 - \sum_j c_j^2\|\mathbf{u}_j\|^2 \;=\; 0 \; .$$

Hence, by repeated applications of the Pythagorean Theorem,

$$\left\|\mathbf{w} - \sum_j c_j\mathbf{u}_j\right\|^2 \;=\; \|\mathbf{w}\|^2 - \left\|\sum_j c_j\mathbf{u}_j\right\|^2 \;=\; \|\mathbf{w}\|^2 - \sum_j c_j^2\|\mathbf{u}_j\|^2 \; .$$

□

Fourier series is fundamentally founded on having orthogonal bases, and so the above result applies. In particular, the collection $\{1, \cos(\pi x/\ell), \sin(\pi x/\ell), \cos(2\pi x/\ell), \sin(2\pi x/\ell), \ldots\}$ is an orthogonal basis for $L^2(-\ell, \ell)$. If we take any N of these, denoting them as $v_j(x) = \sin(j\pi x/\ell)$, $j = 1, \ldots, N$, and let $\mathcal{S}$ be the subspace spanned by them, then for $f \in L^2(-\ell, \ell)$,

$$\text{proj}_{\mathcal{S}} f \;=\; \sum_{j=1}^{N} \text{proj}_{v_j} f \;=\; \sum_{j=1}^{N} \frac{\langle f, v_j \rangle}{\|v_j\|_2^2}\, v_j \;.$$

Complex Fourier Series

Now let us return to the Fourier ring problem

$$u_t = u_{xx}, \quad -\ell < x \le \ell, \;\; t > 0,$$

subject to

$$u(t, -\ell) = u(t, \ell), \quad u_x(t, -\ell) = u_x(t, \ell), \;\; u(0, x) = f(x) \;.$$

In solving this problem, the assumption of separability led us to the eigenvalue problem

$$v'' + \lambda v = 0, \qquad \text{subject to} \qquad v(-\ell) = v(\ell) \quad \text{and} \quad v'(-\ell) = v'(\ell) \;,$$

where $\lambda = -\omega^2$ with $\omega \ge 0$. We found the values $\omega_n = n\pi/\ell$, $n = 0, 1, 2, \ldots$ were those for which nontrivial (eigenfns) solutions exist. For the negative eigenvalues $\lambda_n = -n^2\pi^2/\ell^2$ there were two independent eigenfns, which we took to be

$$\cos\left(\frac{n\pi x}{\ell}\right) \qquad \text{and} \qquad \sin\left(\frac{n\pi x}{\ell}\right) \;.$$

Now we propose to use a different pair of functions as a basis for the eigenspace associated with λ_n:

$$w_n(x) \;=\; e^{i\omega_n x} \;=\; e^{in\pi x/\ell} \qquad \text{and} \qquad w_{-n}(x) \;=\; e^{-i\omega_n x} \;=\; e^{-in\pi x/\ell} \;.$$

Claim 12. For $n > 0$, $w_n(\cdot)$ and $w_{-n}(\cdot)$ are orthogonal (under the inner product of $L^2(-\ell, \ell)$).

The (alternate) solution of the Fourier ring problem is

$$u(t, x) \;=\; \sum_{n=-\infty}^{\infty} c_n e^{-n^2\pi^2 t/\ell^2} w_n(x) \;, \qquad \text{with} \qquad c_n \;=\; \frac{\langle f, w_n(\cdot) \rangle}{\|w_n(\cdot)\|_2^2} \;.$$

The coefficients c_n are obtained by applying the IC

$$f(x) \;=\; u(0, x) \;=\; \sum_{n=-\infty}^{\infty} c_n w_n(x) \;. \tag{22}$$

The collection $\{w_n\}_{n=-\infty}^{\infty}$ is a complete orthogonal basis for $L^2(-\ell, \ell)$, so (22) holds in the *mean-square sense*, and is called the **complex exponential Fourier series** of f. By the projection results above, if $\mathcal{S}_n$ is the eigenspace associated with eigenvalue λ_n, then for $n > 0$ we have

$$c_{-n}w_{-n}(x) + c_n w_n(x) \;=\; \text{proj}_{\mathcal{S}_n} f \;=\; a_n \cos\left(\frac{n\pi x}{\ell}\right) + b_n \sin\left(\frac{n\pi x}{\ell}\right) .$$

Using this, we may establish the following relationships between coefficients of the complex exponential FS and those of the classical FS.

Claim 13. For each $n = 0, 1, 2, \ldots$,

$$\begin{aligned} a_n &= c_n + c_{-n}, & c_n &= \tfrac{1}{2}\left(a_n - ib_n\right), \\ b_n &= i(c_n - c_{-n}), & c_{-n} &= \tfrac{1}{2}\left(a_n + ib_n\right), \end{aligned} \qquad n = 0, 1, 2, \ldots .$$

Orthogonality of Eigenfunctions

We have seen now several instances in which the assumption of separability $u(t, x) = q(t)v(x)$ has led to an eigenvalue problem, one in which

- the eigenvalues are all real (perhaps all even of the same sign), and
- the resulting eigenfunctions are orthogonal.

Q: Why does this happen?

Self-Adjoint Operators

While the setting is simpler, we have the following experience with symmetric matrices (linear operators on $\mathbb{R}^n$):

Result 14. Suppose $\mathbf{A}$ is an n-by-n matrix. If $\mathbf{A}$ is *symmetric*, then

- its eigenvalues are all real, and
- there exists an orthogonal basis of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.

Q: Is there a concept that generalizes the notion of symmetry (for matrices) to linear operators in an inner product space?

A: Yes, *self-adjointness*.

Definition 15. Let L be a linear (differential) operator defined on some (dense) subset of $L^2(a,b)$. A pairing of L with BCs is said to be **self-adjoint** precisely when

$$\Big\langle L[\phi], \psi \Big\rangle \;=\; \Big\langle \phi, L[\psi] \Big\rangle \qquad \text{for all} \qquad \phi, \psi \in \text{dom}(L) \text{ satisfying the BCs.}$$

Example 21:

Let $L[v] = v''$, with homogeneous Dirichlet BCs: $v(a) = 0 = v(b)$. Then L is self-adjoint in $L^2(a,b)$. (See homework.)

■

Example 22:

Let $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) be bounded with boundary sufficiently smooth to support Green's ($n = 2$ case) or the Divergence ($n = 3$ case) Theorem. Take $L[v] = \Delta v$, with homogeneous Dirichlet BCs: $v(\mathbf{x}) = 0$ for $x \in \partial\Omega$. Then

$$\begin{aligned}
\Big\langle L[\phi], \psi \Big\rangle \;&=\; \int_\Omega \Delta\phi(\mathbf{x})\, \psi(\mathbf{x})\, d\mathbf{x} \;=\; \int_{\partial\Omega} (\nabla\phi \cdot \mathbf{n})(\mathbf{x})\, \psi(\mathbf{x})\, d\sigma - \int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\psi(\mathbf{x})\, d\mathbf{x} \\
&=\; -\int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\psi(\mathbf{x})\, d\mathbf{x} \;=\; \int_{\partial\Omega} \phi(\mathbf{x})\, (\nabla\phi \cdot \mathbf{n})(\mathbf{x})\, d\sigma - \int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\psi(\mathbf{x})\, d\mathbf{x} \\
&=\; \int_\Omega \phi(\mathbf{x})\, \Delta\psi(\mathbf{x})\, d\mathbf{x} \;=\; \Big\langle \phi, L[\psi] \Big\rangle \;,
\end{aligned}$$

showing L with homogeneous Dirichlet BCs is self-adjoint.

■

Note that on conclusion the proof gives us is that

$$\Big\langle \Delta\phi, \psi \Big\rangle \;=\; -\Big\langle \nabla\phi, \nabla\psi \Big\rangle ,$$

where the right-hand side is an appropriate definition for an inner product between vector functions in Ω. Taking ψ to be ϕ, and ϕ to be an eigenfn associated with eigenvalue λ, this immediately implies

$$\lambda \|\phi\|_2^2 \;=\; \Big\langle \Delta\phi, \phi \Big\rangle \;=\; -\Big\langle \nabla\phi, \nabla\phi \Big\rangle \;=\; \|\nabla\phi\|_2^2,$$

showing that eigenvalues of the Laplacian operator are non-positive (i.e., Δ is **negative semi-definite**).

Self-adjoint operators are truly an analog of symmetric matrices, as seen in the next result.

Result 16. Let L be a self-adjoint linear operator. Then

(i) all eigenvalues of L are real, and

(ii) eigenvectors (eigenfns) corresponding to distinct eigenvalues are orthogonal.

Proof. To prove (i), let (λ, v) be an eigenpair of L. Then

$$\lambda\|v\|^2 \;=\; \lambda\,\langle v, v\rangle \;=\; \langle \lambda v, v\rangle \;=\; \langle L[v], v\rangle \;=\; \langle v, L[v]\rangle \;=\; \langle v, \lambda v\rangle \;=\; \overline{\lambda}\,\langle v, v\rangle \;=\; \overline{\lambda}\|v\|^2\,.$$

Since v is an eigenvector, $\|v\| \neq 0$ and we may divide through to get $\lambda = \overline{\lambda}$, showing that $\lambda \in \mathbb{R}$.

Now let $\lambda,\ \mu \in \mathbb{R}$ be eigenvalues of L with corresponding eigenvectors u, v, respectively, and assume $\lambda \neq \mu$. Then

$$(\lambda - \mu)\,\langle u, v\rangle \;=\; \lambda\,\langle u, v\rangle - \mu\,\langle u, v\rangle \;=\; \langle \lambda u, v\rangle - \langle u, \mu v\rangle \;=\; \langle L[u], v\rangle - \langle u, L[v]\rangle \;=\; 0\,.$$

Since $\lambda \neq \mu$, it must be $\langle u, v\rangle = 0$, proving (ii). □

The previous result does not establish that a complete orthogonal basis of eigenfns of a self-adjoint operator exists. This is true, however, at least for the Laplacian, by Rellich's Theorem.

Theorem 17 (Rellich's Principle)**.** Let $\Omega \subset \mathbb{R}^n$ be bounded. Then eigenfunctions of the Laplacian operator with homogeneous Dirichlet boundary conditions form a complete orthogonal basis of $L^2(\Omega)$.

Theorem 18 (Rellich-Weyl Principle)**.** All diffusion problems

$$u_t \;=\; \Delta u$$

subject to homogeneous Dirichlet BCs on a bounded domain Ω are uniquely solvable for any given square-integrable initial shape f by orthogonal separation of variables.

Example 23: Sturm-Liouville Eigenvalue Problems (SLEPs)

Many of the eigenvalue problems that come up as a result of separation of variables are classified as **Sturm-Liouville eigenvalue problems**, or SLEPs. For given real-valued functions p, q defined on the interval $[a, b]$ with p differentiable, consider the operator

$$K[v] \;:=\; -\frac{d}{dx}\left[p(x)\frac{dv}{dx}\right] + q(x)v \;=\; -p(x)\frac{d^2v}{dx^2} - p'(x)\frac{dv}{dx} + q(x)v\,. \tag{23}$$

The eigenvalue problem for K is

$$K[v] = \lambda v, \qquad \text{subject to BCs:} \qquad \begin{cases} \alpha_1 v(a) + \alpha_2 v'(a) = 0, \\ \beta_1 v(b) + \beta_2 v'(b) = 0, \end{cases} \tag{24}$$

where $|\alpha_1| + |\alpha_2| > 0$ and $|\beta_1| + |\beta_2| > 0$. Assuming everything is real-valued here, then for sufficiently smooth functions u, v satisfying the boundary conditions we have

$$\begin{aligned} \langle K[u], v\rangle &= -\int_a^b v(x)\,\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)dx + \int_a^b q(x)u(x)v(x)\,dx \\ &= -p(x)u'(x)v(x)\Big|_a^b + \int_a^b \frac{du}{dx}\left(p(x)\frac{dv}{dx}\right)dx + \int_a^b q(x)u(x)v(x)\,dx \\ &= \Big[-p(x)u'(x)v(x)\Big]_a^b + \Big[p(x)u(x)v'(x)\Big]_a^b + \int_a^b \left[-\frac{d}{dx}\left(p(x)\frac{dv}{dx}\right) + q(x)v(x)\right]u(x)\,dx \\ &= \Big[-p(x)u'(x)v(x)\Big]_a^b + \Big[p(x)u(x)v'(x)\Big]_a^b + \langle u, K[v]\rangle\ . \end{aligned}$$

For K to be self-adjoint, we need these boundary terms to vanish. In the case that each α_j, β_j is nonzero ($j = 1, 2$), we have

$$\begin{aligned} &\Big[-p(x)u'(x)v(x)\Big]_a^b + \Big[p(x)u(x)v'(x)\Big]_a^b \\ &= p(a)u'(a)v(a) - p(b)u'(b)v(b) + p(b)u(b)v'(b) - p(a)u(a)v'(a) \\ &= \frac{p(a)}{\alpha_1\alpha_2}[\alpha_2 u'(a)\,\alpha_1 v(a) - \alpha_1 u(a)\,\alpha_2 v'(a)] + \frac{p(b)}{\beta_1\beta_2}[\beta_1 u(b)\,\beta_2 v'(b) - \beta_2 u'(b)\,\beta_1 v(b)] \\ &= \frac{p(a)}{\alpha_1\alpha_2}[\alpha_2 u'(a)\,\alpha_1 v(a) - \alpha_2 u'(a)\,\alpha_1 v(a)] + \frac{p(b)}{\beta_1\beta_2}[\beta_1 u(b)\,\beta_2 v'(b) - \beta_1 u(b)\,\beta_2 v'(b)] \\ &= 0\ , \end{aligned}$$

giving that K is self-adjoint. Thus, the eigenvalues of K are real, and eigenfns associated with distinct eigenvalues are orthogonal.

■

If p, p' and q are continuous on $[a, b]$ with p non-vanishing in this interval, then the SLEP (24) is said to be **regular**. (Otherwise, the SLEP is said to be **singular**, a condition that often occurs because p is zero at an endpoint of the interval.) For regular SLEPs the following theorem reveals even more about the eigenpairs.

Theorem 19. Assume the operator (23) is *regular* on the interval $[a, b]$. Then the SLEP (24)

(i) has infinitely many eigenvalues λ_n, $n = 1, 2, 3, \ldots$, (all of which are real, by the self-adjointness demonstrated in the last example) satisfying $\lim_{n\to\infty} |\lambda_n| = +\infty$, and

(ii) there exists a complete orthogonal system $\{v_n(\cdot)\}_{n=1}^{\infty}$ (basis) in $L^2(a,b)$ consisting of eigenfns of K. That is, every $f \in L^2(a,b)$ can be expanded in a **generalized Fourier series** as

$$f(x) \;=\; \sum_{n=1}^{\infty} c_n v_n(x)\,, \qquad \text{with} \qquad c_n \;=\; \frac{\langle f, v_n(\cdot)\rangle}{\|v_n\|^2}\,,$$

where the series on the right-hand side converges to f in (at least) the *mean-square sense*. In fact, the series converges pointwise on (a,b) to averages of left and right-hand limits of f when $f \in C^1$.

