Finite Difference Discretization of Elliptic Equations: 1D Problem

Lectures 2 and 3

1 Model Problem

1.1 Poisson Equation in 1D

SLIDE 1

Boundary Value Problem (BVP)

$$\boxed{-u_{xx}(x) = f(x)}$$ $\boxed{\text{N1}}$

$$x \in (0,1), \quad u(0) = u(1) = 0, \quad f \in \mathcal{C}^0$$ $\boxed{\text{N2}}$ $\boxed{\text{N3}}$

Describes many simple physical phenomena (e.g.):

- Deformation of an elastic bar $\boxed{\text{N4}}$
- Deformation of a string under tension $\boxed{\text{N5}}$
- Temperature distribution in a bar $\boxed{\text{N6}}$

The Poisson equation in one dimension is in fact an ordinary differential equation. When dealing with ordinary differential equations we will also use the "prime" notation to indicate differentiation. Thus, $u_x \equiv u'$, $u_{xx} = u''$, etc. The Poisson equation will be used here to illustrate numerical techniques for elliptic PDE's in multi-dimensions. Other techniques specialized for ordinary differential equations could be used if we were only interested in the one dimensional case.

Note 1 ***Poisson equation***

The Poisson equation (in $\mathbb{R}^2$) is elliptic, per our classification. It is also coercive, or positive definite, and symmetric (these concepts will be defined more precisely in the Finite Element lectures). These attributes are very important as regards numerical treatment. These properties are reflected in the fact (see first lecture) that the eigenvalues of $-\nabla^2 v$ are real and positive.

Note 2 ***$\mathcal{C}^m$ spaces***

We denote by $\mathcal{C}^m$, more precisely, $\mathcal{C}^m([0,1])$, the set of functions $f(x) : [0,1] \rightarrow \mathbb{R}$ with continuous m derivatives. Thus, $\mathcal{C}^0$ denotes the set of continuous functions. Obviously, $\mathcal{C}^k \subset \mathcal{C}^m$ for $k > m$.

For this problem, the solution u can be written explicitly as

$$u(x) = \int_0^1 G(x,y)f(y)dy$$

where $G(x,y)$ is the Green's function given by

$$G(x,y) = \begin{cases} y(1-x) & if \quad 0 \le y \le x \\ x(1-y) & if \quad x \le y \le 1 \end{cases}$$

To show this, we start by recalling that for any function which is twice differentiable, there are constants C_1 and C_2, such that

$$\begin{aligned} u(x) &= C_1 + \int_0^x u'(y)dy \\ u'(y) &= C_2 + \int_0^y u''(z)dz. \end{aligned}$$

If u satisfies the one dimensional Poisson equation, then

$$u'(y) = C_2 - \int_0^y f(z)dz.$$

Therefore,

$$u(x) = C_1 + C_2 x - \int_0^x \left(\int_0^y f(z)dz\right) dy.$$

Defining

$$F(y) = \int_0^y f(z)dz,$$

we observe that

$$\begin{aligned} \int_0^x \left(\int_0^y f(z)dz\right) dy &= \int_0^x F(y)dy \\ &= [yF(y)]_0^x - \int_0^x yF'(y)dy \\ &= xF(x) - \int_0^x yf(y)dy \\ &= \int_0^x (x-y)f(y)dy, \end{aligned}$$

by proper attention to dummy variables. Finally, we obtain the general solution in the form

$$u(x) = C_1 + C_2 x - \int_0^x (x-y) f(y) dy.$$

For our particular problem we can now impose the boundary conditions $u(0) = u(1) = 0$ to determine the constants C_1 and C_2. Thus, after some arithmetic,

$$u(x) = \int_0^x y(1-x) f(y) dy + \int_x^y x(1-y) f(y) dy,$$

or

$$u(x) = \int_0^1 G(x,y) f(y) dy.$$

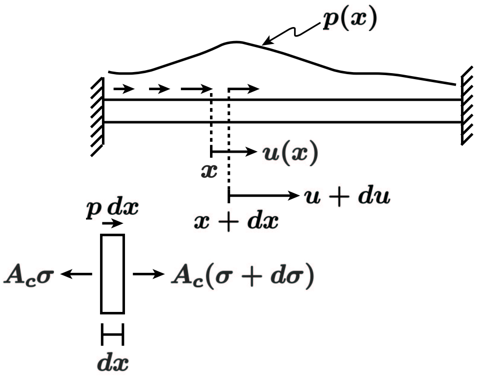
We note that $G(x,y)$ has the following properties:

- G is continuous,
- G is symmetric e.g. $G(x,y) = G(y,x)$,
- $G(x,y) \geq 0$ for all $x, y \in (0,1)$,
- G is a piecewise linear function of x for fixed y and vice versa.

The particular form of expressing the solution, in terms of the Green function, will be revisited when we address the topic of integral equations.

Note 4 ***Elastic bar***

Consider an elastic bar of unit length which is fixed at both ends and subjected to a tangential load per unit length $p(x)$.



Let $\sigma(x)$ and $u(x)$ be the axial stress and tangential displacement at x, respectively. From horizontal equilibrium we have

$$A_c\sigma - A_c(\sigma + d\sigma) = p\ dx$$

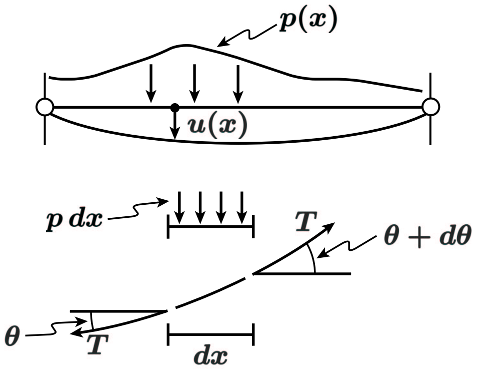
Under the assumption of small displacements and a linearly elastic material we have

$$\sigma = E\frac{(u + du) - u}{dx}, \qquad \text{(Hooke's law)}$$

where E is the modulus of elasticity and A_c is the area of the bar cross section. Differentiating the constitutive equation and combining the two equations to eliminate σ' we obtain the Poisson equation with $f = p/(EA_c)$.

Note 5 ***String under transversal load***

Consider a string of unit length under tension T, which is subjected to a transverse distributed load of magnitude $p(x)$ per unit length.



Let $u(x)$ denote the transverse displacement at point x. Assuming small displacements, so that the tension T can be taken as constant over the whole string, and considering vertical equilibrium we have

$$T(\theta + d\ \theta) - T\theta = p\ dx.$$

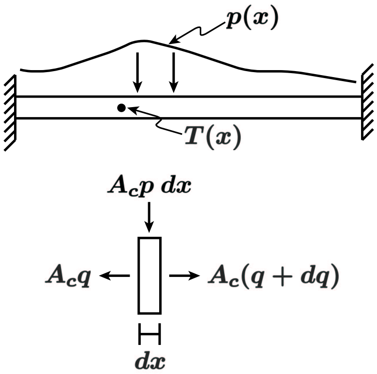
The angle θ can be related to the displacement u simply as

$$\theta = -\frac{du}{dx}.$$

Note the minus sign which is due to the fact that a positive u corresponds to a downwards displacement. Combining the two equations to eliminate the variable θ we obtain the Poisson equation with $f = p/T$.

Note 6 ***Temperature distribution***

Let $u(x)$ and $q(x)$ denote the temperature and heat flux in a homogeneous heat conducting bar of unit length. The bar is subjected to a distributed volumetric heat source $p(x)$ and the temperature is maintained at zero at the end points; the sides of the bar are assumed insulated so that the heat flow is one-dimensional.



