I. INTRODUCTION

1. Motivating examples

1.1 Solving differential equations

Consider the 2-point boundary value problem

$$-y'' + c(x)y = f(x), 0 \le x \le 1, y(0) = \alpha, y(1) = \beta \tag{1.1}$$

Want to solve it numerically on a grid $x_j = jh, \quad h = 1/N, \quad j = 0, \ldots, N$, using finite difference approximations. Let

$$y_j = y(x_j) \ , \quad c_j = c(x_j) \ , \quad f_j = f(x_j)$$

(Known: c_j, f_j, Unknown: y_j.) Finite difference approximation of y'':

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} = y''(x_i) + \frac{h^2}{12} y^{(4)}(x_j) + O(h^4) = y''(x_i) + O(h^2) \tag{1.2}$$

(proof: homework). The notation $F(h) = O(h^p)$ means that there exists some constant $C > 0$ such that for h sufficiently small, $F(h) \le Ch^p$. Substitution (1.2) into (1.1) obtain

$$\frac{-y_{j+1} + 2y_j - y_{j-1}}{h^2} + c_j y_j = f_j + O(h^2) \ , \quad j = 1, \ldots, n-1, y_0 = \alpha, y_n = \beta$$

or

$$\frac{-u_{j+1} + 2u_j - u_{j-1}}{h^2} + c_j u_j = f_j \ , \quad j = 1, \ldots, n-1, u_0 = \alpha, u_n = \beta$$

where $u_j \approx y_j$. Writing this in matrix notation obtain system

$$\begin{bmatrix} 2 + c_1 h^2 & -1 & & & & & \\ -1 & 2 + c_2 h^2 & -1 & & & & \\ & \cdots & \cdots & \cdots & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -1 & 2 + c_{N-2} h^2 & -1 & \\ & & & & -1 & 2 + c_{N-1} h^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} f_1 h^2 + \alpha \\ f_2 h^2 \\ \vdots \\ \vdots \\ f_{n-2} h^2 \\ f_{n-1} h^2 + \beta \end{bmatrix}$$

that is

$$A^h \mathbf{u}^h = \mathbf{f}^h$$

Note: A^h tridiagonal, symmetric Questions we'll answer in this class:

1. Is A^h invertible?
2. How can u^h be computed?
3. Does $u^h \to y$ as $h \to 0$?, ie does the scheme converge? In the homework, as intro matlab exercise, solve this system using the backslash command.

1.2 Discretization and roundoff errors

Finite difference approximations are a good example to introduce the effect of roundoff error, that we will spend part of this class studying.

Using Taylor series expansions we can show that

$$\frac{f(a+h) - f(h)}{h} = f'(a) + O(h)$$

The $O(h)$ error in approximating $f'(a)$ by the finite difference is called the discretization error. However if we plot

$$\left| \frac{f(a+h) - f(a)}{h} - f'(a) \right|$$

for a sample function at a sample base point, for a range of values of h, we obtain figure 1.

The data looks piecewise linear on a log-log scale.

Questions:

If data looks linear on a log-log scale, what can you deduce about it?

Answer: Note that if $y = Ch^p$ then on log-log scale, plot looks linear. p is the slope on the logarithmic scale. In this case we see data has slope 1 in one part (looks like ch), slope -1 in another part (looks like d/h),

(After fitting data:) Why factor 10^{-15}?

Why $y = e$ for small h? What happens for small h?

The figure shows that $\left| \frac{f(a+h) - f(a)}{h} - f'(a) \right| = ch + \frac{d}{h}$ except for very small h. The $O(h)$ behaviour is the discretization error. The remainder must be due to roundoff error. Note that machine representation $fl(x)$ of a number x is not exact, but

$$fl(x) = x(1+\epsilon) \ , \qquad |\epsilon| \leq \epsilon_{mach}$$

Thus $fl \left| \frac{f(a+h) - f(a)}{h} - f'(a) \right| = \left| \frac{f(a+h)(1+\epsilon_1) - f(a)(1+\epsilon_2)}{h(1+\epsilon_3)} - f'(a)(1+\epsilon_4) \right|$

$$\approx \left| \frac{f(a+h) - f(a)}{h} - f'(a) + \frac{f(a+h)\epsilon_1 - f(a)\epsilon_2}{h} \right| = O(h) + O(\frac{\epsilon_{machine}}{h})$$