Separation of Variables: Part II

The general idea:

> Assume the solution of the PDE may be written as a product of functions which isolate the influence of the independent variables.

Example 24: Heat Problem with Mixed BCs

Solve the diffusion problem

$$u_t = ku_{xx}, \quad 0 < x < 1, \quad t > 0, \quad \text{subject to IC: } u(0,x) = f(x), \quad \text{and BCs: } \begin{cases} u(t,0) = 0, \\ u(t,1) + u_x(t,1) = 0. \end{cases}$$

Solution: Assume $u(t,x) = q(t)\varphi(x)$ to get

$$\frac{q'}{kq} \;=\; \frac{\varphi''}{\varphi} \;=\; \lambda,$$

yielding the two ODEs:

$$q' \;=\; \lambda k q, \qquad \text{and} \qquad \varphi'' - \lambda\varphi \;=\; 0.$$

Solving the former ODE (in t) gives $q(t) \;=\; q_0 \exp(k\lambda t)$. The latter ODE is subject to the conditions

$$\varphi(0) \;=\; 0 \qquad \text{and} \qquad \varphi(1) + \varphi'(1) \;=\; 0.$$

To determine the sign of λ, we note that if λ, φ are an eigenpair, then

$$\begin{aligned} \lambda\|\varphi\|_2^2 \;&=\; \int_0^1 (\varphi'')\varphi\,dx \;=\; \varphi'(x)\varphi(x)\Big|_0^1 - \int_0^1 (\varphi')^2\,dx \\ &=\; \varphi'(1)\varphi(1) - \|\varphi'\|_2^2 \;=\; -\varphi^2(1) - \|\varphi'\|_2^2 \qquad (\text{since } \varphi(1) = -\varphi'(1)). \end{aligned}$$

Thus, $\lambda = -[\varphi^2(1) + \|\varphi'\|^2]/\|\varphi\|^2 \le 0$. We note that 0 is not an eigenvalue since, in that case, the general solution for the ODE in x would be $\varphi(x) = ax + b$. But, the BC at $x = 0$ implies $b = 0$ and the BC at $\varphi = 1$ implies $2a = 0$, so there are no nontrivial solutions. Thus, taking $\lambda = -\omega^2$, with $\omega > 0$, the general solution of the ODE in x is

$$\varphi(x) = a\cos(\omega x) + b\sin(\omega x).$$

The condition $\varphi(0) = 0$ implies $a = 0$, so $\varphi(x) = \sin(\omega x)$. The right-end condition implies

$$\omega\cos\omega + \sin\omega \;=\; 0, \qquad \text{or} \qquad \tan\omega \;=\; -\omega,$$

an equation that has infinitely many solutions $\omega_n \sim (2n-1)\pi/2$ as $n \to \infty$. Thus, we have eigenvalues $\lambda_n = -\omega_n^2$, with corresponding eigenfunctions $\varphi_n(x) = \sin(\omega_n x)$, n = 1, 2, ... Note: The script `tanxPlusxZeros.m`, which works in Octave, accepts an integer m input and approximates the first m positive solutions of this equation (the negative solutions are just additive inverses of the positive ones).

Returning to the ODE in t, we see there is a solution $q_n(t) = \exp(-k\omega_n^2 t)$ for each $n = 1, 2, \ldots$ to go with each ϕ_n, so we get the series solution

$$u(x,t) \;=\; \sum_{n=1}^{\infty} c_n e^{-k\omega_n^2 t}\sin(\omega_n x), \qquad \text{where} \qquad c_n \;=\; \frac{\langle f, \sin(\omega_n \cdot)\rangle}{\|\sin(\omega_n \cdot)\|^2}.$$

Octave code for the case $k = 1$, $f(x) = 10x(1-x)$, keeping 20 terms of series solution (grab `tanxPlusxZeros.m` and `prob1Approx.m`):

```
octave:1> function y = f(x)
> y = 10*x.*(1-x);
> end

octave:2> xs = [0:.01:1]';
octave:3> for t = 0:.05:.3, plot(xs, prob1Approx(@f, xs, 20, ts, 1)), axis([0 1 0 2.5]), pause, end
```

■

Example 25: Solid Ball Dropped in Bath

Suppose a solid ball of radius a is dropped into a bath held at a fixed temperature. We wish to solve for the temperature of the ball as it changes in time. Such a problem, in general, relies on 4 independent variables (time and 3 spatial dimensions). We make, however, the simplifying assumption that the initial temperature profile is radially symmetric, and hence assume the solution *remains* radially symmetric for all $t > 0$, relying only on time and distance ρ from the ball's center. The Laplacian in spherical coordinates is

$$\Delta \;=\; \frac{1}{\rho^2}\frac{\partial}{\partial\rho}\left(\rho^2\frac{\partial}{\partial\rho}\right) + \frac{1}{\rho^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\rho^2\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\,.$$

But because of radial symmetry, we have the model problem

$$\begin{aligned} u_t &= \Delta u = \frac{\partial^2 u}{\partial \rho^2} + \frac{2}{\rho}\frac{\partial u}{\partial \rho}, \quad 0 \le \rho < a, \quad t > 0, \\ u(t,a) &= 0, \qquad\qquad\qquad\qquad t > 0, \\ u(0,\rho) &= f(\rho), \qquad\qquad\qquad t > 0. \end{aligned}$$

Solution: Assume $u(\rho,t) = v(\rho)q(t)$. Then our PDE becomes

$$vq' = v''q + \frac{2}{\rho}v'q, \qquad \text{or} \qquad \frac{q'}{q} = \frac{v'' + 2\rho^{-1}v'}{v} = \lambda.$$

Thus, we have two ODEs:

$$q' = \lambda q \qquad \text{and} \qquad v'' + \frac{2}{\rho}v' - \lambda v = 0.$$

By the BC $u(t,a) = 0$, we get $v(a) = 0$. Also, since we do not expect temperatures $u(t,\rho)$ to grow with time (in particular, to become unbounded), we impose the condition that $u(t,0)$ is bounded, which becomes $v(0)$ stays bounded. Now set $w(\rho) = \rho v(\rho)$. With this substitution,

$$v'' + \frac{2}{\rho}v' - \lambda v = 0 \qquad \text{becomes} \qquad w'' - \lambda w = 0,$$

subject to the conditions $w(a) = 0$ and $w(\rho)/\rho$ stays bounded as $\rho \to 0$. This latter condition implies $w(0) = 0$. Thus, for an eigenpair λ, w we have

$$\lambda \|w\|^2 = w(\rho)w'(\rho)\Big|_0^a - \int_0^a (w')^2\,d\rho = -\|w'\|^2,$$

giving that eigenvalues $\lambda \le 0$. The BCs further imply that $\lambda < 0$, so we may write $\lambda = -\beta^2$ for real $\beta > 0$. The usual arguments now lead to eigenvalues $\lambda_n = -(n\pi/a)^2$ with corresponding eigenvectors $w_n(\rho) = \sin(n\pi\rho/a)$, $n = 1, 2, \ldots$, or rather $v_n(\rho) = \sin(n\pi\rho/a)/\rho$, $n = 1, 2, \ldots$.

After solving the corresponding ODE in t, we have the series solution

$$u(\rho,t) = \sum_{n=1}^{\infty} c_n e^{-(n\pi/a)^2 t}\,\frac{\sin(n\pi\rho/a)}{\rho}.$$

We know we need to choose the c_n to satisfy

$$f(\rho) = \sum_{n=1}^{\infty} c_n\,\frac{\sin(n\pi\rho/a)}{\rho}, \qquad \text{which means} \qquad \rho f(\rho) = \sum_{n=1}^{\infty} c_n \sin(n\pi\rho/a).$$

Thus, we may choose the c_n as the usual sine coefficients for $\rho f(\rho)$:

$$c_n = \frac{2}{a}\int_0^a \rho f(\rho)\sin(n\pi\rho/a)\,d\rho.$$

For $f(\rho) \equiv 1$ and radius $a = 1$, we may view the truncated series solution keeping 20 terms at time $t = 0.5$ using the commands

```
octave:1> function y = f(rho)
> y = rho;
> end

octave:2> rhos = [0:.01:1]';
octave:3> plot(rhos, prob2Approx('f', rhos, 0.5, 20, 1))
```

■

Example 26: One-Dimensional Wave Equation on Bounded Interval

Consider the transverse vibrations of a string secured at its ends having known initial displacement and velocity—that is,

$$u_{tt} = c^2 u_{xx}, \quad 0 < x < \ell, \quad t > 0, \qquad \text{with BCs} \qquad u(t,0) = 0 = u(t,\ell),$$

subject to initial conditions

$$u(0,x) = f(x) \qquad \text{and} \qquad u_t(0,x) = g(x)\ .$$

Solution: Assuming separation $u(t,x) = q(t)\varphi(x)$ leads to

$$u(t,x) \ = \ \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{n\pi ct}{\ell}\right) + b_n \sin\left(\frac{n\pi ct}{\ell}\right) \right] \sin\left(\frac{n\pi x}{\ell}\right)\ .$$

Now use the ICs to get expressions for the a_n, b_n.

Demonstrate (truncated) solution when

$$f(x) \ = \ \begin{cases} x, & 0 < x < 1/2, \\ 1-x, & 1/2 \le x \le 1, \end{cases} \qquad g(x) = 0\ ,$$

using code from `dirichletWaveProb.m`.

■

Example 27: Laplace's Equation in a Rectangular Domain

Consider the problem

$$\Delta u = 0, \quad 0 < x < a, \quad 0 < y < b, \quad \text{subj. to} \quad u(x,0) = f(x), \ \ u(x,b) = 0, \ \ u(0,y) = 0, \ \ u(a,y) = 0.$$

■

Maximum Principles

Definition 20. Suppose Ω is an open, connected subset of $\mathbb{R}^n$. A real-valued function $u\colon \Omega \to \mathbb{R}$ is said to be **harmonic** if it satisfies $\Delta u = 0$ in Ω.

Harmonic functions have a remarkable property.

Result 21 (Mean Value Principle)**.** Suppose $\Omega \subset \mathbb{R}^n$ is open and connected, and that u is harmonic in Ω and piecewise continuous on $\partial\Omega$. Then for each $x_0 \in \Omega$ and for each ball B centered at x_0 of radius $r > 0$ such that $\overline{B} \subset \Omega$, the value $u(x_0)$ is the average of values on ∂B (and in B):

$$u(x_0) \;=\; \frac{\displaystyle\int_{\partial B} u\, d\sigma}{\displaystyle\int_{\partial B} d\sigma} \;=\; \frac{\displaystyle\int_{B} u\, d\mathbf{x}}{\displaystyle\int_{B} d\mathbf{x}}\,.$$

Proof. [Sketch.] Let $a(r)$ be the average value of u on the sphere $S = \partial B$ centered at x_0:

$$a(r) \;:=\; \frac{\displaystyle\int_S u\, d\sigma}{\displaystyle\int_S d\sigma}\,. \qquad \text{Then} \qquad a'(r) \;=\; \cdots \;=\; \frac{\displaystyle\int_B \Delta u\, d\mathbf{x}}{\displaystyle\int_S d\sigma} \;=\; 0\,,$$

and so $a(r) = C$ (a constant). By continuity,

$$u(x_0) \;=\; \lim_{r\to 0} a(r) \;=\; C = a(r)$$

for any $r > 0$. □

The next result, a corollary to the *mean value principle*, indicates that one expects to find maximum and minimum values for harmonic functions on the boundary of the region Ω.

Theorem 22 (Maximum Principle)**.** An harmonic function on Ω (open, connected in $\mathbb{R}^n$) cannot attain a maximum (nor a minimum) in Ω unless it is constant.

Proof. [Idea.] Suppose u attains a maximum value M at $\mathbf{x}_0 \in \Omega$. Set $v(\mathbf{x}) = M - u(\mathbf{x})$. Note that $\Delta v = -\Delta u = 0$, so v is harmonic in Ω, and $v(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \Omega$. Let $r > 0$ be such that the ball B of radius r centered at x_0 lies entirely inside Ω. Then

$$0 \;=\; v(\mathbf{x}_0) \;=\; \frac{\displaystyle\int_B v\, d\mathbf{x}}{\displaystyle\int_B d\mathbf{x}}\,,$$

showing that $v(\mathbf{x}) = 0$ throughout B (and, hence, $u(\mathbf{x}) \equiv M$ for $\mathbf{x} \in B$). □

The *maximum principle* may be used to deduce uniqueness of solutions to Poisson's equation.

Example 28: Uniqueness of Solution for Poisson's Equation with Dirichlet BCs

Suppose Ω is an open, bounded, connected region of $\mathbb{R}^n$. Consider the Poisson problem with Dirichlet boundary conditions

$$\Delta u = f, \quad \mathbf{x} \in \Omega, \qquad \text{subject to} \qquad u(\mathbf{x}) = g(\mathbf{x}) \ \text{ for } \ x \in \partial\Omega.$$

Suppose u, v both solve this problem, and let $w := u - v$. Then,

$$\Delta w \ = \ \Delta u - \Delta v \ = \ f - f \ = \ 0 \, ,$$

showing that w is harmonic in Ω. By the maximum principle,

$$\min_{\overline{\Omega}} w \ = \ \min_{\partial\Omega} w \ = \ 0 \ = \ \max_{\partial\Omega} w \ = \ \max_{\overline{\Omega}} w \, .$$

Said another way, $u \equiv v$ in $\overline{\Omega}$.

■

Notes:

- It is also possible to show that any solution of Poisson's equation subject to mixed inhomogeneous (Dirichlet on part, Neumann on the rest) BCs is unique. Solutions to Poisson's equation subject to Neumann BCs are only unique up to an additive constant.
- Proving existence of a solution to Poisson's equation is more difficult. A sufficient condition (see the 1st paragraph of [Olver], p. 217) is that $f \in C^1(\Omega)$.

There is a version of the maximum principle for the diffusion equation—the evolutionary equation for which Poisson's equation represents the equilibrium problem.

Theorem 23 (Maximum Principle for Diffusion Equations)**.** Consider the forced heat equation

$$u_t \ = \ \gamma u_{xx} + F(t,x), \quad a < x < b, \quad t > 0$$

($\gamma > 0$ constant). Assume the source term is nowhere positive $F(t,x) \leq 0$ for all $(t,x) \in R = [a,b] \times [0,c]$. Then the global maximum (and minimum) of $u(t,x)$ on the domain R occurs either at $t = 0$ or $x = a$ or $x = b$.