The stationary temperature distribution can be obtained by considering the energy balance

$$A_c(q + dq) - A_c q = A_c p\ dx,$$

and the empirical relation between the temperature and the heat flux

$$q = -kT'. \qquad \text{(Fourier's law)}$$

In the above equations, k is the heat conductivity and A_c is the bar cross sectional area. Defining $f = p/k$ and eliminating q from the above equations we obtain the Poission equation.

1.1.1 Solution Properties

SLIDE 2

- The solution $u(x)$ always **exists**
- $u(x)$ is always **"smoother"** than the data $f(x)$

(see first lecture). In particular, if f has m continuous derivatives, u will have $m + 2$ continuous derivatives. Thus, if $f \in \mathcal{C}^0$, then $u \in \mathcal{C}^2$.

- If $f(x) \geq 0$ for all x, then $u(x) \geq 0$ for all x

Follows from the positivity of Green's function.

- $||u||_\infty \leq (1/8)||f||_\infty$ $\boxed{\text{N7}}$
- Given $f(x)$ the solution $u(x)$ is **unique** $\boxed{\text{N8}}$

Note 7 ***Continuous stability estimate***

We recall that for a function $u : \Omega \to \mathbb{R}$

$$||u||_\infty = \sup_{x \in \Omega} |u(x)|,$$

where Ω is the domain of definition. For example, the $||\cdot||_\infty$−norm of the functions x, $x(1-x)$, $e^{\sqrt{x}}$ and $\sin(\pi x)$, in the interval $\Omega \equiv [0,1]$ is 1, 1/4, e and 1, respectively.
Since G is non-negative we have

$$|u(x)| \leq \int_0^1 G(x,y)|f(y)|dy \leq ||f||_\infty \int_0^1 G(x,y)dy = ||f||_\infty \frac{1}{2}x(1-x).$$

Therefore

$$||u||_\infty = \sup_{x \in [0,1]} |u(x)| \leq \frac{1}{8}||f||_\infty.$$

This estimate is a consequence of the fact that the solution u depends continuously on the data f. In other words, we can say that if f is small so is u.

Note 8 ***Solution uniqueness***

Uniqueness of the solution follows directly from the above estimate. If we have two solutions u_1 and u_2 which satisfies the Poisson problem for a given f, we have that $u_1'' - u_2'' = (u_1 - u_2)'' = 0$. This implies that $u_1 - u_2$ satisfy the Poisson problem for $f = 0$. Thus, we can use the above stability estimate to show that $||u_1 - u_2||_\infty = 0$. Therefore, $u_1 = u_2$ (We note that the same conclusion can be reached by integrating $(u_1 - u_2)'' = 0$ twice and imposing the appropriate boundary conditions.)

2 Numerical Solution

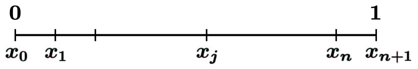
2.1 Finite Differences

2.1.1 Discretization

SLIDE 3

Subdivide interval $(0,1)$ into $n+1$ equal subintervals

$$\Delta x = \frac{1}{n+1}$$



$$x_j = j\Delta x, \qquad \hat{u}_j \approx u_j \equiv u(x_j)$$

$$\text{for} \quad 0 \leq j \leq n+1$$

When v_j corresponds to the values of a continuous function $v(x)$, at point x_j, we will not make any distiction between v_j and $v(x_j)$. We will use $\hat{u}_j$ to denote the approximation to u_j. We will use the underscore to indicate vector. Thus, $\underline{v}$ denotes the vector $\{v_j\}_{1\leq j\leq n}$.

2.1.2 Approximation

SLIDE 4

For example ...

$$\begin{aligned} v''(x_j) &\approx \frac{1}{\Delta x}(v'(x_{j+1/2}) - v'(x_{j-1/2})) \\ &\approx \frac{1}{\Delta x}(\frac{v_{j+1}-v_j}{\Delta x} - \frac{v_j - v_{j-1}}{\Delta x}) \\ &= \frac{v_{j+1}-2v_j+v_{j-1}}{\Delta x^2} \\ & \quad \text{for } \Delta x \text{ small} \end{aligned}$$

A more formal derivation of difference approximations to function derivatives will be consider later.

2.1.3 Equations

SLIDE 5

$-u_{xx} = f$ suggests ...

$$-\frac{\hat{u}_{j+1}-2\hat{u}_j+\hat{u}_{j-1}}{\Delta x^2} = f(x_j) \quad 1 \leq j \leq n$$

$$\hat{u}_0 = \hat{u}_{n+1} = 0$$

$$\Longrightarrow \qquad \boxed{A\,\underline{\hat{u}} = \underline{f}}$$

SLIDE 6

$$A = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}, \quad \underline{\hat{u}} = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_{n-1} \\ \hat{u}_n \end{pmatrix}, \quad \underline{f} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) \end{pmatrix}$$

(Symmetric)

$$A \in \mathbb{R}^{n\times n} \qquad \underline{\hat{u}},\ \underline{f} \in \mathbb{R}^n$$

2.1.4 Solution

SLIDE 7

Is A non-singular ?

For any $\underline{v} = \{v_1, v_2, \ldots, v_n\}^T$

$$\underline{v}^T\ A\ \underline{v} = \frac{1}{\Delta x^2}(v_1^2 + \sum_{i=2}^{n}(v_i - v_{i-1})^2 + v_n^2)$$

Hence $\boxed{\underline{v}^T\ A\ \underline{v} > 0, \text{ for any } \underline{v} \not\equiv 0}$ (A is **SPD**) $\boxed{\text{N9}}$

$A\ \underline{\hat{u}} = \underline{f}$: $\underline{\hat{u}}$ **exists** and is **unique** $\boxed{\text{N10}}$

Note 9 ***Matrix Properties***

The matrix A has a number of properties that will be exploited in the analysis. We give below the definition of some matrix classes and their main properties.

Symmetric Positive Definite (SPD)
We say that a matrix M is positive definite if $\underline{v}^T M \underline{v} > 0$ for any non-zero vector $\underline{v}$. For symmetric matrices this condition is equivalent to requiring that all the eigenvalues of the matrix be positive. To show this we note that if A is symmetric and has real coefficients, it can be written as $M = Q^T \Lambda Q$, where Λ is the diagonal matrix of eigenvalues and Q is an orthonormal transformation, i.e. $Q^{-1} = Q^T$. Then, $\underline{v}^T M \underline{v} = \underline{v}^T Q^T \Lambda Q \underline{v} = \underline{w}^T \Lambda \underline{w} > 0$ for any $\underline{v} \not\equiv 0$ (or any $\underline{w} = Q\underline{v} \not\equiv 0$ since Q is non-singular), implies that all the entries in Λ must be greater than zero. Obviously, any matrix which is SPD is also non-singular and therefore invertible, $M^{-1} = Q\Lambda^{-1}Q^T$.

Diagonal Dominant
We say that a matrix $M = \{m_{ij}\}_{1\leq i,j,\leq n}$ is diagonally dominant if

$$|m_{ii}| \geq \sum_{j \neq i}^{n} |m_{ij}|, \quad \text{for all } i.$$

If the inequality is satisfied with strict inequality, we say that the matrix is strictly diagonally dominant. It can be shown that strictly diagonally dominant matrices are always invertible.
We observe that our matrix A is not strictly diagonally dominant since for all the rows, except the first and last, the equality holds. Matrices that are diagonally dominant and such that: 1) for at least one row the inequality is satisfied in a strict sense, and 2) there is no partition $I_1 \cup I_2$ of $\{1, 2, \ldots, n\}$ such that $m_{i_1 i_2} = 0$ for all $i_1 \in I_1$ and $i_2 \in I_2$, are called irreducible. We can readily verify that A is a diagonally dominant matrix in irreducible form.

M–matrix
A matrix M is called an M–matrix if it satisfies

$$m_{ii} > 0, \quad m_{ij} \leq 0, \quad \text{for all } i \neq j, \quad \sum_{j=1}^{n} m_{ij} > 0 \quad \text{for all} \quad i.$$

It can be shown that if M is a symmetric matrix all the entries in M^{-1} are non-negative numbers.
A is not an M–matrix, since the above strict inequality is not satisfied for every row. However, it can also be shown that if the last inequality above is replaced by an equality, and M is an irreducible diagonally dominant matrix, then all the coefficients of M^{-1} are non-negative. In Note 11 below, we will prove the non-negativity of the coefficients of A^{-1}.

Note 10 ***Thomas' Algorithm***

Gaussian elimination can be efficiently applied to a non-singular tridiagonal system of the form

$$\begin{pmatrix} \alpha_1 & \gamma_1 & 0 & \cdots & 0 \\ \beta_2 & \alpha_2 & \gamma_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots & 0 \\ \vdots & \ddots & \beta_{n-1} & \alpha_{n-1} & \gamma_{n-1} \\ 0 & \cdots & 0 & \beta_n & \alpha_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix}$$

using the following algorithm:

```
$\delta_1 = \alpha_1$
$c_1 = b_1$
for $k = 2, 3, \ldots, n$     (upper triangular form)
    $m_k = \beta_k / \delta_{k-1}$
    $\delta_k = \alpha_k - m_k \gamma_{k-1}$
    $c_k = b_k - m_k c_{k-1}$
$v_n = c_n / \delta_n$          (backsubstitution)
for $k = n-1, n-2, \ldots, 1$
    $v_k = (c_k - \gamma_k v_{k+1}) / \delta_k$
```

The above algorithm would break down if any of the δ_ks becomes zero. It can be shown [TV] that this is not the case for irreducible diagonally dominant matrices such as A. Later in this course we will be devoting time to the solution of linear systems of equations.

2.1.5 Example

SLIDE 8

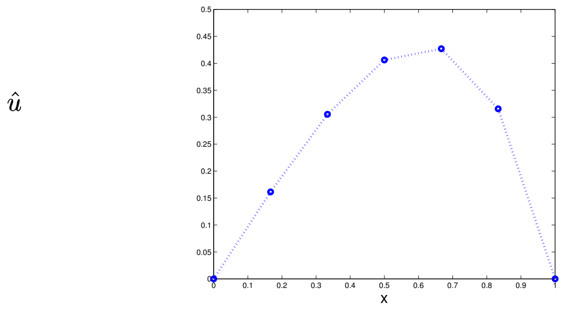
$$-u_{xx} = (3x + x^2)e^x, \qquad x \in (0, 1)$$

with

$$u(0) = u(1) = 0.$$

Take $\quad n = 5, \;\; \Delta x = 1/6 \;\; \ldots$

SLIDE 9



2.1.6 Convergence ?

SLIDE 10

1. Does the discrete solution $\hat{u}$ retain the qualitative propeties of the continuous solution $u(x)$?

2. Does the solution become more accurate when $\Delta x \to 0$?

3. Can we make $|u(x_j) - \hat{u}_j|$ for $0 \le j \le n+1$ arbitrarily small?

3 Discretization Error Analysis

3.1 Properties of A^{-1}

SLIDE 11

Let

$$A^{-1} = \{\alpha_{ij}\}_{1 \leq i,j \leq n}$$

- Non-negativity $\boxed{\text{N11}}$

$$\alpha_{ij} \geq 0, \qquad \text{for} \qquad 1 \leq i, j \leq n$$

- Boundedness $\boxed{\text{N12}}$

$$0 \leq \sum_{j=1}^{N} \alpha_{ij} \leq \frac{1}{8}, \qquad \text{for} \qquad 1 \leq i \leq n$$

Note 11 ***Positivity of the coefficients of*** A^{-1}

We introduce first the following notation: for $\underline{v} \in \mathbb{R}^n$ we say that $\underline{v} \geq 0$ if $v_i \geq 0$ for $1 \leq i \leq n$.

We shall show that if $A\underline{w} = \underline{v}$ and $\underline{v} \geq 0$ this implies that $\underline{w} \geq 0$. This will prove that all the coefficients of A^{-1} are positive since we can choose $\underline{v}$ identically zero except for $v_k = 1$, thus showing that v, which is equal to the k-th column of A^{-1}, is positive.

Let i_0 be the index of the smallest component of $\underline{w}$

$$w_{i_0} = \min_{1 \leq i \leq n} w_i .$$

It is easy to see that i_0 should be either 1 or n, otherwise

$$2w_{i_0} - w_{i_0+1} - w_{i_0-1} \leq 0$$

which obviously contradicts our original assumption. Finally, if $i_0 = 1$, then $2w_1 - w_2 \geq 0$, which implies that $w_1 \geq (w_2 - w_1) \geq 0$ and therefore $\underline{w} \geq 0$. An analogous argument can be used for the case when $i_0 = n$.

Note 12 ***Bound on the coefficients of*** A^{-1}

We note that the function $v(x) = \frac{x(1-x)}{2}$ satisfies

$$-\frac{v(x_{j+1}) - 2v(x_j) + v(x_{j-1})}{\Delta x^2} = 1.$$

This can be verified directly, or deduced from the expression derived in note 13 with $v''(x) = 1$ and $v^{(4)}(x) = 0$.

The above means that the vectors $\underline{v}$ and $\underline{w} = (1, \ldots, 1)^T$ satisfy $\underline{v} = A^{-1}\underline{w}$ from which

$$\sum_{j=1}^{n} \alpha_{ij} = v(x_i) \leq \max_{0 \leq x \leq 1} v(x) = \frac{1}{8} \ .$$

3.2 Qualitative Properties of $\hat{u}$

3.2.1 $f \geq 0 \quad \rightarrow \quad \hat{u} \geq 0$

SLIDE 12

$$\underline{\hat{u}} = A^{-1} \ \underline{f}$$

If

$$f_j = f(x_j) \geq 0 \ , \qquad \text{for} \quad 1 \leq j \leq n$$

Then

$$\hat{u}_i = \sum_j \alpha_{ij} f_j \geq 0 \ , \qquad \text{for} \quad 1 \leq i \leq n$$

3.2.2 Discrete Stability

SLIDE 13

$$\underline{\hat{u}} = A^{-1} \ \underline{f}$$

$$\begin{aligned} ||\underline{\hat{u}}||_\infty &= \max_i |\hat{u}_i| = \max_i (|\sum_j \alpha_{ij} f_j|) \\ &\leq \max_i (\sum_j \alpha_{ij}) \max_i |f_i| \\ &\leq \frac{1}{8} ||\underline{f}||_\infty \end{aligned}$$

For a vector $\underline{v} \in \mathbb{R}^n$ *the* $||\cdot||_\infty$ *norm is defined as* $||\underline{v}||_\infty = \max_{1 \leq i \leq n} |v_i|$.

3.3 Truncation Error

SLIDE 14

For any $v \in \mathcal{C}^4$ we can show that $\boxed{\text{N13}}$

$$\frac{v(x_{j+1}) - 2v(x_j) + v(x_{j-1})}{\Delta x^2} = v''(x_j) + \frac{\Delta x^2}{12} v^{(4)}(x_j + \theta \Delta x)$$

$$-1 \leq \theta \leq 1$$

Take $u \equiv v \qquad (-u'' = f)$

$$-\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{\Delta x^2} = f(x_j) \underbrace{- \frac{\Delta x^2}{12} u^{(4)}(x_j + \theta_j \Delta x)}_{\tau_j}$$

Here, τ_j is referred to as the truncation error, and will be defined more precisely later.