As a corollary to this result, we have the following:

Corollary 24. Suppose $u(t,x)$ solves the heat equation (with $\gamma > 0$)

$$u_t = \gamma u_{xx}, \quad a \le x \le b, \quad 0 \le t \le c.$$

Let m and M be, respectively, the minimum and maximum values for the initial and boundary temperatures—that is,

$$m \ \le \ u(t,x) \ \le \ M \quad \text{for} \quad (t,x) \in \Big\{(0,x)\,\Big|\,a \le x \le b\Big\} \cup \Big\{(t,a)\,\Big|\,0 \le t \le c\Big\} \cup \Big\{(t,b)\,\Big|\,0 \le t \le c\Big\} \ .$$

Then $m \le u(t,x) \le M$ for all (t,x) in the rectangle $[a,b] \times [0,c]$.

Well-Posed Problems

Definition 25 (Jacques Hadamard, paper of 1923)**.** An initial/boundary value problem is said to be **well-posed** if it meets these criteria:

- a solution of the problem exists,
- there is no more than one solution, and
- the solution depends continuously on the initial and/or boundary data (a condition also known as **stability**). That is, a small change in initial/boundary data should yield a small change in solution.

Example 29: Cauchy Problem for Laplace's Equation is Unstable

Consider the problem

$$\Delta u = 0, \quad (x,y) \in \mathbb{R} \times (0,\infty), \qquad \text{subject to} \qquad u(x,0) = f(x), \ \ u_y(x,0) = g(x).$$

Notice that $u_0(x,y) \equiv 0$ satisfies this problem when $f(x) = g(x) = 0$. Consider the related problems in which

$$f(x) = f_n(x) = \frac{1}{n}\cos(nx), \qquad g(x) = g_n(x) \equiv 0 \ ,$$

which produce solutions $u_n(x,y) = \frac{1}{n}\cos(nx)\cosh(ny)$. The larger n, the closer our BCs are to the homogeneous ones. Nevertheless, for any fixed n,

$$u_n(0,y) \ = \ \frac{\cosh(ny)}{n} \ \to \ \infty \quad \text{as} \quad y \to \infty.$$

■

It is left to homework to show that Poisson's equation is stable on a bounded domain Ω.

Example 30: Backwards Heat Problem

Consider the heat problem

$$u_t = u_{xx}, \quad 0 < x < \pi, \quad 0 < t < 1, \quad \text{subject to homog. Dirichlet BCs:} \quad u(t,0) = 0 = u(t,\pi).$$

Suppose, also, that while we presume $w(x) := u(0,x) \in L^2(0,\pi)$, we have the temperature profile at $t = 1$: $u(1,x) = f(x)$.

Note: Up to the application of a temporal snapshot of the temperature profile (previously assumed to be available at $t = 0$), previous work separating variables is relevant:

$$u(t,x) \;=\; \sum_{n=1}^{\infty} c_n e^{-n^2 t} \sin(nx) \; .$$

Employing that $u(1,x) = f(x)$, we take the $L^2(0,\pi)$ inner product with $sin(m\cdot)$ to get

$$\langle f, \sin(m\cdot)\rangle \;=\; \left\langle \sum_{n=1}^{\infty} c_n e^{-n^2} \sin(n\cdot), \sin(m\cdot) \right\rangle \;=\; \cdots = c_m e^{-m^2} \| \sin(m\cdot)\|_2^2 \; ,$$

giving that

$$c_n \;=\; e^{m^2} \frac{\langle f, \sin(m\cdot)\rangle}{\| \sin(m\cdot)\|_2^2} \;=\; \frac{2}{\pi} e^{m^2} \langle f, \sin(m\cdot)\rangle \; .$$

Claim 26. This problem is unstable.

Proof. Clearly if we take $f \equiv 0$, we get each $c_n = 0$, and hence the zero soln $u \equiv 0$. Now fix N and let $f(x) = \frac{1}{N}\sin(Nx)$. Clearly we can make $\|f\|_\infty$ (or $\|f\|_2$) as small as we like by making N large. We have each $c_n = 0$, $n \neq N$, so the corresponding solution of the heat problem is

$$u(t,x;N) \;=\; \frac{1}{N} e^{N^2(1-t)} \sin(Nx) \quad \Rightarrow \quad \|w(x)\|_\infty \;=\; \frac{1}{N} e^{N^2} \; .$$

□

The 2nd Law of Thermodynamics (Clausius) says

> A transformation whose only final result is to transfer heat from a body at a given temperature to a body at a higher temperature is impossible; i.e., the transfer is only possible at the expense of some organizational effort."

Based on this, Hadamard concluded problems like the backward heat equation are not physical. Quoting Poincaré, he said "The physical world not only provides us with problems to solve, . . . it also suggests to us the solutions" (i.e., if we can/should solve them, and how to do so). By such edicts from an influential mathematician (Hadamard proved the **prime number theorem**, which describes the asymptotic distribution of prime numbers), many shied away from such problems.

■

Some other examples of ill-posed problems:

- **Sideways heat equation**. Imagine trying to find out the temperature on the heat shield of the space shuttle during re-entry. A sensor on the surface would quickly burn up. Nevertheless, the use of heat readings from a sensor embedded under the shield to recover temperatures on the surface is an ill-posed problem.
- **Gravitational intensity problem**. One wishes to use readings of gravitational intensity at the surface of the earth to discover locations of ore deposits in the earth, another ill-posed problem.
- **Seismic exploration of marine oil deposits**. One generates sound waves and listens to the echo. Using knowledge of how various materials reflect sound waves, one wishes to detect areas of likely oil deposits.

Example 31: Uniqueness of Solution for 1D Dirichlet Heat Problem, Bounded Domain

Consider the forced heat problem

$$u_t = \gamma u_{xx} + f(t,x), \quad 0 < x < \ell, \quad t > 0, \quad \text{subject to} \quad \begin{cases} \text{Dirichlet BCs:} & u(t,0) = \alpha(t), \\ & u(t,\ell) = \beta(t), \\ \text{IC:} & u(0,x) = \varphi(x). \end{cases} \tag{25}$$

Suppose u, v are both solutions of (25), and set $w = u - v$. Then w satisfies the related (to (25)) problem

$$w_t = \gamma w_{xx}, \quad 0 < x < \ell, \quad t > 0, \quad \text{subject to} \quad \begin{cases} \text{Dirichlet BCs:} & w(t,0) = 0, \\ & w(t,\ell) = 0, \\ \text{IC:} & w(0,x) = 0. \end{cases} \tag{26}$$

By Corollary 24, $w \equiv 0$, meaning that $u = v$.

An alternate argument (Don't do! It is assigned for HW). We will show $w(t,x) \equiv 0$ using the following *energy* argument. Define

$$E(t) := \int_0^\ell w^2(t,x)\,dx\,.$$

Then

$$E'(t) = \frac{d}{dt}\int_0^\ell w^2(t,x)\,dx = \int_0^\ell \frac{\partial}{\partial t} w^2(t,x)\,dx = 2\int_0^\ell w w_t\,dx$$
$$= 2\gamma\int_0^\ell w w_{xx}\,dx = 2\gamma w w_x\Big|_{x=0}^{x=\ell} - 2\gamma\int_0^\ell w_x^2\,dx = -2\gamma\int_0^\ell w_x^2\,dx \le 0\,.$$

Thus

$$0 \le \int_0^\ell w^2(t,x)\,dx = E(t) \le E(0) = \int_0^\ell w^2(0,x)\,dx = \int_0^\ell 0\,dx = 0$$

for all $t > 0$. Thus $0 \equiv w(t,x) = u(t,x) - v(t,x)$ for all (t,x).

■

Inhomogeneous Problems

Many problems, as initially posed, do not yield themselves to the technique of *separation of variables*. Certain "tricks of the trade" must be applied first. Here is a short list of some scenarios and standard tricks that are used.

1. **Nonhomogeneous models**.
   Consider the problem

   $$u_t = A[u] + F(t,x)\,, \tag{27}$$

   with zero boundary/initial conditions. As for linear ODEs, it is reasonable to consider the associated homogeneous model

   $$u_t = A[u]\,, \tag{28}$$

   again with zero BCs/ICs.

   Eigenfunction expansion. This technique involves

   - first finding eigenfns $\{v_n(\cdot)\}_{n=1}^\infty$ associated with homogeneous problem (28)
   - expanding fns of (t,x) in series with time-varying coefficients:

   $$s(t,x) = \sum_{n=1}^\infty s_n(t)v_n(x) \qquad \text{with} \qquad s_n(t) = \frac{\langle s(t,\cdot), v_n\rangle}{\|v_n\|_2^2}\,.$$

   Example 32:

Consider the heat problem

$$u_t - \gamma u_{xx} = f(t,x), \quad 0 < x < \pi, \quad t > 0, \quad \text{subject to} \quad \begin{cases} \text{BCs:} & u(t,0) = 0 = u(t,\pi), \\ \text{ICs:} & u(0,x) = 0. \end{cases}$$

Handling the homogeneous version of the problem, previous work separating variables tells us to expand in a series of eigenfns $v_n(x) = \sin(nx)$ corresponding to eigenvalues $\lambda_n = -n^2$, $n = 1, 2, \ldots$. Now assume

$$f(t,x) \;=\; \sum_{n=1}^{\infty} f_n(t) \sin(nx)\,, \qquad \text{with} \qquad f_n(t) \;=\; \frac{2}{\pi} \int_0^{\pi} f(t,x) \sin(nx)\, dx\,,$$

and

$$u(t,x) \;=\; \sum_{n=1}^{\infty} g_n(t) \sin(nx)\,,$$

with the $g_n(\cdot)$ to be determined, so that

$$\frac{\partial}{\partial t} u(t,x) \;=\; \sum_{n=1}^{\infty} g_n'(t) \sin(nx) \qquad \text{and} \qquad \frac{\partial^2}{\partial x^2} u(t,x) \;=\; -\sum_{n=1}^{\infty} n^2 g_n(t) \sin(nx)\,.$$

Inserting these into the appropriate expressions of our PDE, we get

$$\sum_{n=1}^{\infty} g_n'(t) \sin(nx) + \gamma \sum_{n=1}^{\infty} n^2 g_n(t) \sin(nx) \;=\; \sum_{n=1}^{\infty} f_n(t) \sin(nx)\,.$$

Collecting coefficients of the independent eigenfns, we get a sequence of linear ODEs

$$g_n'(t) + n^2 \gamma g_n(t) \;=\; f_n(t)\,, \quad \text{with soln} \quad g_n(t) \;=\; g(0) + \int_0^t f_n(\tau) e^{-n^2 k(t-\tau)}\, d\tau\,,$$

for $n = 1, 2, \ldots$. Note that, since

$$0 \;=\; u(0,x) \;=\; \sum_{n=1}^{\infty} g_n(0) \sin(nx)\,,$$

we have $g_n(0) = 0$ for each n.

■

2. **Nonhomogeneous BCs**.
Now suppose we have the unforced problem

$$u_t = A[u], \quad 0 < x < \ell, \quad t > 0, \quad \text{subject to} \quad \begin{cases} \text{BCs:} & u(t,0) = \alpha(t), \\ & u(t,\ell) = \beta(t), \\ \text{IC:} & u(0,x) = \phi(x). \end{cases}$$

Let

$$u^*(t,x) = \frac{x}{\ell}\beta(t) + \frac{\ell - x}{\ell}\alpha(t) = \alpha(t) + \frac{\beta(t) - \alpha(t)}{\ell}\, x \tag{29}$$

Note: In cases where $\alpha(t)$, $\beta(t)$ are simply constants, $u^*(t,x)$ is the **equilibrium solution**, or solution of the corresponding static problem $A[u] = 0$ with these BCs. Set $v(t,x) = u(t,x) - u^*(t,x)$. Notice that

$$\left.\begin{array}{rcl} v(t,0) &=& u(t,0) - u^*(t,0) \;=\; \alpha(t) - \alpha(t) \;=\; 0, \\ v(t,\ell) &=& u(t,\ell) - \beta(t) \;=\; 0, \\ v(0,x) &=& u(0,x) - u^*(0,x) \;=\; \phi(x) - \left[\frac{x}{\ell}\,\beta(0) + \frac{\ell - x}{\ell}\,\alpha(0)\right]. \end{array}\right\} \tag{30}$$

If, further, we assume that the operator A annihilates linear functions (i.e., that u^* is in the kernel of A, as is the case when $A = \partial^2/\partial x^2$), then we have $A[v] = A[u]$, which gives that

$$v_t \;=\; u_t - u_t^* \;=\; A[u] - h(t,x) \;=\; A[v] - h(t,x),$$

where $h(t,x) = \frac{\partial u^*}{\partial t}(t,x)$. Thus, we have exchanged a problem with nonzero BCs for one with a forcing term.

Based upon 1 and 2 above, the following paradigm is proposed for solving evolutionary (and may be adapted for static) problems. Suppose A is a linear differential operator that annihilates linear functions is x (so $A[u^*] = 0$, for u^* in (29). Given the problem

$$u_t = A[u] + f(t,x), \quad 0 < x < \ell, \quad t > 0, \quad \text{subject to} \quad \left\{\begin{array}{rl} \text{BCs:} & u(t,0) = \alpha(t), \\ & u(t,\ell) = \beta(t), \\ \text{IC:} & u(0,x) = \varphi(x), \end{array}\right. \tag{31}$$

we set $v = u - u^*$. Then v satisfies

$$v_t \;=\; u_t - u_t^* \;=\; A[u] - u_t^* \;=\; A[v] + F(t,x)\ ,$$

along with BC/ICs (30), where $F(t,x) = f(t,x) - \partial u^*/\partial t$. This problem in v we split into two problems

$$v_t \;=\; A[v] + F(t,x), \quad 0 < x < \ell, \quad t > 0, \quad \text{subject to} \quad \left\{\begin{array}{rl} \text{BCs:} & v(t,0) = 0 = v(t,\ell), \\ \text{IC:} & v(0,x) = 0, \end{array}\right. \tag{32}$$

and

$$v_t \;=\; A[v], \quad 0 < x < \ell, \quad t > 0, \quad \text{subject to} \quad \left\{\begin{array}{rl} \text{BCs:} & v(t,0) = 0 = v(t,\ell), \\ \text{IC:} & v(0,x) = \phi(x) - u^*(0,x). \end{array}\right. \tag{33}$$

The sum of solutions to (32) and (33) is a function that solves (31).