Note 13 ***Difference approximation***

For functions in $\mathcal{C}^4$ we can use Taylor series expansions and write

$$\begin{aligned} v_{j+1} &= v_j + \Delta x v'(x_j) + \frac{\Delta x^2}{2} v''(x_j) + \frac{\Delta x^3}{6} v'''(x_j) + \frac{\Delta x^4}{24} v^{(4)}(x_j + \theta_j^+ \Delta x) \\ v_{j-1} &= v_j - \Delta x v'(x_j) + \frac{\Delta x^2}{2} v''(x_j) - \frac{\Delta x^3}{6} v'''(x_j) + \frac{\Delta x^4}{24} v^{(4)}(x_j + \theta_j^- \Delta x). \end{aligned}$$

Here we have used the Mean Value Theorem to truncate the expansion. The values of θ_j^+ and θ_j^- are unknown, however they should satisfy $0 \le \theta_j^+ \le 1$ and $-1 \le \theta_j^- \le 0$. Combining these two expressions we obtain

$$\frac{v_{j+1} - 2v_j + v_{j-1}}{\Delta x^2} = v''(x_j) + \frac{\Delta x^2}{24} (v^{(4)}(x_j + \theta_j^+ \Delta x) + v^{(4)}(x_j + \theta_j^- \Delta x)).$$

Noting that $v^{(4)}(x)$ is a continuous function, we obtain the desired result.

3.4 Error Equation

SLIDE 15

Let $\boxed{e_j = u(x_j) - \hat{u}_j}$ be the **discretization error**.

$$\begin{aligned} -\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{\Delta x^2} &= f(x_j) + \tau_j \\ -\frac{\hat{u}_{j+1} - 2\hat{u}_j + \hat{u}_{j-1}}{\Delta x^2} &= f(x_j) \end{aligned}$$

Subtracting

$$-\frac{e_{j+1} - 2e_j + e_{j-1}}{\Delta x^2} = \tau_j, \qquad 1 \le j \le n$$

and $\quad e_0 = e_{n+1} = 0$

SLIDE 16

$$\boxed{A\, \underline{e} = \underline{\tau}}$$

$$\underline{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_N \end{pmatrix}, \qquad \underline{\tau} = \frac{\Delta x^2}{12} \begin{pmatrix} u^{(4)}(x_1 + \theta_1 \Delta x) \\ u^{(4)}(x_2 + \theta_2 \Delta x) \\ \vdots \\ \vdots \\ u^{(4)}(x_N + \theta_N \Delta x) \end{pmatrix}$$

3.5 Convergence

SLIDE 17

Using the discrete stability estimate on $A\,\underline{e} = \underline{\tau}$

$$||\underline{e}||_\infty \leq \frac{1}{8}||\underline{\tau}||_\infty$$

or

$$\boxed{\max_{1\leq i\leq n}|u(x_i) - \hat{u}_i| \leq \frac{\Delta x^2}{96}\max_{0\leq x\leq 1}|u^{(4)}(x)|}$$

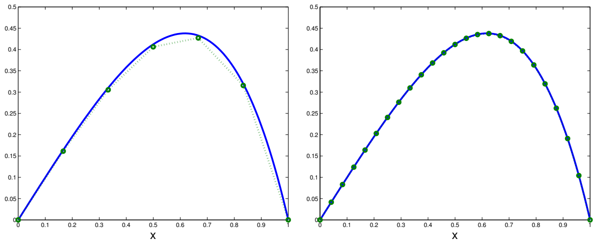
A-priori Error Estimate

We note that the maximum of $u^{(4)}(x)$ over the interval $0 \leq x \leq 1$, will certainly be larger or equal to $u^{(4)}(x_j + \theta_j\Delta x)$ for all j.

3.6 Numerical Example

SLIDE 18

$$-u_{xx} = (3x + x^2)e^x, \quad x \in (0,1), \quad u(0) = u(1) = 0$$



$\Delta x = 1/6$ $\qquad$ $\Delta x = 1/24$

SLIDE 19

EXAMPLE : $\quad -u_{xx} = (3x + x^2)e^x, \;\; x \in (0,1)$

$n+1$	$\|\|\underline{u} - \underline{\hat{u}}\|\|_\infty$
3	0.0227
6	0.0059
12	0.0015
24	$3.756e-04$
48	$9.404e-05$
96	$2.350e-05$
192	$5.876e-06$

Asymptotically,

$$||\underline{u} - \underline{\hat{u}}||_\infty \approx C\Delta x^\alpha$$

$C = 0.216623$
$\alpha = 2.000$

3.7 Summary

SLIDE 20

- For a simple model problem we can produce numerical approximations of **arbitrary accuracy**.

- An **a-priori error estimate** gives the asymptotic dependence of the solution error on the discretization size Δx.

4 Generalizations

4.1 Definitions

SLIDE 21

Consider a linear elliptic **differential equation**

$$\boxed{\mathcal{L}\, u = f}$$

Here we will assume that appropriate boundary conditions are given on u so that the solution is uniquely defined. We will also assume that these boundary conditions are exactly satisfied by the discrete approximation. In our previous example $\mathcal{L}$ is identified with $-\frac{d^2}{dx^2}$.

and a **difference scheme**

$$\boxed{\hat{\mathcal{L}}\, \underline{\hat{u}} = \underline{\hat{f}}}$$

$\hat{\mathcal{L}}$ can be thought of as a matrix operating on the vector $\underline{\hat{u}}$ or a difference operator acting on the grid function $\underline{\hat{u}}$. In our previous example $\hat{\mathcal{L}}$ is identified with the matrix A. We also note that for some schemes $\underline{\hat{f}}$ may be different from $\underline{f}$, i.e. the vector whose components are the values of the function f evaluated at the grid points.

4.2 Consistency

SLIDE 22

The difference scheme is **consistent** with the differential equation if:

For **all** smooth functions v

$$(\hat{\mathcal{L}}\underline{v} - \underline{\hat{f}})_j - (\mathcal{L}v - f)_j \;\rightarrow 0, \quad \text{for} \;\; j = 1, \ldots, n$$

when $\Delta x \rightarrow 0$.

$(\hat{\mathcal{L}}\underline{v} - \underline{\hat{f}})_j - (\mathcal{L}v - f)_j = \mathcal{O}(\Delta x^p)$ for all j

$\Rightarrow$ p is **order of accuracy**

We say that a function $g(\Delta x)$ is $\mathcal{O}(\Delta x^p)$, when $\Delta x \to 0$, if there exists constants C and Δx_0 such that for all $\Delta x < \Delta x_0$, $|g(\Delta x)| < C\Delta x^p$.

4.3 Truncation Error

Slide 23

$$(\hat{\mathcal{L}}\underline{u} - \underline{\hat{f}})_j - \underbrace{(\mathcal{L}u - f)_j}_{=0} = \tau_j, \quad \text{for} \quad j = 1, \ldots, n$$

or,

$$\hat{\mathcal{L}}\underline{u} - \underline{\hat{f}} = \underline{\tau} \ .$$

The truncation error results from inserting the exact solution into the difference scheme.

$$\boxed{\text{Consistency} \ \Rightarrow \ ||\underline{\tau}||_\infty = \mathcal{O}(\Delta x^p)}$$

The above statement is obviously true since, from consistency, each component is $\mathcal{O}(\Delta x^p)$.

4.4 Error Equation

Slide 24

Original scheme

$$\hat{\mathcal{L}} \ \underline{\hat{u}} = \underline{\hat{f}}$$

Consistency

$$\hat{\mathcal{L}} \ \underline{u} = \underline{\hat{f}} + \underline{\tau}$$

The error $\underline{e} = \underline{u} - \underline{\hat{u}}$ satisfies

$$\boxed{\hat{\mathcal{L}}\underline{e} = \underline{\tau} \ .}$$

4.5 Stability

SLIDE 25

Matrix norm

$$||M||_\infty = \sup_{\underline{v} \in \mathbb{R}^n} \frac{||M\underline{v}||_\infty}{||\underline{v}||_\infty}$$

$\boxed{\text{N14}}$

The difference scheme is **stable** if

$$\boxed{||\hat{\mathcal{L}}^{-1}||_\infty \le C \quad \text{(independent of } \Delta x)}$$

SLIDE 26

$$\begin{aligned} ||M||_\infty &= \sup_{||\underline{v}||_\infty = 1} ||M\underline{v}||_\infty \\ &= \sup_{||\underline{v}||_\infty = 1} (\max_i |\sum_{j=1}^n m_{ij} v_j|) \\ &= \max_i (\sup_{||\underline{v}||_\infty = 1} |\sum_{j=1}^n m_{ij} v_j|) \quad v_j = \text{sign}(m_{ij}) \\ &= \max_i \sum_{j=1}^n |m_{ij}| \quad \textbf{(max row sum)} \end{aligned}$$

We see that the sum of the absolute values of the rows of A^{-1} is in fact $||A^{-1}||_\infty$, and what we have in fact shown is that the scheme in our previous example is stable since $||A^{-1}||_\infty < 1/8$, independently of h.

Note 14 ***Matrix norms***

The infinity norm used for vectors and matrices is not the only possible choice, but it is, in most cases, the most convenient when dealing with finite difference schemes.

We can generalize the infinity norm already defined by introducing the so-called p norms. For a vector $\underline{v} \in \mathbb{R}^n$ the p norm is defined as

$$||\underline{v}||_p = (\sum_{j=1}^n |v_j|^p)^{1/p} \quad \text{for } \; p = 1, 2, \ldots < \infty.$$

and for $p = \infty$,

$$||\underline{v}||_\infty = \max_j |v_j|.$$

Associated, or subordinate, to these vector norms we can define for a matrix $M \in \mathbb{R}^{n \times n}$ the corresponding p norm as

$$||M||_p = \sup_{\underline{v}\in\mathbb{R}^n} \frac{||M\underline{v}||_p}{||\underline{v}||_p}.$$

From the definition, it is clear that for any p norm

$$||M\underline{v}||_p \leq ||M||_p ||\underline{v}||_p,$$

for any matrix M and vector $\underline{v}$. We have shown above that the infinity norm is the maximum row sum of absolute values. Similarly, it can be shown that

$$||M||_1 = \max_j \sum_{i=1}^n |m_{ij}|, \quad \text{(max column sum)}$$

and

$$||M||_2 = \sqrt{\mu_{max}}$$

where μ_{max} is the maximum eigenvalue of $M^T M$.
Additional properties satisfied by these p matrix norms are:

- $||M||_p \geq 0$ for all M, with equality if and only if $M = 0$.
- $||cM||_p = |c| ||M||_p$ for all $c \in \mathbb{R}$ and all M.
- $||M + N||_p \leq ||M||_p + ||N||_p$ for all M and N.
- $||M \cdot N||_p \leq ||M||_p ||N||_p$ for all M and N.

4.6 Convergence

SLIDE 27

Error equation

$$\underline{e} = \hat{\mathcal{L}}^{-1}\ \underline{\tau}$$

Taking norms

$$\begin{aligned} ||\underline{e}||_\infty &= ||\hat{\mathcal{L}}^{-1}\ \underline{\tau}||_\infty \\ &\leq ||\hat{\mathcal{L}}^{-1}||_\infty\ ||\underline{\tau}||_\infty \\ &\leq \underbrace{||\hat{\mathcal{L}}^{-1}||_\infty\ C}_{C_1}\ \Delta x^p = C_1\ \Delta x^p \end{aligned}$$

We note that proving convergence in a different norm would require being able to show that, in that particular norm, $||\underline{\tau}|| \to 0$, and more critically, that $||\hat{\mathcal{L}}^{-1}||$ is bounded from above and independently of Δx. Later on, once we are equiped with the tools required to compute the $p = 2$ norm of a matrix, we will illustrate how to prove convergence in the $p = 2$ norm.

4.7 Summary

SLIDE 28

Consistency + Stability ⇒ Convergence

$$\begin{array}{ccccc} \text{Convergence} & & \text{Stability} & & \text{Consistency} \\ ||\underline{e}||_\infty & \leq & ||\hat{\mathcal{L}}^{-1}||_\infty & \cdot & ||\underline{\tau}||_\infty \end{array}$$

We emphasize that consistency establishes the relationship between the numerical scheme and the differential equation. Stability on the other hand is a property of the numerical scheme alone, and guarantees that small perturbations to the right hand side produce small perturbations to the computed solution. The idea that stability plus consistency implies convergence, with perhaps slightly different definitions, is central to numerical analysis and will re-appear several times throughout this course.

It is also possible to show that for optimal convergence (i.e. $\underline{\tau}$ and $\underline{e}$ converging at the same rate), a consistent numerical scheme must be stable.

5 The Eigenvalue Problem

5.1 Model Problem

5.1.1 Statement

SLIDE 29

Find nontrivial (u, λ) such that

$$\begin{aligned} -u_{xx} = \lambda u, \qquad & x \in (0,1) \\ u(0) = u(1) = 0; \end{aligned}$$

denote solutions (u^k, λ^k), $\;k = 1, 2, \ldots,\;$ with

$$0 \leq \lambda^1 \leq \lambda^2 \leq \ldots$$

$\boxed{\text{N15}}$

Note 15 ***Implications of SPD Operator***

As we have indicated, and will later prove in the finite element unit of the course, $-u_{xx}$ is an SPD operator. It can then be shown that it must have positive real eigenvalues, and that the eigenfunctions (which may be chosen real) satisfy the following orthogonality relation

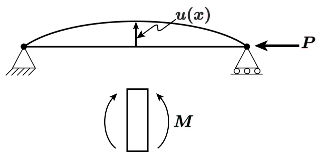
$$\int_0^1 u^k u^l \, dx = C\delta_{kl},$$

where the costant C can be made equal to 1 by proper normalization. For now, we shall simply demonstrate these properties by exact solution of our particular problem.

5.2 Application

5.2.1 Axially Loaded Beam

SLIDE 30



- "Small" Deflection

$$EIu_{xx} = M_{internal}$$

- External Force

$$M_{external} = -Pu$$

Equilibrium $\Rightarrow u_{xx} + \frac{P}{EI}u = 0$ $\boxed{\lambda = P/EI}$

$$\boxed{-u_{xx} = \lambda u, \qquad u(0) = u(1) = 0}$$

In the above expression E is the Young elasticity modulus and I is the moment of inertia of the beam section. The internal moment is proportional the the curvature which, consistent with the small deflection assumption is approximated by u''. Equilibrium states that the internal and external moments must be equal.

5.3 Exact Solution

SLIDE 31

$$-u_{xx} - \lambda u = 0$$
$$\Downarrow$$
$$u = A \sin \sqrt{\lambda} x + B \cos \sqrt{\lambda} x$$

$$\begin{aligned} u(0) &= 0 \Rightarrow B = 0 \\ u(1) &= 0 \Rightarrow A = 0 \ \textbf{ or } \ \lambda = k^2\pi^2, k = 1, 2, \ldots \end{aligned}$$

SLIDE 32

Thus (choose $A = 1$)

$$\left.\begin{array}{r} u^k = \sin k\pi x \\ \lambda^k = k^2\pi^2 \end{array}\right\} \quad k = 1, 2 \ldots$$

Larger $k \Rightarrow$ more oscillatory $u^k \Rightarrow$ larger λ.