Bessel Functions in Separation of Variables

Consider the vibrations of a circular membrane (drum), modeled by the equation

$$\frac{\partial^2 u}{\partial t^2} \;=\; c^2 \Delta u.$$

The geometry suggests polar coordinates, or that $u = u(t, r, \theta)$. We assume the membrane is fixed at the circular boundary, and the initial position and velocity are given:

$$u(t, a, \theta) = 0, \qquad \begin{aligned} u(t, r, -\pi) &= u(t, r, \pi), \\ u_t(t, r, -\pi) &= u_t(t, r, \pi), \end{aligned} \qquad \text{and} \qquad \begin{aligned} u(0, r, \theta) &= f(r, \theta), \\ u_t(0, r, \theta) &= g(r, \theta). \end{aligned}$$

Solution: After rescaling, we may take $a = 1$. We first assume $u(x, t) = T(t)v(r, \theta)$. Then

$$u_{tt} = c^2 \left[\frac{1}{r} \frac{\partial}{\partial r}(r u_r) + \frac{1}{r^2} u_{\theta\theta} \right] \qquad \text{becomes} \qquad T'' v = c^2 T \left(v_{rr} + \frac{1}{r} v_r + \frac{1}{r^2} v_{\theta\theta} \right),$$

or

$$\frac{T''}{c^2 T} = \frac{v_{rr} + r^{-1} v_r + r^{-2} v_{\theta\theta}}{v} = \lambda.$$

This leads to the two "simpler" DEs

$$T'' = \lambda c^2 T\,, \tag{34}$$

$$v_{rr} + \frac{1}{r} v_r + \frac{1}{r^2} v_{\theta\theta} = \lambda v\,, \qquad v(1, \theta) = 0, \quad v(r, \pi) = v(r, -\pi), \quad v_\theta(r, \pi) = v_\theta(r, -\pi). \tag{35}$$

The space of functions for which we hope to obtain a complete orthogonal basis is $L^2(\mathbb{D})$, where $\mathbb{D}$ is the disc centered at the origin of radius 1. To show eigenvalues are real and nonpositive, we employ the inner product of that space. Specifically, if (λ, v) form an eigenpair of our 2-dimensional Laplacian operator L, then

$$\begin{aligned} \lambda \|v\|^2 = \langle \lambda v, v \rangle = \langle L[v], v \rangle &= \int_0^{2\pi} \int_0^1 (v_{rr} + r^{-1} v + r^{-2} v_{\theta\theta}) v \, r \, dr \, d\theta \\ &= \int_0^{2\pi} \int_0^1 v_{rr}(rv) \, dr \, d\theta + \int_0^{2\pi} \int_0^1 v_r v \, dr \, d\theta + \int_0^1 r^{-1} \int_0^{2\pi} v_{\theta\theta} v \, d\theta \, dr \\ &= \int_0^{2\pi} \left\{ \Big[r v_r v \Big]_0^a - \int_0^1 v_r (r v_r + v) \, dr + \int_0^1 v_r v \, dr \right\} d\theta + \int_0^1 r^{-1} \left\{ \Big[v_\theta v \Big]_0^{2\pi} - \int_0^{2\pi} v_\theta^2 \, d\theta \right\} dr \\ &= -\int_0^{2\pi} v_r^2 \, r \, dr \, d\theta - \int_0^1 \int_0^{2\pi} r^{-1} v_\theta^2 \, d\theta \, dr \leq 0. \end{aligned}$$

Moreover, if $\lambda = 0$ it is clear that v_r and v_θ are zero throughout $\mathbb{D}$, which implies v is constant in $\mathbb{D}$. But since v is zero on the boundary $r = a$, it must be zero throughout $\mathbb{D}$. Thus, the only eigenvalues $\lambda < 0$. Let us write $\lambda = -\alpha^2$ for $\alpha > 0$.

Now if we assume $v(r, \theta) = p(r)q(\theta)$, we get

$$p'' q + \frac{1}{r} p' q + \frac{1}{r^2} p q'' + \alpha^2 p q = 0, \qquad \text{or} \qquad \frac{r^2 p'' + r p'}{p} + \alpha^2 r^2 = -\frac{q''}{q} = \mu.$$

That is, we have simplified (35) further

$$q'' = -\mu q, \qquad \text{subject to periodic BCs} \qquad q(-\pi) = q(\pi), \quad q'(-\pi) = q'(\pi), \tag{A}$$

$$r^2 p'' + r p' + \alpha^2 r^2 p = \mu p, \qquad \text{subject to} \qquad p(1) = 0. \tag{B}$$

By previous work, we have that the BVP for q results in eigenvalues $\mu_m = m^2$, $m = 0, 1, 2, \ldots$, with corresponding eigenfunctions

$$\{\cos(m\theta)\}_{m=0}^{\infty} \qquad \text{and} \qquad \{\sin(m\theta)\}_{m=1}^{\infty}.$$

With these choices of μ_m, our ODE (A) (one for each m) is $r^2p'' + rp' + \alpha^2 r^2 p = m^2 p$, or

$$\frac{d}{dr}(rp') + \left(\alpha^2 r - \frac{m^2}{r}\right)p = 0. \tag{36}$$

Now let $z = r\alpha$. Then

$$p' = \frac{d}{dr}p = \left(\frac{d}{dz}p\right)\frac{dz}{dr} = \alpha\frac{dp}{dz} \qquad \Rightarrow \qquad rp' = z\frac{dp}{dz},$$

and so

$$\frac{d}{dr}(rp') = \frac{d}{dz}\left(z\frac{dp}{dz}\right)\frac{dz}{dr} = \left(\frac{dp}{dz} + z\frac{d^2p}{dz^2}\right)\alpha.$$

Thus, (36) becomes

$$\left(z\frac{d^2p}{dz^2} + \frac{dp}{dz}\right)\alpha + \left(z\alpha - m^2\alpha z^{-1}\right)p = 0, \qquad \text{or} \qquad z^2\frac{d^2p}{dz^2} + z\frac{dp}{dz} + (z^2 - m^2)p = 0.$$

where the latter is obtained from the former via multiplying through by $z\alpha$. It (the latter) is a DE known as the m^{th} **order Bessel equation** (see Chapter 5, perhaps Section 5.7, of Boyce & DiPrima). It has non-constant coefficients, and so is not solvable via methods covered in MATH 231 (unless instructor covers *series solutions* methods like those of Chapter 5, B & D). As a 2nd order ODE, it has two independent solutions which, for $m = 0, 1, 2, \ldots$ (i.e., m a nonnegative integer), are $J_m(z)$, called the **Bessel function of the first kind of m^{th} order**, and $Y_m(z)$, called the **Bessel function of the second kind of m^{th} order**. The general solution is, thus,

$$p(r) = c_1 J_m(z) + c_2 Y_m(z) = c_1 J_m(r\alpha) + c_2 Y_m(r\alpha).$$

These two types of functions may be plotted in Octave using commands like

```
xs = 0:.01:30;
plot(xs, besselj(2, xs)      % Bessel fn. of 1st kind of order 2
plot(xs, bessely(0, xs)      % Bessel fn. of 2nd kind of order 0
```

Note that the Bessel fns of second kind satisfy $|Y_m(z)| \to \infty$ as $z \to 0$. With $u(t, r, \theta) = T(t)p(r)q(\theta)$, we see that an implicit condition we should impose is boundedness of u at the origin, which implies boundedness of p at $r = 0$. We get this only by taking $c_2 = 0$. Imposing the condition $p(1) = 0$, we get

$$J_m(\alpha) = 0\,,$$

which means α is a (positive) zero of $J_m(\cdot)$, of which there are countably infinitely many. Write α_{mn}, $n = 1, 2, \ldots$, for these zeros. (Olver writes them as ζ_{mn}.) Then there is a problem of the form (B) for

each $\mu = \mu_m$, $m = 0, 1, 2, \ldots$ and, for each fixed m (B) has infinitely many solutions $p_{mn}(r) = J_m(\alpha_{mn} r)$, $n = 1, 2, \ldots$.

To summarize the work thus far, problem (35) has a doubly-infinite collection of eigenvalues $\lambda = \lambda_{mn} = -\alpha_{mn}^2$ for $m = 0, 1, 2, \ldots$, $n = 1, 2, \ldots$. For the case $m = 0$, there is one independent eigenmode

$$v_{0n}(r, \theta) = J_0(\alpha_{0n} r), \qquad \text{for each} \quad n = 1, 2, \ldots,$$

while for $m > 0$, each λ_{mn}, $n = 1, 2, \ldots$, yields two independent eigenmodes

$$v_{mn}(r, \theta) = J_m(\alpha_{mn} r) \cos(m\theta) \qquad \text{and} \qquad \tilde{v}_{mn}(r, \theta) = J_m(\alpha_{mn} r) \sin(m\theta)\,.$$

Now problem (34) must be solved for $\lambda_{mn} = -\alpha_{mn}^2$:

$$T'' + c^2 \alpha_{mn}^2 T = 0\,, \qquad \text{has independent solns} \qquad \cos(c\alpha_{mn} t) \qquad \text{and} \qquad \sin(c\alpha_{mn} t)\,.$$

Thus, we finally arrive at a solution (modula ICs)

$$\begin{aligned} u(t, r, \theta) = & \sum_{n=1}^{\infty} [a_{0,n} \cos(c\alpha_{0,n} t) + c_{0,n} \sin(c\alpha_{0,n} t)] J_0(\alpha_{0,n} r) \\ & + \sum_{m,n=1}^{\infty} \Big\{ [a_{m,n} \cos(c\alpha_{m,n} t) + c_{m,n} \sin(c\alpha_{m,n} t)] \cos(m\theta) \\ & \qquad + [b_{m,n} \cos(c\alpha_{m,n} t) + d_{m,n} \sin(c\alpha_{m,n} t)] \sin(m\theta) \Big\} J_m(\alpha_{m,n} r) \end{aligned} \tag{37}$$

Note that, in the *radially symmetric case* (when $u = u(t, r)$ is independent of θ), we have the simpler general solution

$$u(t, r) = \sum_{n=1}^{\infty} [a_{0,n} \cos(c\alpha_{0,n} t) + c_{0,n} \sin(c\alpha_{0,n} t)] J_0(\alpha_{0,n} r)\,. \tag{38}$$

To find the coefficients in (37), we have

$$f(r, \theta) = u(0, r, \theta) = \sum_{n=1}^{\infty} J_0(\alpha_{0,n} r) + \sum_{m,n=1}^{\infty} [a_{m,n} \cos(m\theta) + b_{m,n} \sin(m\theta)] J_m(\alpha_{m,n} r),$$

so taking the $L^2(\mathbb{D})$ inner product with various eigenmodes v_{mn} (or $\tilde{v}_{mn}$) yields

$$a_{0,k} = \frac{\langle f(\cdot,\cdot), J_0(\alpha_{0,k}\,\cdot)\rangle}{\|J_0(\alpha_{0,k}\,\cdot)\|_2^2} = \frac{\displaystyle\int_{-\pi}^{\pi}\int_0^1 f(r,\theta) J_0(\alpha_{0,k} r)\, r\, dr\, d\theta}{\pi\, J_1^2(\alpha_{0,n})}\,, \qquad k = 1, 2, \ldots,$$

$$a_{m,k} = \frac{\langle f(\cdot,\cdot), J_m(\alpha_{m,k}\,\cdot)\cos(m\,\cdot)\rangle}{\|J_m(\alpha_{m,k}\,\cdot)\|_2^2} = \frac{\displaystyle\int_{-\pi}^{\pi}\int_0^1 f(r,\theta) J_m(\alpha_{m,k} r)\cos(m\theta)\, r\, dr\, d\theta}{(\pi/2)\, J_{m+1}^2(\alpha_{m,k})}\,, \qquad m, k = 1, 2, \ldots,$$

$$b_{m,k} = \frac{\langle f(\cdot,\cdot), J_m(\alpha_{m,k}\,\cdot)\sin(m\,\cdot)\rangle}{\|J_m(\alpha_{m,k}\,\cdot)\|_2^2} = \frac{\displaystyle\int_{-\pi}^{\pi}\int_0^1 f(r,\theta) J_m(\alpha_{m,k} r)\sin(m\theta)\, r\, dr\, d\theta}{(\pi/2)\, J_{m+1}^2(\alpha_{m,k})}\,, \qquad m, k = 1, 2, \ldots.$$

To get the other coefficients, we first note that

$$\begin{aligned} u_t(t,r,\theta) &= c\sum_{n=1}^{\infty} \alpha_{0,n} J_0(\alpha_{0,n} r)\left[c_{0,n}\cos(c\alpha_{0,n}t) - a_{0,n}\sin(c\alpha_{0,n}t)\right] \\ &\quad + c\sum_{m,n=1}^{\infty} \alpha_{m,n}\left\{\left[c_{m,n}\cos(c\alpha_{m,n}t) - a_{m,n}\sin(c\alpha_{m,n}t)\right]\cos(m\theta)\right. \\ &\qquad \left. + \left[d_{m,n}\cos(c\alpha_{m,n}t) - b_{m,n}\sin(c\alpha_{m,n}t)\right]\sin(m\theta)\right\} J_m(\alpha_{m,n} r) \end{aligned}$$

Thus,

$$\begin{aligned} g(r,\theta) &= u_t(0,r,\theta) \\ &= c\sum_{n=1}^{\infty} c_{0,n}\alpha_{0,n} J_0(\alpha_{0,n} r) + c\sum_{m,n=1}^{\infty} \alpha_{m,n} J_m(\alpha_{m,n} r)\left[c_{m,n}\cos(m\theta) + d_{m,n}\sin(m\theta)\right]. \end{aligned}$$

Taking inner products on both sides with the various eigenmodes yields

$$c_{0,k} = \frac{\displaystyle\int_{-\pi}^{\pi}\int_0^1 g(r,\theta) J_0(\alpha_{0,k} r)\, r\, dr\, d\theta}{c\alpha_{0,k}\pi\, J_1^2(\alpha_{0,k})}, \qquad k = 1,2,\ldots,$$

$$c_{m,k} = \frac{\displaystyle\int_{-\pi}^{\pi}\int_0^1 g(r,\theta) J_m(\alpha_{m,k} r)\cos(m\theta)\, r\, dr\, d\theta}{c\alpha_{m,k}(\pi/2)\, J_{m+1}^2(\alpha_{m,k})}, \qquad m,k = 1,2,\ldots, \qquad \text{and}$$

$$d_{m,k} = \frac{\displaystyle\int_{-\pi}^{\pi}\int_0^1 g(r,\theta) J_m(\alpha_{m,k} r)\sin(m\theta)\, r\, dr\, d\theta}{c\alpha_{m,k}(\pi/2)\, J_{m+1}^2(\alpha_{m,k})}, \qquad m,k = 1,2,\ldots.$$

Finite Difference Approximations to 2nd Order Problems

Approximating derivatives

We have discussed the following finite difference approximations to derivatives of functions:

1^{st} derivatives: $f'(x) = \dfrac{f(x+h)-f(x)}{h} + O(h)$	called a *forward difference* when $h > 0$
$f'(x) = \dfrac{f(x)-f(x-h)}{h} + O(h)$	called a *backward difference* when $h > 0$
$f'(x) = \dfrac{f(x+h)-f(x-h)}{2h} + O(h^2)$	called a *centered difference*
2^{nd} derivatives: $f''(x) = \dfrac{f(x-h)-2f(x)+f(x+h)}{h^2} + O(h^2)$	another *centered difference*

During lecture

- Demonstrate order of convergence in Excel by computing various approximations to $f'(0)$ with $f(x) = e^x$. Start at $h = 1$, and keep halving this stepsize, looking at the ratio of current error to prior error.
- Show how the forward difference formula, along with the estimate of its truncation error as $O(h)$, arises from Taylor's Theorem with remainder.

Applying to Poisson's equation (elliptic PDEs)

Consider first the Dirichlet Poisson problem on a rectangle—that is,

$$-\Delta u = f, \quad \text{in } R = \Big\{(x,y)\,\Big|\, 0 < x < a, \quad 0 < y < b\Big\}, \quad \text{subject to} \quad u = g \quad \text{on } \partial R.$$

Assume a uniform partition of the rectangle in both the x and y-direction—that is,

$$\begin{aligned} &0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = a, \quad \text{with each} \quad x_j - x_{j-1} = \Delta x, \\ &0 = y_0 < y_1 < \cdots < y_M < x_{M+1} = b, \quad \text{with each} \quad y_m - y_{m-1} = \Delta y. \end{aligned}$$

For simplicity, let us take $\Delta x = \Delta y = h$. Let us denote our

$$\begin{aligned} &\text{approximation to } u(x_j, y_m) = u(jh, mh) \quad \text{as} \quad u_{jm}, \\ &\text{value of } f(x_j, y_m) = f(jh, mh) \quad\quad\quad\quad\;\; \text{as} \quad f_{jm}. \end{aligned}$$

Note that

$$\begin{aligned} u_{0,m} &= g(0, mh) &&= g(0, y_m), && m = 0, 1, \ldots, M+1, \\ u_{N+1,m} &= g((N+1)h, mh) &&= g(a, y_m), && m = 0, 1, \ldots, M+1, \\ u_{j,0} &= g(jh, 0) &&= g(x_j, 0), && j = 0, 1, \ldots, N+1, \\ u_{j,M+1} &= g(jh, (M+1)h) &&= g(x_j, b), && j = 0, 1, \ldots, N+1. \end{aligned}$$

Using a centered difference approximation for both u_{xx} and u_{yy}, we have

$$-\frac{u(x-h,y) - 2u(x,y) + u(x+h,y)}{h^2} - \frac{u(x,y-h) - 2u(x,y) + u(x,y+h)}{h^2} \;=\; f(x,y)\,,$$

or, applying this to the point (x_j, y_m),

$$4u_{j,m} - u_{j-1,m} - u_{j,m-1} - u_{j+1,m} - u_{j,m+1} = h^2 f_{j,m}\,, \qquad j = 1,\ldots,N,\ \ m = 1,\ldots,M.$$

This problem has the same number of constraints as unknowns, and is linear in those unknowns, and so we should formulate and solve it as a matrix problem $\mathbf{AU} = \mathbf{d}$. The most challenging aspect is lexicographic. We choose to set $U_k = u_{j,m}$ where $k = (m-1) * N + j$, with the result that the (MN)-by-(MN) coefficient matrix $\mathbf{A}$ is both sparse and is $(2N+1)$-banded:

$$\mathbf{A} = \begin{bmatrix}
4 & -1 & 0 & \cdots & 0 & -1 & 0 & \cdots & & & & & & & \\
-1 & 4 & -1 & 0 & \cdots & 0 & -1 & 0 & \cdots & & & & & & \\
0 & -1 & 4 & -1 & 0 & \cdots & 0 & -1 & 0 & \cdots & & & & & \\
 & & \ddots & \ddots & \ddots & & & & \ddots & & & & & & \\
\cdots & 0 & -1 & 0 & \cdots & 0 & -1 & 4 & -1 & 0 & \cdots & 0 & -1 & 0 & \cdots \\
 & & & & & & \ddots & \ddots & \ddots & & & & \ddots & & \\
 & & & & & \cdots & 0 & -1 & 0 & \cdots & 0 & -1 & 4 & -1 & 0 \\
 & & & & & & \cdots & 0 & -1 & 0 & \cdots & 0 & -1 & 4 & -1 \\
 & & & & & & & \cdots & 0 & -1 & 0 & \cdots & 0 & -1 & 4
\end{bmatrix}.$$

Actually, the form presented above is misleading. The entries in the first super and subdiagonals of $\mathbf{A}$ are not always (-1); there is an occasional zero, appearing every N^{th} entry. Here is a rudimentary Octave code which creates the correct form of $\mathbf{A}$:

```
superdiag = [];
for m = 1:M
  superdiag = [superdiag; ones(N-1, 1)];
  if (m != M)
    superdiag = [superdiag; 0];
  end
end
A = diag(4*ones(M*N, 1)) - diag(superdiag, -1) - diag(superdiag, 1) ...
     - diag(ones((M-1)*N, 1), -N) - diag(ones((M-1)*N, 1), N);
```

The right-hand side vector $\mathbf{d}$ is made up both of the inhomogeneity f and the (known) boundary values. Let us write

$$\begin{aligned} u(0,y) &= \ell BC(y), & \qquad u(x,0) &= bBC(x), \\ u(a,y) &= rBC(y), & \qquad u(x,b) &= tBC(x). \end{aligned}$$

Assuming that functions giving the BCs have been implemented in Octave in such a way that they all return column vectors, and there is a function implementing the imhomogenity $f(x,y)$ (so that it is able to accept matrix arguments for x and y, as when these inputs have been built using the `meshgrid()` command), the following Octave code excerpt will build an appropriate right-hand side vector $\mathbf{d}$:

```
[xs, ys] = meshgrid(0:h:a, 0:h:b);
d = h^2 * f(xs(2:M+1, 2:N+1), ys(2:M+1, 2:N+1))(:);
d(1:N) += bBC(xs(1, 2:N+1))';
d((M - 1)*N + (1:N)) += tBC(xs(1, 2:N+1))';
d(1:N:M*N) += lBC(ys(2:M+1, 1));
d(N:N:M*N) += rBC(ys(2:M+1, 1));
```

Example 33: Various Instances of the Dirichlet Poisson Problem on a Rectangle

- First we do several instances of the Dirichlet Laplace problem (i.e., $f \equiv 0$ in Poisson's equation) on the square $[0,1] \times [0,1]$ with just one boundary of the rectangle nonzero. The pair of scripts `psset1.m` and `poissonSolver.m` finds and graphs the solution with

$$\begin{aligned} u(0,y) &= 0, & u(x,0) &= x^3(1-x), \\ u(a,y) &= 0, & u(x,b) &= 0. \end{aligned}$$

  Switching to `psset3.m` yields the solution to the Laplace problem with

$$\begin{aligned} u(0,y) &= 0, & u(x,0) &= \begin{cases} x, & 0 \le x \le 1/2, \\ 1-x, & 1/2 < x \le 1, \end{cases} \\ u(a,y) &= 0, & u(x,b) &= 0. \end{aligned}$$

- The file `psset2.m` solves the Dirichlet Laplace problem on the rectangle (non-square) $[0,1/2] \times [0,1]$, with BCs

$$\begin{aligned} u(0,y) &= 4y, & u(x,0) &= 16x^2, \\ u(a,y) &= 4, & u(x,b) &= 4, \end{aligned}$$

  all of which are nonhomogeneous.

  Lecture highlight: My routine `poissonSolver.m` carries out the details of building and solving the linear system of equations. In its first implementation, I was working from the belief that both the first superdiagonal and subdiagonal of the matrix $\mathbf{A}$ consisted entirely of (-1)'s. This previous implementation is preserved in the file `psBad.m`. It is instructive to call that routine instead of `poissonSolver.m` from `psset2.m`. The resulting surface, the graph of what is supposed to be an harmonic function, has an interior maximum. Thus, awareness of theory (the *maximum principle*) informs us the routine has an error.

- Finally, in `psset4.m` we solve the Dirichlet Poisson problem with inhomogeneity

$$f(x,y) = [(3x + x^2)y(1-y) + (3y + y^2)x(1-x)]e^{x+y} ,$$

  and zero boundary conditions on all four sides of the rectangle. The exact solution in this instance is

$$u(x,y) = x(1-x)y(1-y)e^{x+y} .$$

Lecture highlight: Some features of Octave to mention include

- `meshgrid()` and `mesh()` functions
  These are just the kinds of plotting (and related) functions one would want in order to visualize a (numeric) solution on a grid.
- `view(azimuth, elevation)` function, relevant for 3D plots
- sparse matrices and matrix calculations
  The file `poissonSolver.m` contains code (originally commented out) which converts the coefficient matrix to a sparse one and allows for a comparison in solution time. You may wish to add a `spy(A)` command in order to get a picture of how sparse **A** is.
- While the method may be applied to the Poisson problem on non-rectangular domains, the details are more difficult to carry out. See [] for an example.

■

Finite Differences on Heat Problems (Parabolic PDEs)

Now consider the 1D heat equation

$$u_t = \gamma u_{xx}, \;\; 0 < x < \ell, \;\; t > 0, \quad \text{subject to} \quad \begin{cases} \text{BCs:} & u(t,0) = \alpha(t), \\ & u(t,\ell) = \beta(t), \\ \text{IC:} & u(0,x) = f(x). \end{cases} \tag{39}$$

For some fixed choice of Δt, Δx consider the uniform mesh

$$0 = t_0 < t_1 < t_2 < \cdots, \qquad 0 = x_0 < x_1 < \cdots < x_n = \ell,$$

with each $t_{j+1} - t_j = \Delta t$, $x_{m+1} - x_m = \Delta x$. Let us approximate t-derivatives with a forward difference approximation, and x-derivatives with a centered difference approximation. Denoting approximate values of $u(t_j, x_m)$ by u_{jm}, we have

$$\frac{u_{j+1,m} - u_{j,m}}{\Delta t} \;=\; \gamma \, \frac{u_{j,m-1} - 2u_{j,m} + u_{j,m+1}}{(\Delta x)^2},$$

or

$$u_{j+1,m} \;=\; \delta u_{j,m-1} + (1 - 2\delta) u_{j,m} + \delta u_{j,m+1}, \tag{40}$$

where $\delta = \gamma \Delta t / (\Delta x)^2$, a difference equation which holds for $j = 0, 1, 2, \ldots$, and $m = 1, 2, \ldots, n-1$. (See Problem 2 on Exam 1.) Let $\mathbf{u}^{(j)} = (u_{j,1}, u_{j,2}, \ldots, u_{j,n-1})$. Then $\mathbf{u}^{(0)}$ is given by the IC, while each

$$\mathbf{u}^{(j+1)} \;=\; \mathbf{A}\mathbf{u}^{(j)} + \mathbf{b}^{(j)}, \tag{41}$$

where

$$\mathbf{A} \;=\; \begin{bmatrix} 1-2\delta & \delta & & & & \\ \delta & 1-2\delta & \delta & & & \\ & \ddots & \ddots & \ddots & & \\ & & \delta & 1-2\delta & \delta & \\ & & & \delta & 1-2\delta & \end{bmatrix} \qquad \text{and} \qquad \mathbf{b}^{(j)} \;=\; \begin{bmatrix} \delta\alpha(t_j) \\ 0 \\ \vdots \\ 0 \\ \delta\beta(t_j) \end{bmatrix}.$$

Take note of where the IC/BCs figure in to these equations.

Example 34: Finite Difference Solution of Heat Equation

Solve the problem

$$u_t = u_{xx}, \quad 0 < x < 1, \quad t > 0, \quad \text{subject to} \quad \begin{cases} \text{BCs:} & u(t,0) = 0 = u(t,\ell), \\ \text{IC:} & u(0,x) = f(x), \end{cases}$$

where

$$f(x) \;=\; \begin{cases} -x, & 0 \le x \le 1/5, \\ x - 2/5, & 1/5 < x \le 7/10, \\ 1 - x, & 7/10 < x \le 1. \end{cases}$$

Note: Olver poses this problem first in Section 4.1 (see p. 123), and returns to it in Section 10.2, where he solves it using this same divided difference approach.

Solution. In class, use the OCTAVE script `heatSet1.m` that employs (41) (encoded in `fdHeatSolver.m`) to solve this heat problem with fixed $\Delta x = 0.1$ and several different time steps: $\Delta t = 0.01, 0.005$. Students may recall we had a condition on the size of time step when applying finite differences to the (1^{st} order) transport equation. In the above, finite difference solutions look meaningful only for the smaller of the two time steps.

■

Von Neumann Stability Analysis

A numerical algorithm is called **stable** if ...

The idea behind Von Neumann's method for analyzing the stability of a numerical algorithm is to:

- Consider a discrete eigenmode solution at the j^{th} time step.
- Use the algorithm to find what has happened to this solution at the $(j+1)^{\text{st}}$ time step. It will be a scalar multiple of the discrete eigenmode at time step j. Look at the scalar multiplier to see that it is less than 1 in absolute value. (Otherwise, the contribution of this eigenmode is growing in time.)

Let us apply this idea to the heat problem $u_t = \gamma u_{xx}$ on $(0,\pi) \times (0,\infty)$. With

- homogeneous Dirichlet BCs we get eigenfunctions $\sin x$, $\sin(2x)$, $\sin(3x)$, ...,
- homogeneous Neumann BCs we get eigenfunctions 1, $\cos x$, $\cos(2x)$, $\cos(3x)$, ...,

and resulting eigensolutions are time-varying rescalings of these. Both collections are both found in e^{ikx}, $k = 0, 1, \ldots$, one as the real part, the other as the imaginary.