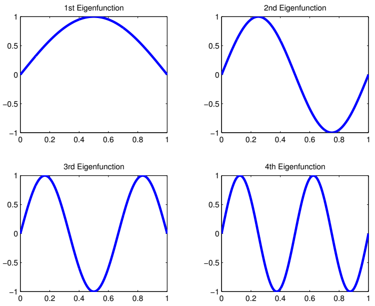
Note the results are quite similar to the periodic case investigated in the first lecture. However now, due to the Dirichlet conditions, the eigenvalues have multiplicity one, and the zero eigenvalue has been eliminated.

For the axially loaded beam, the most relevant eigenvalue is the lowest, since this determines the buckling load. In this case $P/EI = \pi^2$, or $P = EI\pi^2$.

Note 16 **Link to $u_t = u_{xx}$**

We recall from the first lecture that the $-\lambda^k$ correspond to the exponential temporal decay rates of the spatial modes $u^k(x)$ in the separation-of-variables solution of the heat equation. Physically, higher modes have more spatial oscillations, which thus more readily "diffuse out" the heat, and which thus lose their "energy" more quickly. Note that as $k \to \infty$, the decay rate goes to infinity, that is, the timescales go to zero.

SLIDE 33



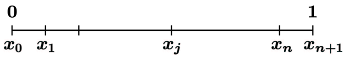
5.4 Discrete Equations

5.4.1 Difference Formulas

SLIDE 34

$$-u_{xx} = \lambda u, \qquad u(0) = u(1) = 0$$

$$\Downarrow$$



$$\boxed{\Delta x = \frac{1}{n+1}}$$

$$\frac{-1}{\Delta x^2}(\hat{u}_{j-1} - 2\hat{u}_j + \hat{u}_{j+1}) = \hat{\lambda}\hat{u}_j, \qquad j = 1, \ldots, n$$

$$\hat{u}_0 = \hat{u}_{n+1} = 0$$

5.4.2 Matrix Form

SLIDE 35

$$A = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}, \qquad \underline{\hat{u}} = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_{n-1} \\ \hat{u}_n \end{pmatrix}$$

$n \times n$ **SPD**

$$A\,\underline{\hat{u}} = \hat{\lambda}\,\underline{\hat{u}} \quad \Rightarrow \quad \underline{\hat{u}}^k, \hat{\lambda}^k, \quad k = 1, 2, \ldots, n$$

$\boxed{\text{N17}}$

$\boxed{\text{N18}}$

Note 17 ***Implications of SPD Matrix***

The fact that SPD operators have real positive eigenvalues and orthogonal eigenfunctions directly implies that SPD matrices have real positive eigenvalues and orthogonal eigenvectors, since we can view a matrix as an operator from $\mathbb{R}^n$ to $\mathbb{R}^n$. This fact is proven in any text on linear algebra. In our current context it tells us that the approximate eigenvalues (the eigenvalues of A) have the same essential features as the exact eigenvalues (the eigenvalues of $-u_{xx}$).

Note 18 ***Number of eigenvalues of*** A

Since A is an $n \times n$ matrix, it must perforce have n eigenvalues (though in general some may have multiplicity greater than unity — note that even in that case an SPD matrix is still diagonalizable). It is immediately clear that we can not possibly hope to approximate all the (i.e., infinite number of) eigenvalues of $-u_{xx}$. We will understand this better shortly; in fact, it is a blessing, as we shall see.

5.5 Error Analysis

5.5.1 Analytical Solution: $\underline{\hat{u}}^k, \hat{\lambda}^k$

SLIDE 36

Claim that

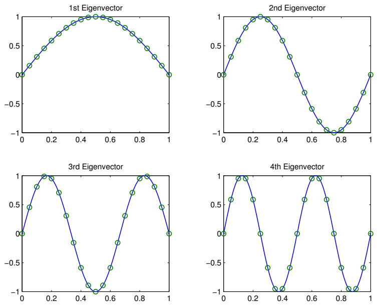
$$\boxed{\underline{\hat{u}}^k \equiv \underline{u}^k}$$

$$\begin{aligned} \hat{u}_j^k &= u^k(x_j) = \sin(k\pi x_j) \\ &= \sin(k\pi j \Delta x) = \sin(\frac{k\pi j}{n+1}), \quad j = 1, \ldots, n \end{aligned}$$

Note $\hat{u}_0^k = \hat{u}_{n+1}^k = 0$ since $\sin(0) = \sin(k\pi) = 0$.

We shall prove the below by direct substitution, and at the same time find the $\hat{\lambda}^k$. *Discrete Fourier analysis will often work with matrices that have particularly simple forms (e.g., Toeplitz). There are of course ways to determine the accuracy of eigenvalues in the more general case; this is discussed briefly in the context of finite element methods.*

SLIDE 37



SLIDE 38

What are $\hat{\lambda}^k$?

$$-\frac{1}{\Delta x^2}\{\hat{u}^k_{j-1} - 2\hat{u}^k_j + \hat{u}^k_{j+1}\}$$

$$= -\frac{1}{\Delta x^2}\{\sin(k\pi(x_j - \Delta x)) - 2\sin(k\pi x_j) + \sin(k\pi(x_j + \Delta x))\}$$

$$= -\frac{1}{\Delta x^2}\{\underbrace{\sin(k\pi x_j - k\pi\Delta x) + \sin(k\pi x_j + k\pi\Delta x)}_{2\cos(k\pi\Delta x)\sin(k\pi x_j)} - 2\sin(k\pi x_j)\}$$

Recall that $\sin(\alpha - \beta) + \sin(\alpha + \beta) = 2\sin\alpha\cos\beta$.

SLIDE 39

Thus:

$$-\frac{1}{\Delta x^2}\{\hat{u}^k_{j-1} - 2\hat{u}^k_j + \hat{u}^k_{j+1}\}$$

$$= -\frac{1}{\Delta x^2}\{2\cos(k\pi\Delta x)\sin(k\pi x_j) - 2\sin(k\pi x_j)\}$$

$$= \underbrace{\frac{2}{\Delta x^2}\{1 - \cos(k\pi\Delta x)\}}_{\hat{\boldsymbol{\lambda}}^k}\ \underbrace{\sin(k\pi x_j)}_{\hat{\boldsymbol{u}}^k_j}.$$

$$\boxed{A\underline{\hat{u}}^k = \hat{\lambda}^k\underline{\hat{u}}^k}$$

5.5.2 Conclusions

SLIDE 40

Low modes

For fixed k, $\Delta x \to 0$:

$$\begin{aligned}
\hat{\lambda}^k &= \frac{2}{\Delta x^2}\{1-\cos(k\pi\Delta x)\} \\
&= \frac{2}{\Delta x^2}\{1-(1-\frac{1}{2}k^2\pi^2\Delta x^2+\mathcal{O}(\Delta x^4))\} \\
&= k^2\pi^2+\mathcal{O}(\Delta x^2)
\end{aligned}$$

second-order convergence, $\hat{\lambda}^k \;\to\; \lambda^k$.

Recall that for $x \to 0$, $\cos(x) \approx 1-\frac{x^2}{2}+\frac{x^4}{24}\ldots$.