Suppose the solution at some time step is a pure eigenmode e^{ikx}; that is, $u_{j,m} = e^{ikx_m}$. Then

$$\begin{aligned} u_{j+1,m} &= \delta u_{j,m-1} + (1-2\delta)u_{j,m} + \delta u_{j,m+1} = \delta e^{ikx_{m-1}} + (1-2\delta)e^{ikx_m} + \delta e^{ikx_{m+1}} \\ &= \delta e^{ik(x_m - \Delta x)} + (1-2\delta)e^{ikx_m} + \delta e^{ik(x_m+\Delta x)} = [\delta e^{-ik\Delta x} + (1-2\delta) + \delta e^{ik\Delta x}]e^{ikx_m} \\ &= \lambda u_{j,m} \, , \end{aligned}$$

where the multiplier λ satisfies

$$\begin{aligned} \lambda &= \delta e^{-ik\Delta x} + (1-2\delta) + \delta e^{ik\Delta x} \\ &= \delta[\cos(k\Delta x) - i\sin(k\Delta x)] + (1-2\delta) + \delta[\cos(k\Delta x) + i\sin(k\Delta x)] \\ &= 1 - 2[1-\cos(k\Delta x)]\delta = 1 - 4\delta\sin^2(k\Delta x/2) \, . \end{aligned}$$

Argue that stability requires $|\lambda| \le 1$, yielding **CFL condition**

$$\Delta t \le \frac{(\Delta x)^2}{2\gamma} \, .$$

An Implicit Algorithm

Let us alter the previous finite difference approach slightly, employing a backward difference approximation for u_t:

$$u_t(t,x) \approx \frac{u(t,x) - u(t-\Delta t, x)}{\Delta t} \, ,$$

but using the same centered difference approximation for u_{xx}. Inserting these finite differences at the point (t_j, x_m) yields the difference equation

$$\frac{u_{j,m} - u_{j-1,m}}{\Delta t} = \gamma \frac{u_{j,m-1} - 2u_{j,m} + u_{j,m+1}}{(\Delta x)^2} \, .$$

Making the change $j \mapsto j+1$ and doing some algebraic manipulations so as to get *unknowns* (at the $(j+1)^{\text{st}}$ time step) and *knowns* (at the j^{th} time step) on opposite sides of the equation, we have

$$-\delta u_{j+1,m-1} + (1+2\delta)u_{j+1,m} - \delta u_{j+1,m+1} = u_{j,m} \, , \tag{42}$$

where $\delta = \gamma\Delta t/(\Delta x)^2$ as before.

The previous finite difference approach (40) is said to be **explicit** because the (only) unknown is solved for explicitly in terms of known quantities. In contrast, (42) involves multiple unknowns, and so is an **implicit** scheme.

Taking (again) $\mathbf{u}^{(j)} = (u_{j,1}, u_{j,2}, \ldots, u_{j,n-1})$, we see that the implicit finite difference scheme (42) applied to the homogeneous Dirichlet heat problem (39) has matrix formulation

$$\mathbf{A}\mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} + \mathbf{b}^{(j)},$$

where

$$\mathbf{A} = \begin{bmatrix} 1+2\delta & -\delta & & & \\ -\delta & 1+2\delta & -\delta & & \\ & \ddots & \ddots & \ddots & \\ & & -\delta & 1+2\delta & -\delta \\ & & & -\delta & 1+2\delta \end{bmatrix} \quad \text{and} \quad \mathbf{b}^{(j)} = \begin{bmatrix} \delta\alpha(t_{j+1}) \\ 0 \\ \vdots \\ 0 \\ \delta\beta(t_{j+1}) \end{bmatrix}.$$

Example 35: Implicit Finite Difference Algorithm Solution of Heat Equation

Solve the problem

$$u_t = u_{xx}, \quad 0 < x < 1, \quad t > 0, \quad \text{subject to} \quad \begin{cases} \text{BCs:} & u(t,0) = 0 = u(t,\ell), \\ \text{IC:} & u(0,x) = f(x), \end{cases}$$

where

$$f(x) = \begin{cases} -x, & 0 \le x \le 1/5, \\ x - 2/5, & 1/5 < x \le 7/10, \\ 1 - x, & 7/10 < x \le 1. \end{cases}$$

Solution. In class, use the Octave script `impHeatSet1.m` that employs (42) (encoded in `implicitFDHeatSolver.m`) to solve this heat problem with fixed $\Delta x = 0.1$ and increasing choices for time steps: $\Delta t = 0.01, 0.05, 0.1$. Each of these choices yields a solution that appears to be meaningful.

■

Stability Analysis of Implicit Scheme: Heat Equation

Because it is a *known* quantity in (42) which is solved for, this time we will assume the solution at the $(j+1)^{\text{st}}$ time step is a pure eigenmode—that is, that for some fixed natural number k, $u_{j+1,m} = e^{ikx_m}$. Then, after (42), we have

$$\begin{aligned} u_{j,m} &= -\delta e^{ik(x_m - \Delta x)} + (1+2\delta)e^{ikx_m} - \delta e^{ik(x_m + \Delta x)} \\ &= \left[-\delta e^{-ik\Delta x} + (1+2\delta) - \delta e^{ik\Delta x}\right] e^{ikx_m} \\ &= \lambda u_{j+1,m}\,, \end{aligned}$$

where

$$\begin{aligned} \lambda &= -\delta(e^{-ik\Delta x} + e^{ik\Delta x}) + 1 + 2\delta = -2\delta\cos(k\Delta x) + 1 + 2\delta \\ &= 1 + 4\delta\left[\frac{1-\cos(k\Delta x)}{2}\right] = 1 + 4\delta\sin^2\left(\frac{k\Delta x}{2}\right) \ge 1 \end{aligned}$$

for all $\delta \geq 0$. Hence, the implicit scheme is *unconditionally stable*—i.e., there is no CFL condition—as

$$u_{j+1,m} = \frac{1}{\lambda}\, u_{j,m} \ ,$$

with $0 < \lambda^{-1} \leq 1$ for all choices of $\delta = \gamma \Delta t/(\Delta x)^2$.

Neumann Boundary Conditions

See the following files/exercises:

- `/Users/scofield/courses/333/exercises/tveito-Winther/exercises/ch02/e14_mod.tex`
- `/Users/scofield/courses/333/exercises/olver/c10/10.2.07.tex`

Finite Differences on Wave Problems (Hyperbolic PDEs)

Consider the one-dimensional wave problem

$$u_{tt} = c^2 u_{xx}, \ \ 0 < x < \ell, \ \ t > 0, \quad \text{subject to} \quad \begin{cases} \text{BCs:} & u(t,0) = \alpha(t), \\ & u(t,\ell) = \beta(t), \\ \text{ICs:} & u(0,x) = f(x), \\ & u_t(0,x) = g(x). \end{cases} \tag{43}$$

For fixed $\Delta t > 0$ and $\Delta x > 0$, we will assume a uniform grid

$$0 = t_0 < t_1 < t_2 < \ldots, \qquad 0 = x_0 < x_1 < x_2 < \ldots < x_n = \ell,$$

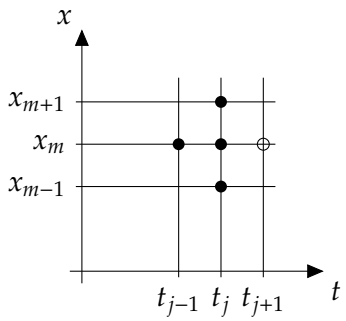
with each $t_j - t_{j-1} = \Delta t$ and each $x_m - x_{m-1} = \Delta x$. Let us apply centered difference approximations to the 2nd derivatives u_{tt}, u_{xx}. We write $u_{j,m}$ for our approximation to the solution $u(t_j, x_m)$, which satisfies

$$\frac{u_{j-1,m} - 2u_{j,m} + u_{j+1,m}}{(\Delta t)^2} \ = \ c^2 \, \frac{u_{j,m-1} - 2u_{j,m} + u_{j,m+1}}{(\Delta x)^2},$$

or, taking $\sigma = c\Delta t/\Delta x$,

$$u_{j+1,m} \ = \ \sigma^2 u_{j,m+1} + 2(1-\sigma^2)u_{j,m} + \sigma^2 u_{j,m-1} - u_{j-1,m}, \tag{44}$$

for $j = 1, 2, \ldots$, $m = 1, 2, \ldots, n-1$.



The computation molecule for (44) appears at right, with the node at (t_{j+1}, x_m) drawn as an open circle showing it is the one unknown when (44) is applied. This algorithm is rightly called an **explicit three-level scheme**. Again writing $\mathbf{u}^{(j)} = (u_{j,1}, u_{j,2}, \ldots, u_{j,n-1})$ for the vector of unknowns

at the j^{th} time step, we rewrite (44) in matrix form $\mathbf{u}^{(j+1)} = \mathbf{A}\mathbf{u}^{(j)} - \mathbf{u}^{(j-1)} + \mathbf{b}^{(j)}$, $j = 1, 2, \ldots$, where

$$\mathbf{A} = \begin{bmatrix} 2(1-\sigma^2) & \sigma^2 & 0 & \cdots & 0 \\ \sigma^2 & 2(1-\sigma^2) & \sigma^2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & & \sigma^2 \\ 0 & \cdots & 0 & \sigma^2 & 2(1-\sigma^2) \end{bmatrix}, \quad \text{and} \quad \mathbf{b}^{(j)} = \begin{bmatrix} \sigma^2\alpha(t_j) \\ 0 \\ \vdots \\ 0 \\ \sigma^2\beta(t_j) \end{bmatrix}.$$

Naturally, we take $\mathbf{u}^{(0)} = (f(x_1), f(x_2), \ldots, f(x_{n-1}))$ but, to get things going, we require something other than (44) to find the entries of $\mathbf{u}^{(1)}$. That is where the initial velocity $g(x)$ gets used. The simplest way to start off is to use a forward difference approximation

$$g(x_m) = u_t(0, x_m) \approx \frac{u(t_1, x_m) - u(0, x_m)}{\Delta t} = \frac{u(t_1, x_m) - f(x_m)}{\Delta t} \qquad \Rightarrow \qquad u_{1,m} = f(x_m) + g(x_m)\Delta t\,.$$

The drawback to this is that we are using $O((\Delta t)^2)$ and $O((\Delta x)^2)$ approximations for derivatives everywhere else in the scheme; an $O(\Delta t)$ forward difference approximation can potentially introduce much greater error than the other approximations, and would appear at the very start of the method!

If we assume the wave equation holds on the initial line where $t = 0$ and that we have sufficient differentiability, we may employ Taylor's theorem to write

$$\begin{aligned} u(\Delta t, x_m) &= u(0, x_m) + (\Delta t)u_t(0, x_m) + \frac{(\Delta t)^2}{2}u_{tt}(0, x_m) + \frac{(\Delta t)^3}{6}u_{ttt}(\tilde{t}, x_m) \quad (\text{with } 0 < \tilde{t} < \Delta t) \\ &= \phi(x_m) + g(x_m)\Delta t + \frac{c^2(\Delta t)^2}{2}u_{xx}(0, x_m) + O((\Delta t)^3) \\ &= \phi(x_m) + g(x_m)\Delta t + \frac{c^2(\Delta t)^2}{2}f''(x_m) + O((\Delta t)^3) \\ &= \phi(x_m) + g(x_m)\Delta t + \frac{c^2(\Delta t)^2}{2} \cdot \frac{f(x_{m-1}) - 2f(x_m) + f(x_{m+1})}{(\Delta x)^2} + O((\Delta x)^2) + O((\Delta t)^3)\,, \end{aligned}$$

where the last expression replaces its predecessor so that we do not need to numerically differentiate f. Thus, we take

$$u_{1,m} = \frac{\sigma^2}{2}f(x_{m-1}) + (1-\sigma^2)f(x_m) + \frac{\sigma^2}{2}f(x_{m+1}) + g(x_m)\Delta t\,, \qquad m = 1, 2, \ldots, n-1.$$

Numerically, one finds this algorithm behaves poorly if σ is too large, suggesting a CFL condition must be met. In fact, numerical stability requires that

$$\sigma := c\Delta t/\Delta x \leq 1, \tag{45}$$

which is equivalent to saying $c \leq \Delta x/\Delta t$, the same condition required for stability when solving the first order transport equation using finite differences. (For an justification of this CFL condition based on the *numerical domain of dependence*, see Olver, p. 459, or Stanoyevitch, pp. 548–549.)

Algorithm for solving (43): Choose Δt, Δx in such a manner that $\sigma := c\Delta t/\Delta x \leq 1$ and $n := \ell/\Delta x$ is an integer, and set $t_j = j\Delta t$, $x_m = m\Delta x$. Initialize the algorithm by setting $\mathbf{u}^{(0)} = (f(x_1), f(x_2), \ldots, f(x_{n-1}))$ and $\mathbf{u}^{(1)} = \frac{1}{2}\mathbf{A}\mathbf{u}^{(0)} + \frac{1}{2}\mathbf{b}^{(0)} + (\Delta t)\mathbf{g}$, where

$$\mathbf{A} = \begin{bmatrix} 2(1-\sigma^2) & \sigma^2 & 0 & \cdots & 0 \\ \sigma^2 & 2(1-\sigma^2) & \sigma^2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & & \sigma^2 \\ 0 & \cdots & 0 & \sigma^2 & 2(1-\sigma^2) \end{bmatrix}, \quad \mathbf{b}^{(j)} = \begin{bmatrix} \sigma^2\alpha(t_j) \\ 0 \\ \vdots \\ 0 \\ \sigma^2\beta(t_j) \end{bmatrix}, \quad \text{and} \quad \mathbf{g} = \begin{bmatrix} g(x_1) \\ g(x_2) \\ g(x_3) \\ \vdots \\ g(x_{n-1}) \end{bmatrix}.$$

Then, for $j = 1, 2, \ldots$, set $\mathbf{u}^{(j+1)} = \mathbf{A}\mathbf{u}^{(j)} - \mathbf{u}^{(j-1)} + \mathbf{b}^{(j)}$.

According to Stanoyevitch, p. 548, this algorithm "can be proved to converge to the exact solution of the wave problem (43) (as the partitions become more and more refined) provided that, in addition to the required differentiability assumptions, the ... CFL condition" (45) holds. The approximations which led to this algorithm are often applied to obtain algorithms for more general hyperbolic linear PDEs and, indeed, according to Stanoyevitch, are often applied to nonlinear problems as well.

Some General Comments on Finite Difference Methods

- Finite difference methods are easiest to apply in rectangular (or multidimensional rectangular) regions.
- In our discussion, we have applied finite difference methods to the simplest manifestations of the prototypical problems: Poisson's equation in two dimensions, the homogeneous one-dimensional heat equation, and the homogeneous one-dimensional wave equation. There is nothing to bar us from applying the same approaches to other representatives from their respective classes: elliptic, parabolic and hyperbolic (linear) PDEs, homogeneous and non-homogeneous. They are even successfully applied to nonlinear PDEs. Nevertheless, we should exercise caution, applying what we know of CFL conditions, keeping an eye out for numerical instability and, so far as we are able, checking that our solutions are reasonable. Regarding numerical methods for nonlinear PDEs, Stanoyevitch says

  > In general those that are based on conservation laws (physical principles) are the most successful. This seems to imply that a purely mathematical approach to the numerical solution of nonlinear PDEs is not sufficient; an additional requirement is a certain knowledge of the physical principles governing the phenomena that are modeled by the PDEs.[2]

[2] Stanoyevitch, p. 549.

Classifying PDEs

We have focused a great deal on the prototypical equations: Poisson's equation (of which Laplace's equation is the homogeneous version), the heat equation and the wave equation. These each are, respectively, representatives from three different families or classes of PDEs: elliptic PDEs, parabolic PDEs, and hyperbolic PDEs. Consider a PDE of the form

$$Au_{xx} + Bu_{xt} + Cu_{tt} + F(x, t, u, u_x, u_t) \;=\; 0\,, \tag{46}$$

where A, B, and C are constants, not all of which are zero. The **principal part** of this equation,

$$Au_{xx} + Bu_{xt} + Cu_{tt}\,,$$

is linear, and is the basis for classification. We define the **discriminant**

$$\Delta \;:=\; B^2 - 4AC\,,$$

and say that (46) is

- **elliptic** if $\Delta < 0$,
- **parabolic** if $\Delta = 0$, and
- **hyperbolic** if $\Delta > 0$.

The *names* for these classifications are based on the planar curves arising from algebraic equations of the form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F \;=\; 0\,,$$

which, in non-degenerate cases, are ellipses, parabolas and hyperbolas when $\Delta < 0$, $\Delta = 0$, and $\Delta > 0$ respectively. The *reason* for the classifications is that, "as it turns out, all parabolic equations are diffusion-like, all hyperbolic equations are weave-like, and all elliptic equations are static."[3]

Some remarks:

- In the static (elliptic) case, we have only spatial variables, and we would typically replace t in (46) with y.
- When the coefficients A, B, C are non-constant, depending on t, x, and perhaps even u (so the principal part is nonlinear), then Δ can change sign. We can still talk about regions where the problem is elliptic with $\Delta = \Delta(t, x) < 0$ (linear principal part case), etc.
- While we discuss these classifications in the context of two independent variables, "the terminology, underlying properties, and associated physical models carry over to second order partial differential equations in higher dimensions."[4]

[3] Logan, p. 45.
[4] Olver, p. 168.

- While first order PDEs do not fall into this discussion, their solutions exhibit wave-like behavior, and so they are usually grouped with hyperbolic PDEs.

Following Logan (see pp. 46–48), we propose a linear change of variables

$$\tau = at + bx, \quad \xi = ct + dx, \qquad \text{or} \qquad \begin{bmatrix} \tau \\ \xi \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix}, \tag{47}$$

assuming the determinant $ad - bc \neq 0$, so the transformation may be inverted. If we then take $U(\tau, \xi) = U(at + bx, ct + dx) = u(t, x)$, then (after some tedious chain rule calculations)

$$\begin{aligned} Au_{xx} + Bu_{xt} + Cu_{tt} \quad = \quad & (Ab^2 + Bab + Ca^2)U_{\tau\tau} + [2Abd + B(ad + bc) + 2Cac]U_{\tau\xi} \\ & + (Ad^2 + Bcd + Cc^2)U_{\xi\xi} \end{aligned} \tag{48}$$

Hyperbolic case $\Delta > 0$: We propose to put (46) into the form

$$U_{\tau\xi} + G(\xi, \tau, U, U_\tau, U_\xi) = 0\,. \tag{49}$$

Recall that it was through an invertible change of variables like (47) that we transformed the wave equation into the form (49) and arrived at d'Alembert's formula. Thus, if what we propose is possible, then any hyperbolic PDE may, under a transformation (47), be written in such a way as to have the same principal part as the wave equation.

Now, if $A = C = 0$, then we get this trivially by taking $b = c = 0$, $a = 1$, $d = 1/B$, so that $\tau = t$ and $\xi = x/B$. Hence, we assume that at least one of A, C is nonzero; without loss of generality, $C \neq 0$. Then by taking $b = d = 1$,

$$a = \frac{-B + \sqrt{\Delta}}{2C}, \qquad \text{and} \qquad c = \frac{-B - \sqrt{\Delta}}{2C}$$

then the right-hand side of (48) becomes a constant multiple of $U_{\tau\xi}$, from which we arrive at (49).

Parabolic case $\Delta = 0$: Here, we propose to put (46) into the form

$$U_{\xi\xi} + G(\xi, \tau, U, U_\tau, U_\xi) = 0\,, \tag{50}$$

which is like the heat equation in principal part. Note that, if $B = 0$, precisely one of A or C is zero while the other is nonzero (to avoid the case where our PDE is not even 2^{nd} order), which means the PDE (46) is already in the form (50). Thus, we assume $B \neq 0$ which necessitates $A, C \neq 0$ as well. Then taking

$$\xi = x, \qquad \tau = x - \frac{B}{2C}\,t$$

achieves the form (50).

Elliptic case $\Delta < 0$: In this case we have both $A, C \neq 0$. Then the transformation

$$\tau = -\frac{B}{2C}\,t + x, \qquad \xi = -\frac{\sqrt{-\Delta}}{2C}\,t$$

turns (46) into

$$-\frac{\Delta}{4C}\left(U_{\tau\tau}+U_{\xi\xi}\right)+G(\tau,\xi,U,U_\tau,U_\xi) \;=\; 0\,,$$

which (after dividing through by the common constant multiplier $(-\Delta/(4C))$) has principal part equal to that of Laplace's equation for U.

Olver says that "the field of partial differential equations splits into . . . four subfields, (hyperbolic, parabolic, and elliptic PDEs, and) the last containing all the equations, including higher order equations, that do not fit into the preceding categorization.[5]

Finite Element Methods

Solving Via Minimization

We start with several definitions.

Definition 27. Let $\mathcal{V}$ be an inner product space. A linear operator $L\colon D\subset\mathcal{V}\to\mathcal{V}$ is said to be **positive definite** if $\langle v, Lv\rangle > 0$ for every $v\neq 0$ in D. If $\langle v, Lv\rangle \geq 0$ for every $v\in D$, then L is said to be **positive semidefinite**.

The first step is to show that some operator equations may be solved by minimizing a related quadratic functional.

Theorem 28. Suppose $\mathcal{V}$ is a real inner product space, and $K\colon\mathcal{V}\to\mathcal{V}$ is a self-adjoint positive definite linear operator. Suppose the operator equation

$$K[u]=f$$

has a solution. Then this solution, call it $u_\star$, is unique. Moreover, if we define an associated quadratic functional

$$Q[u] \;:=\; \frac{1}{2}\langle u, K[u]\rangle - \langle f, u\rangle \tag{51}$$

for all admissable $u\in\mathcal{V}$, then $Q[u_\star] < Q[u]$ for all admissable $u\neq u_\star$.

Proof. To establish uniqueness of solution, suppose $u, v\in\mathcal{V}$ are such that $K[u]=K[v]=f$. Then

$$\langle u-v, K[u-v]\rangle \;=\; \langle u-v, K[u]-K[v]\rangle \;=\; \langle u-v, f-f\rangle \;=\; \langle u-v, 0\rangle \;=\; 0\,.$$

[5] Olver, p. 168.

By positive definiteness, $u = v$.

Now note that, for any admissible u

$$\begin{aligned} Q[u] &= \frac{1}{2}\langle u, K[u]\rangle - \langle u, f\rangle \;=\; \frac{1}{2}\langle u, K[u]\rangle - \langle u, k[u_\star]\rangle \\ &= \frac{1}{2}\langle u, K[u]\rangle - \frac{1}{2}\langle u, K[u_\star]\rangle - \frac{1}{2}\langle u, K[u_\star]\rangle \\ &= \frac{1}{2}\langle u, K[u-u_\star]\rangle - \frac{1}{2}\langle u_\star, K[u]\rangle \\ &= \frac{1}{2}\langle u, K[u-u_\star]\rangle - \frac{1}{2}\Big\{\langle u_\star, K[u-u_\star]\rangle + \langle u_\star, K[u_\star]-K[u]\rangle\Big\} - \frac{1}{2}\langle u_\star, K[u]\rangle \\ &= \frac{1}{2}\langle u-u_\star, K[u-u_\star]\rangle - \frac{1}{2}\langle u_\star, K[u_\star]\rangle \;. \end{aligned}$$

But $\langle u-u_\star, K[u-u_\star]\rangle \geq 0$ and achieves its minimum value zero when and only when $u = u_\star$. Thus, Q also is minimized (uniquely) when $u = u_\star$. □

Example 36:

Consider the ODE/BVP

$$-y'' = f(x), \;\; 0 < x < \ell, \qquad \text{subject to BCs} \qquad y(0) = 0 = y(\ell). \tag{52}$$

We are working here with the operator $K[y] = -y''$, the one-dimensional negative Laplacian, subject to homogeneous Dirchlet BCs, so it is self-adjoint and positive definite. To see the latter of these assertions, note that for each ϕ that is twice-differentiable in $(0,\ell)$ with $\phi'' \in L^2(0,\ell)$ (the natural inner product space for us to work in), we have

$$\Big\langle \phi, K[\phi]\Big\rangle \;=\; -\int_0^\ell \phi(x)\phi''(x)\,dx \;=\; -\phi(x)\phi'(x)\Big|_0^\ell + \int_0^\ell [\phi'(x)]^2\,dx \;=\; \Big\langle\!\Big\langle \phi, \phi\Big\rangle\!\Big\rangle \;\geq\; 0,$$

where $\langle\!\langle\cdot,\cdot\rangle\!\rangle$ denotes the *1-dimensional Dirichlet inner product*

$$\Big\langle\!\Big\langle \phi, \psi\Big\rangle\!\Big\rangle \;:=\; \int_0^\ell \phi'(x)\psi'(x)\,dx\;. \tag{53}$$

Note that $\langle\!\langle\cdot,\cdot\rangle\!\rangle$ is not truly an inner product in many contexts, as one can have $\Big\langle\!\Big\langle \phi, \phi\Big\rangle\!\Big\rangle = 0$ with $\phi \neq 0$. However, with our BCs, $\Big\langle\!\Big\langle \phi, \phi\Big\rangle\!\Big\rangle = 0$ implies $\phi \equiv 0$.

Thus, the solution (if one exists) of our problem is the function $y_\star$ minimizing

$$Q[y] \;:=\; \frac{1}{2}\langle y, Ky\rangle - \langle f, y\rangle \;=\; \frac{1}{2}\langle\!\langle y, y\rangle\!\rangle - \langle f, y\rangle\;.$$

■

It is interesting to note that, while the operator K requires its arguments to be (at least piecewise) twice differentiable, Q (in its formulation involving the Dirichlet inner product) only requires arguments (admissible functions) from

$$\mathcal{A} \;=\; \left\{ v\colon [0,\ell] \to \mathbb{R} \,\middle|\, v \text{ is continuous, } v' \text{ is PWC and bdd., and } v(0) = 0 = v(\ell) \right\} .$$

As was the case when we solved IBVPs using Fourier series, if we solve (52) by a process which minimizes the associated functional Q, the result may be a *weak solution*.

Example 37: Poisson Problem in the Plane with Dirichlet BCs

Consider the problem

$$-\Delta u = f, \;\; (x,y) \in \Omega, \qquad \text{with} \qquad u = 0 \;\text{ for }\; (x,y) \in \partial\Omega.$$

Here we assume that Ω is a bounded, connected region in $\mathbb{R}^n$ (say, $n = 2$ or 3) with piecewise smooth boundary $\partial\Omega$. Working under the inner product of $L^2(\Omega)$, our operator $K = -\Delta$ (with the prescribed BCs) is once again self-adjoint and positive definite. The argument for the latter is similar to the above

$$\begin{aligned} \Big\langle \phi, K[\phi] \Big\rangle \;&=\; -\int_\Omega \phi(\mathbf{x})\, \Delta\phi(\mathbf{x})\, d\mathbf{x} \;=\; -\int_{\partial\Omega} \phi(\mathbf{x})\, (\nabla\phi \cdot \mathbf{n})(\mathbf{x})\, d\sigma + \int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\phi(\mathbf{x})\, d\mathbf{x} \\ &=\; \int_\Omega \|\nabla\phi(\mathbf{x})\|^2\, d\mathbf{x} \;=\; \Big\langle\!\!\Big\langle \phi, \phi \Big\rangle\!\!\Big\rangle \;\geq\; 0 \,, \end{aligned}$$

where the *n-dimensional Dirichlet inner product* (truly an inner product because of the BCs) is given by

$$\Big\langle\!\!\Big\langle \phi, \psi \Big\rangle\!\!\Big\rangle \;:=\; \int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\psi(\mathbf{x})\, d\mathbf{x} \tag{54}$$

Hence, the solution to our problem, when it exists, is the unique minimizer $u_\star$ of the functional

$$Q[u] \;:=\; \frac{1}{2}\,\langle u, Ku \rangle - \langle f, u \rangle \;=\; \frac{1}{2}\,\langle\!\langle u, u \rangle\!\rangle - \langle f, u \rangle \;.$$

■

Note that, in the case $\Omega \subset \mathbb{R}^2$,

$$Q[u] \;=\; \iint_\Omega \left(\frac{1}{2}\, u_x^2 + \frac{1}{2}\, u_y^2 - fu \right) dx\, dy$$

and is defined (at least) for all functions u which are piecewise C^1 in Ω and satisfy the homogeneous Dirichlet BCs.