SLIDE 41

High modes:

For $k = n$,

$\boxed{\Delta x = \frac{1}{n+1}}$

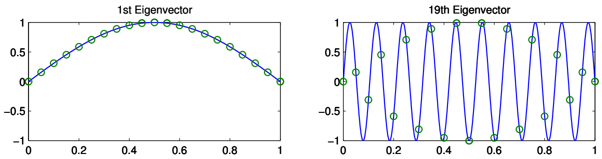
$$\begin{aligned}
\hat{\lambda}^n &= \frac{2}{\Delta x^2}\{1-\cos(\frac{n\pi}{n+1})\} \\
&= 4(n+1)^2 \quad \text{as} \quad \Delta x \to 0
\end{aligned}$$

$$\neq n^2\pi^2 = \lambda^n.$$

High modes ($k \approx n$) **are not** accurate.

SLIDE 42

Low modes vs. high modes
Example : $n = 19$, $\Delta x = 1/20$



SLIDE 43

Low modes vs. high modes

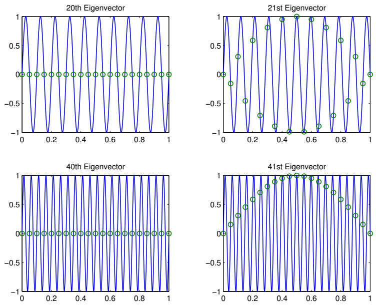
$\boxed{\text{N19}}$

$k \ll n$	$k \approx n$
$\hat{u}^k$ resolved	$\hat{u}^k$ not resolved
$\hat{\lambda}^k$ accurate	$\hat{\lambda}^k$ not accurate
$\hat{\lambda}^k - \lambda^k \sim \mathcal{O}(\Delta x^2)$	$\hat{\lambda}^k - \lambda^k$ is $\mathcal{O}(1)$

$\boxed{\text{BUT: as } \Delta x \to 0,\ n \to \infty, \text{ so any fixed mode } k \text{ converges.}}$

Note 19 — *Aliasing*

We observe that for k close to n, although the eigenvector agrees with the exact solution at the grid points, in between the grid points it misses much of the variation in u^k — and thus we can not expect an accurate eigenvalue. This also "explains" why the discrete approximation can only represent a finite number of modes; the higher modes are simply not seen by the mesh (i.e., they are aliased to the lower modes). The figure below shows some of these higher modes and in particular we observe that the mode 41 has a grid representation which is identical to that of mode 1.



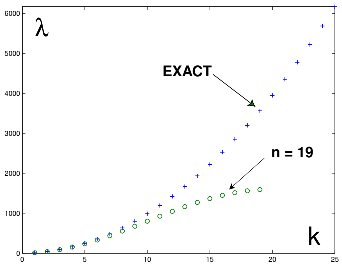
As mentioned in note 17 the matrix A has a full set of orthogonal eigenvectors. It can be verified by direct caculation that the n eigenvectors $\sin(k\pi x_j)$, for $k = 1, \ldots, n$, satisfy the follwoing orthogonality relation,

$$2\Delta x \sum_{j=1}^{n} \sin(k\pi j \Delta x) \sin(l\pi j \Delta x) = \delta_{kl}.$$

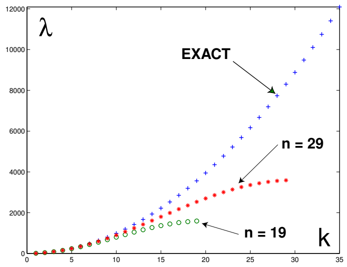
Therefore, any grid function, and in particular the "higher eigenvectors" shown in the above picture, can be uniquely represented as a linear combination of the first n eigenvectors.

Finally, note that although $\hat{\lambda}^n$ is not accurate, it does scale correctly with n, that is, like n^2.

Slide 44



Slide 45



The above pictures illustrate the fact that the eigenvalue for any given mode k, can always be approximated with arbitrary accuracy by making Δx sufficiently small.

5.6 Condition Number of A

SLIDE 46

For a SPD matrix M, the condition number κ_M is given by

$$\boxed{\kappa_M = \frac{\text{maximum eigenvalue of } M}{\text{minimum eigenvalue of } M}.}$$

Thus, for our A matrix,

$$\kappa_A \to \frac{4n^2}{\pi^2} \quad \text{as} \quad \Delta x \to 0$$

grows (in $\mathbb{R}^1$) as number of grid points squared. $\boxed{\text{N20}}$

Importance: understanding solution procedures.

In general, the higher the condition number, the worse. In direct methods, it can sometimes cause numerical stability problems; in iterative methods, it typically implies a slower convergence rate (or the necessity of developing a more sophisticated iterative procedure).

We see here why the finiteness of the number of approximate eigenvalues is a numerical blessing. The "condition number" of $-u_{xx}$ is of course infinite, since the lowest eigenvalue is π^2 and the highest eigenvalue is unbounded. By introducting a numerical approximation, we not only reduce the problem to a finite number of degrees-of-freedom, we also reduce the "stiffness" to a finite value.

Note 20 **_Link to discretization of_ $u_t = u_{xx}$**

If we discretize u_{xx} by finite differences, we are left with a problem still continuous in time, but discrete in space (a "semi-discretization). In so doing, we have eliminated the arbitrarily short timescales, thus making (explicit) treatment in time possible; this is discussed in great detail in the unit on parabolic equations. We mention this here only to again highlight the positive effect of "truncating" the spectrum through discretization.

5.7 Link to $-u_{xx} = f$

5.7.1 Discretization

We show here how the analysis above allows us to obtain an error estimate in a modified $||\cdot||_2$*-norm, for the finite difference treatment of the above equilibrium problem, and more generally, helps us understand consistency and stability. We recall that the* $||\cdot||_2$*-norm of a vector* $v \in \mathbb{R}^n$ *is the Euclidean norm defined as* $||v||_2 = (\sum_{j=1}^{n} v_i^2)^{1/2}$.

SLIDE 47

Recall: $-u_{xx} = f \quad \Rightarrow$

$$-\frac{1}{\Delta x^2}(\hat{u}_{j-1} - 2\hat{u}_j + \hat{u}_{j+1}) = f_j, \quad j = 1, \ldots, n$$

$$\hat{u}_0 = \hat{u}_{n+1} = 0$$

or

$$\boxed{A\underline{\hat{u}} = \underline{f}\ .}$$

Note that the A above is the same A for which we have found the eigenvalues; that is, the eigenvalues of A are $\hat{\lambda}_j, j = 1, \ldots, n$.

SLIDE 48

Error equation: $\underline{e} = \underline{u} - \underline{\hat{u}}$

$$\boxed{A\underline{e} = \underline{\tau},}$$

We have seen that, if the solution u is sufficiently regular, the truncation error of our scheme will satisfy,

$$|\tau_j| \le \max_{x \in (0,1)} \frac{\Delta x^2}{12} u^{(4)}(x) \equiv c_\tau \Delta x^2, \quad \text{for} \quad j = 1, \ldots$$

$\rightarrow 0$ as $\Delta x \rightarrow 0$ (**consistency**).

5.7.2 Norm Definition

SLIDE 49

We will use the "modified" $\| \cdot \|_2$ norm $\boxed{\text{N21}}$

$$\|\underline{v}\|^2 \equiv \boldsymbol{\Delta x} \sum_{i=1}^{n} \underline{v}^T \underline{v} \quad \text{for} \quad \underline{v} \in \mathbb{R}^n$$

$$\boxed{\|\underline{v}\| = \sqrt{\Delta x} \|\underline{v}\|_2}$$

Thus, from consistency

$$\|\underline{\tau}\| \le c_\tau \Delta x^2.$$

Note 21 ***Norm choice***

We choose our norm with the Δx premultiplication to make sure that, as $\Delta x \to 0, v_i = v(x_i)$ for some given function $v(x)$ tends to a constant (in fact, the integral of the square of $v(x)$ over $(0,1)$). This is, in essence, an approximation to the continuous $p = 2$ norm of a function. Recall that for a fucntion $v(x)$ defined for $x \in (0,1)$, the continuous $p = 2$ norm is

$$\|v\|_2 = \{ \int_0^1 v^2(x)\, dx \}^{1/2}.$$

If we were to not include the Δx prefactor, our norm would actually be the sum of an increasing number of pointwise errors, and hence not a very good measure of the accuracy (e.g., would certainly not converge at the same rate as the error in the $p = \infty$ norm).

5.7.3 $\| \cdot \|$ Convergence

SLIDE 50

Ingredients:

1. Rayleigh Quotient: $\boxed{\text{N22}}$

$$\hat{\lambda}^1 \le \frac{\underline{v}^T A \underline{v}}{\underline{v}^T \underline{v}} \le \hat{\lambda}^n, \quad \text{for all} \quad \underline{v} \in \mathbb{R}^n$$

2. Cauchy-Schwarz Inequality: $\boxed{\text{N23}}$

$$\underline{v}^T \underline{w} \le (\underline{v}^T \underline{v})^{\frac{1}{2}} \; (\underline{w}^T \underline{w})^{\frac{1}{2}} \quad \text{for all} \quad \underline{v} \in \mathbb{R}^n$$

Note 22 ***Rayleigh Quotient***

The Rayleigh quotient result given above is proven in most elementary lineary algebra books. The proof is simple: since for an SPD system the eigenvectors form a complete basis, we can write any $\underline{v}$ as $Q\underline{w}$, where Q is the orthonormal matrix of eigenvectors of A (i.e., the eigenvectors are assumed here to be normalized so that $Q^TQ = I$). We then have that

$$\frac{\underline{v}^T A \underline{v}}{\underline{v}^T \underline{v}} = \frac{\underline{w} Q^T A Q \underline{w}}{\underline{w}^T Q^T Q \underline{w}} = \frac{\sum_{i=1}^n w_i^2 \hat{\lambda}^i}{\sum_{i=1}^n w_i^2}$$

since $AQ = Q\Lambda$, where Λ is the diagonal matrix containing the eigenvalues. Then we note that

$$\hat{\lambda}^1 = \frac{\sum_{i=1}^n \hat{\lambda}^1 w_i^2}{\sum_{i=1}^n w_i^2} \leq \frac{\sum_{i=1}^n w_i^2 \hat{\lambda}^i}{\sum_{i=1}^n w_i^2} \leq \frac{\sum_{i=1}^n \hat{\lambda}^n w_i^2}{\sum_{i=1}^n w_i^2} = \hat{\lambda}^n,$$

which completes the proof.

Note 23 ***Cauchy-Schwarz Inequality***

Proof:

$$0 \leq (\underline{v} + \alpha\underline{w})^T(\underline{v} + \alpha\underline{w}) = \underline{v}^T\underline{v} + 2\alpha\underline{v}^T\underline{w} + \alpha^2\underline{w}^T\underline{w} \qquad \forall\alpha \in \mathbb{R}$$

setting $\alpha = -\dfrac{(\underline{v}^T\underline{v})^{1/2}}{(\underline{w}^T\underline{w})^{1/2}}$

$$0 \leq 2(\underline{v}^T\underline{v}) - \frac{2(\underline{v}^T\underline{v})^{1/2}}{(\underline{w}^T\underline{w})^{1/2}}\,(\underline{v}^T\underline{w})$$

and multiplying by $\frac{(\underline{w}^T\underline{w})^{1/2}}{2(\underline{v}^T\underline{v})^{1/2}}$ completes the proof,

$$0 \leq (\underline{v}^T\underline{v})^{1/2}\,(\underline{w}^T\underline{w})^{1/2} - (\underline{v}^T\underline{w}).$$

SLIDE 51

Convergence proof:

$$\boxed{A\underline{e} = \underline{\tau}}$$

$$\underline{e}^T A\underline{e} = \underline{e}^T\underline{\tau}$$

$$\underbrace{\hat{\lambda}^1(\underline{e}^T\underline{e})}_{\times\Delta x} \leq \underbrace{(\underline{e}^T\underline{e})^{\frac{1}{2}}}_{\Delta x^{1/2}}\underbrace{(\underline{\tau}^T\underline{\tau})^{\frac{1}{2}}}_{\Delta x^{1/2}}$$

$$\hat{\lambda}^1||\underline{e}||^2 \leq ||\underline{e}||\;||\tau||$$

SLIDE 52

$$\boxed{\Rightarrow ||\underline{e}|| \leq \frac{1}{\hat{\lambda}^1}||\tau|| \leq \frac{c_\tau}{\hat{\lambda}^1}\Delta x^2}$$

N24 N25 N26

Note 24 ***Presence of*** $\hat{\lambda}^1$

We note that, strictly speaking, there is a Δx effect in our stability estimate, since $\hat{\lambda}^1$ rather than λ^1 appears in our stability constant. However, it is clear that $\hat{\lambda}^1$ approaches a constant as $\Delta x \to 0$, and since stability is an asymptotic concept, the proof of convergence remains effectively unchanged — the scheme is optimal. (In some cases — though not here — the stability constant may even degrade slowly (decrease to zero slowly) as $\Delta x \to 0$; if the degradation is sufficiently slow (compared the order of the scheme), little damage is done. In short, our very strict definition of stability can (must) be relaxed in some situations. Bear in mind that the goal is not stability according to any strict definition per se, but rather rapid convergence (which of course requires some sense of stability).)

Finally, we note that a slight modification of our approximation to the identity matrix on the right-hand side of our discrete eigenvalue problem — as will be obtained when considering finite elements — guarantees a $\hat{\lambda}^1$ which approaches λ^1 from above. In this case (where we must solve a generalized eigenvalue problem), we can bound $1/\hat{\lambda}^1$ by $1/\lambda^1$, and stability is obtained in the standard form. For our finite-difference scheme, our Taylor series estimate (see also slides 44 and 45) indicates that eigenvalues in fact approach from below, which means that $1/\hat{\lambda}^1$ approached $1/\lambda^1$ from above.

Note 25 ***Consistency and Stability***

It seems a bit strange that the high eigenmodes can be in error by order unity, but that the method nevertheless converges. More precisely, from the orthogonality of the eigenvectors, we can show that

$$\hat{u}_j = 2\Delta x \sum_{k=1}^{n} \{ \frac{1}{\hat{\lambda}^k} (\sum_{i=1}^{n} f_i \hat{u}_i^k) \} \hat{u}_j^k .$$

In effect, the last term is the discrete Fourier transform of $\underline{f}$, and our expression is thus analogous to that obtained in the first in the periodic case. It would therefore appear that there will be a large error committed in the higher modes. It is essential to recall that as $n \to \infty$ this higher mode is pushed further and further out; if f is smooth, the Fourier coefficients will decay, and convergence will indeed be obtained.

SLIDE 53

Alternative Derivation

Since $\boxed{\text{N27}}$

$$\|A^{-1}\|_2 = \frac{1}{\hat{\lambda}^1}$$

The proof of the above follows from the Rayleigh quotient result.

From error equation

$$||\underline{e}||_2 \leq ||A^{-1}||_2 ||\underline{\tau}||_2 .$$

Multiplying by $\sqrt{\Delta x}$

$$||\underline{e}|| \leq \frac{1}{\hat{\lambda}^1} ||\underline{\tau}|| .$$

Note 27 ***Relation to*** $||\cdot||_\infty$ ***Estimate***

We have already presented a general stability estimate in terms of the norm of A^{-1}. In fact, the $p = 2$ norm of an SPD matrix is the maximum eigenvalue; since the maximum eigenvalue of a matrix M is the inverse of the minimum eigenvalue of M^{-1}, we directly obtain the stability (and convergence) result which we have derived here "from scratch."