FEM: General Rayleigh-Ritz Approach

We have established that a linear operator equation $K[u] = f$ subject to homogeneous Dirichlet BCs, with K self-adjoint and positive definite, may be recast in the **variational form**

$$\min_{u \in D} Q[u], \tag{55}$$

where Q is a **quadratic functional**, and this minimization occurs over some collection D of admissible functions in a larger inner product space $\mathcal{V}$ (probably an L^2 space). Even though the admissible functions D (generally) do not constitute the full space, D is (usually) an infinite-dimensional subspace. The idea behind the Rayleigh-Ritz[6] approach to FEM is to severely restrict the scope of our search for a minimizer. Instead of searching throughout D, we limit our search to admissible functions lying in some *finite-dimensional* subspace $\mathcal{W}$. In particular, we may fix a choice of independent functions $\phi_1, \phi_2, \ldots, \phi_n \in D$ and take $\mathcal{W} = \text{span}(\{\phi_1, \ldots, \phi_n\})$. We then look to solve

$$\min_{u \in \mathcal{W}} Q[u]. \tag{56}$$

Since each $u \in \mathcal{W}$ has the form

$$u(x) \;=\; \sum_{j=1}^{n} c_j \phi_j(x)\;,$$

(56) is really about choosing the best coefficients $\mathbf{c} = (c_1, \ldots, c_n)$. In an abuse of notation, we now write our quadratic functional as if the input is $\mathbf{c}$:

$$Q(\mathbf{c}) \;:=\; \frac{1}{2}\langle u, K[u]\rangle - \langle f, u\rangle\;, \qquad \text{with} \qquad u = u(x;\mathbf{c}) = \sum_{j=1}^{n} c_j \phi_j(x)\;.$$

Plugging the latter expression for u into the functional yields

$$\begin{aligned} Q(\mathbf{c}) \;&=\; \frac{1}{2}\left\langle \sum_i c_i\phi_i, K\Big[\sum_j c_j\phi_j\Big]\right\rangle - \left\langle f, \sum_i c_i\phi_i\right\rangle \\ &=\; \frac{1}{2}\sum_i\sum_j c_i c_j \left\langle \phi_i, K[\phi_j]\right\rangle - \sum_i c_i \left\langle f, \phi_i\right\rangle \\ &=\; \frac{1}{2}\,\mathbf{c}^T\mathbf{M}\mathbf{c} - \mathbf{c}^T\mathbf{b}\;, \end{aligned} \tag{57}$$

where the **stiffness matrix** $\mathbf{M} = (m_{ij})$ and **load vector** $\mathbf{b} = (b_1, \ldots, b_n)$ are given by

$$m_{ij} \;=\; \left\langle \phi_i, K[\phi_j]\right\rangle, \qquad b_i \;=\; \left\langle f, \phi_i\right\rangle\;, \tag{58}$$

for $i = 1, \ldots, n$, and $j = 1, \ldots, n$. The problem of minimizing a quadratic functional (57) may be a new one to us, but it is a elementary problem in optimization. The **stiffness matrix M**, like the

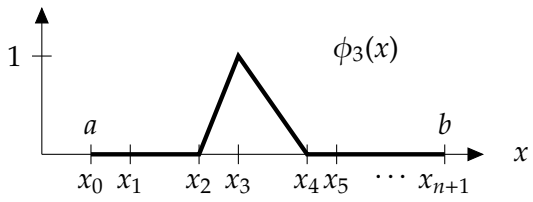
[6] For short blurbs about Rayleigh and Ritz, see Stanoyevitch, p. 426.

underlying operator K, is symmetric (self-adjoint) and positive definite (so nonsingular), and the minimizer is known to be the unique solution of

$$\mathbf{Mc} = \mathbf{b}, \qquad \text{that is,} \qquad \mathbf{c} = \mathbf{M}^{-1}\mathbf{b}.$$

Example 38: Hat Function Basis for One-Dimensional ODEs/BVPs

It seems one of the most common bases to use in the case of a 1-dimensional ODEs/BVP on $[a, b]$ with homogeneous Dirichlet BCs is one consisting of **hat functions**. Let us take the (possibly non-uniform) partition



$$a = x_0 < x_1 < x_2 < \cdots < x_{n+1} = b,$$

with $h_k = x_{k+1} - x_k$ for $k = 0, \ldots, n$, and for $j = 1, 2, \ldots, n$ let $\phi_j(x)$ be the continuous function which is linear on each subinterval $[x_k, x_{k+1}]$ and whose values at mesh points are given by $\phi_j(x_m) = \delta_{jm}$ (Kronecker delta). A plot of $\phi_4(x)$ appears above at right.

Now recall that the problem (52)

$$-y'' = f(x), \;\; 0 < x < \ell, \qquad \text{subject to BCs} \qquad y(0) = 0 = y(\ell),$$

may be solved by minimizing the functional

$$Q[v] \;:=\; \frac{1}{2}\langle\!\langle v, v\rangle\!\rangle - \langle f, v\rangle \;=\; \frac{1}{2}\int_0^\ell \left[\left(\frac{dv}{dx}\right)^2 - f(x)v(x)\right] dx\,,$$

over the set of admissible functions

$$\mathcal{A} \;=\; \left\{v\colon [0,\ell] \to \mathbb{R} \,\middle|\, v \text{ is continuous, } v' \text{ is PWC and bdd., and } v(0) = 0 = v(\ell)\right\}\,.$$

The **piecewise linear Rayleigh-Ritz method** assumes a partition of the interval $[0, \ell]$ and seeks to minimize our functional over the finite-dimensional collection of piecewise linear functions $\mathcal{W} = \text{span}(\{\phi_1, \phi_2, \ldots, \phi_n\})$. That is, we take as our approximate solution

$$\tilde{y}(x) \;=\; \sum_{j=1}^{n} c_j \phi_j(x)\,,$$

where the c_j's are the entries of the vector $\mathbf{c}$ which satisfies $\mathbf{Mc} = \mathbf{b}$, with $\mathbf{M} = (m_{ij})$ and $\mathbf{b}$ having entries given by

$$m_{ij} \;=\; \left\langle \phi_i, \phi_j'' \right\rangle \;=\; \left\langle\!\!\left\langle \phi_i, \phi_j \right\rangle\!\!\right\rangle \;=\; \int_0^\ell \phi_i'(x)\phi_j'(x)\,dx \;=\; \cdots \;=\; \begin{cases} \dfrac{1}{h_{i-1}} + \dfrac{1}{h_i}, & i = j, \\ -\dfrac{1}{h_{\min\{i,j\}}}, & |j-i| = 1, \\ 0, & |j-i| > 1, \end{cases} \tag{59}$$

$$b_i \;=\; \left\langle f, \phi_i \right\rangle \;=\; \cdots \;=\; \frac{1}{h_{i-1}}\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})\,dx + \frac{1}{h_i}\int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x)\,dx\,. \tag{60}$$

Some remarks:

- The matrix **M** is tridiagonal (sparse), which is highly desirable, as the number of **elements** (in this case, subintervals of the original domain $[0, \ell]$) may typically be quite large. This sparsity is owing to the fact that the **support** of the basis functions is fairly small and overlaps with the support of just two other basis functions. In settings where more elaborate basis functions are used, this property is still desirable.
- It is generally wise to place more nodes in regions where the (known) coefficient functions of the differential equation undergo more activity.
- When $\max_i h_i$ is small, we might use the *trapezoid rule* to evaluate the integrals in (60) for the b_i's, we get

$$b_i \approx \frac{1}{2h_{i-1}}[0 + f(x_i)h_{i-1}]h_{i-1} + \frac{1}{2h_i}[f(x_i)h_i + 0]h_i = \frac{1}{2} f(x_i)(h_{i-1} + h_i) .$$

After Stanoyevitch, p. 435ff, apply these ideas to (52) with $\ell = 1$ and

$$f(x) = 100 \sin\left(\operatorname{sign}(x - 0.5) \exp\left(\frac{1}{4|x - 0.5|^{1.05} + 0.3}\right)\right) \exp\left(\frac{1}{4|x - 0.5|^{1.2} + 0.2} - 100(x - 0.5)^2\right) .$$

The code for doing so is found in the file `stanoP436.m`. Some new Octave/Matlab commands of note:

`diff()`, `sign()`, `end` as an index to a vector

The first few lines of the code contain a choice between two types of meshes, one that is uniform, and one that is *adaptive*, being rather coarse where f is well behaved but much finer in the region $0.35 \le x \le 0.65$ where f is highly oscillatory. There is also a switch controlling whether the elements of the load vector **b** are computed using accurate numerical integration or with the trapezoid rule approximation mentioned above. Results are quite different between the two methods using the uniform mesh, but about identical for the adaptive mesh.

■

FEM in Two Spatial Dimensions

We now turn to problems in higher (spatial) dimensions. Just going to two dimensions represents a significant step; that is as far as we will go. Let us continue to assume we are dealing with static (elliptic) problems, and that these are accompanied by homogeneous BCs on bounded, open and connected domains Ω.

Triangulation

We focus first on a two-dimensional analog to the hat functions of the previous example. The way we got those hat functions was to begin with mesh points spaced (perhaps unevenly) throughout a one-dimensional interval. In two dimensions we will want not only mesh points, but a **triangulation** of those mesh points.

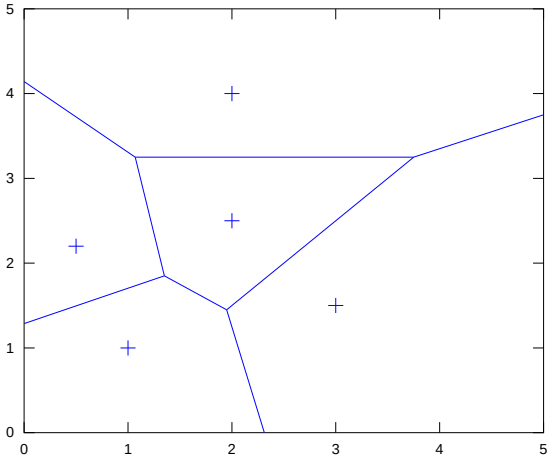
Definition 29. Let $P_1, \ldots, P_n$ be points in the plane $\mathbb{R}^2$. The **Voronoi region** $V(P_i)$ corresponding to point P_i is the set $\left\{Q \in \mathbb{R}^2 \,:\, |Q - P_i| < |Q - P_k| \text{ for } k \neq i\right\}$. A plot of the borders between Voronoi regions is called a **Voronoi diagram**.

Example 39:

Consider the points $P_1 = (1, 1)$, $P_2 = (3, 1.5)$, $P_3 = (2, 4)$, $P_4 = (0.5, 2.2)$, and $P_5 = (2, 2.5)$ in the plane. The following Octave commands display the **Voronoi diagram** for these points.

```
pts = [1 1; 3 1.5; 2 4; 0.5 2.2; 2 2.5];
xVals = pts(:, 1);
yVals = pts(:, 2);
voronoi(xVals, yVals)
axis([0 5 0 5])
```

All but one of the Voronoi regions, of course, are unbounded.



■

Imagine a set $\mathcal{P}$ of points like the one in the previous example. There are many ways the points of $\mathcal{P}$ may be joined by edges to form triangles. One algorithm, known as **Delaunay triangulation**, joins those points (and only those) whose Voronoi regions share a common edge. (Try drawing such triangles in the diagram above.) Here are some properties of the Delaunay triangulation:

- Consider a triangle resulting from the Delaunay triangulation. This triangle joins three points, say p_1, p_2 and p_3, of $\mathcal{P}$. There exists a point $q \in \mathbb{R}^2$ equidistant from p_1, p_2 and p_3, but whose distance to every other point in $\mathcal{P}$ is larger.
- The minimum angle in each of the triangles arising from Delaunay triangulation is as large is possible from any triangulation of the same set $\mathcal{P}$ of points.

We demonstrate Octave commands for carrying out Delaunay triangulation next in the context of

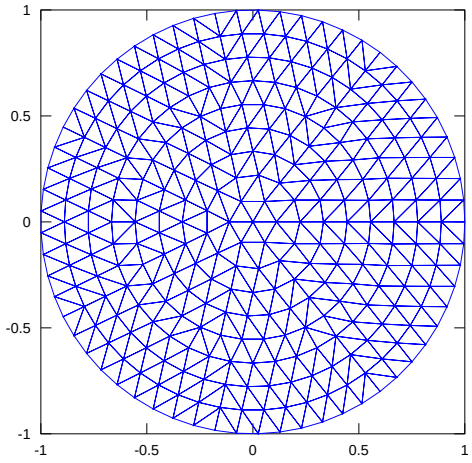
several difference collections of points in the unit circle.

Example 40: Points Distributed Uniformly in the Unit Circle

Suppose we wish to place N points inside the unit circle so that the distances between them are roughly the same. The main code computes positions for such points (with $N = 300$), placing them in vectors `x` and `y`. Once that is completed, we invoke (in the final three lines) the `delaunay()` command to generate a triangulation and `trimesh()` to graph the result.

```
% program taken from Stanoyevitch, p. 614
clear x y
N = 300;
delta = sqrt(pi/N);
x(1) = 0; y(1) = 0;
nodecount = 1;
ncirc = floor(1 / delta);
minrad  = 1 / ncirc;
for j = 1:ncirc
  rad = j * minrad;
  nnodes = floor(2*pi*rad / delta);
  anglegap = 2*pi / nnodes;
  for k = 1:nnodes
    x(nodecount + 1) = rad * cos(k*anglegap);
    y(nodecount + 1) = rad * sin(k*anglegap);
    nodecount += 1;
  end
end

tri = delaunay(x, y);
trimesh(tri, x, y)
axis('equal')
```



■

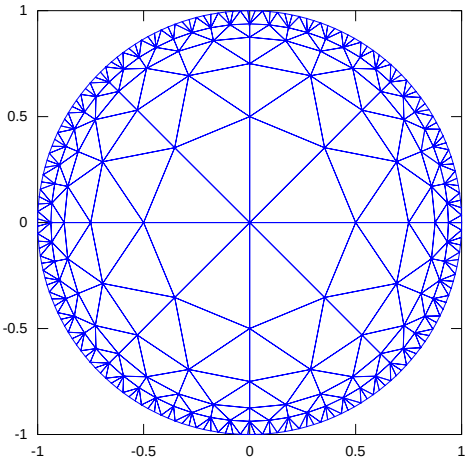
Example 41: Points Distributed More Densely as They Approach the Boundary

We may have reason to require just a few mesh points far from the boundary, but more densely packed ones near it. The resulting Delaunay triangulation may be reminiscent of a painting by M.C. Escher.

```
xb(1) = 0; yb(1) = 0;
oldnodes = 1;     % current number of nodes
rnodes = 299      % remaining number of nodes
newnodes = 299    % # nodes to add next circle
radcount = 1;     % counter for circles
while (newnodes < rnodes / 2)
  rad = 1 - 2^(-radcount);
  for j = 1:newnodes
    xb(oldnodes + j) = rad*cos(2*pi*j/newnodes);
    yb(oldnodes + j) = rad*sin(2*pi*j/newnodes);
  end
  oldnodes += newnodes;
  rnodes -= newnodes;
  radcount += 1;
  newnodes *= 2;
end

% deploy remaining nodes on boundary
for j = 1:rnodes
  xb(oldnodes + j) = cos(2*pi*j/rnodes);
  yb(oldnodes + j) = sin(2*pi*j/rnodes);
end

tri = delaunay(xb, yb);
trimesh(tri, xb, yb)
axis('equal')
```



■

Pyramid Functions

The two-dimensional analog to a one-dimensional hat function is a **pyramid function**. Suppose $\mathcal{P}$ is a set made up of points both on the boundary of and interior to a domain Ω. Let P_j be one of the interior points (that is, not on $\partial\Omega$). The corresponding pyramid function ϕ_j takes the value 1 at P_j, is 0 at every other point in $\mathcal{P}$, and is continuous and piecewise linear (that is, it is made up of patches of planes). More specifically, above every triangle from the triangulation of Ω lies a triangular patch of plane.