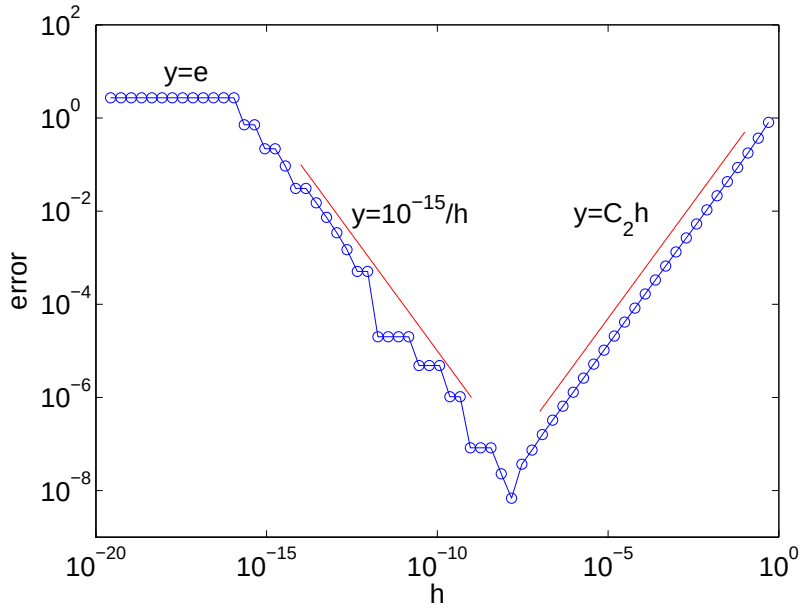


FIG 1: Absolute error in the finite difference approximation of the first derivative of f(x), $f(x) = e^x$ and $a = 1$.

2. Goals in this class

Study numerical methods to solve a problem

1. Matrix factorizations
   - introduce factorizations ($LU, SVD, QR, XDX^{-1}, QTQ^*$)
   - derive algorithms to compute them
   - study their stability, operation count
2. Direct methods using factorizations to solve
   - Ax=b, A invertible
   - Least squares problem (Ax=b, A over or under determined)
   - Eigenvalue problem
3. Iterative methods to solve Ax=b (may complement Trefethen and Bau by another book, by Kelley)

3. Review of Basics

$\mathbf{x}$: n-dimensional column vector ($\mathbf{x}^T$: n-dimensional row vector)

A: $m \times n$ matrix with entries a_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$

Matrix-vector Multiplication $A\mathbf{x} = \mathbf{b}$: m-dimensional column vector

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & a_{3n} \\ : & : & & : \\ : & : & & : \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ : \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ : \\ : \\ b_m \end{bmatrix}$$

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \ldots i = 1, \ldots, m$$

Note: if $\mathbf{a}_j$ is the jth column of A, then $\mathbf{b} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots x_n\mathbf{a}_n$ is a linear combination of the columns of A. Thus $\mathbf{b}$ is in the column space of A. Also, as a result:

Matrix-Matrix Multiplication Every column of $C = AB$ is a linear combination of the columns of$A$.

column space: space spanned by columns $=\{\mathbf{b}|\mathbf{b} = A\mathbf{x}$ for some $\mathbf{x}\}$ =range(A)= $R(A)$

rank of A: $rank(A) = dim(range(A))$

row space: space spanned by rows

null space of A: $N(A) = \{\mathbf{x}|A\mathbf{x} = \mathbf{0}\}$, Remember $\dim(N(A)) + \dim(R(A)) = n$

I expect you to be fully comfortable with the terms: basis, linear combination, vector space, span, dimension,..

Invertibility. Theorem: Let $A = C_{n\times n}$. The following conditions are equivalent:

(i) A is invertible. ie, there exists A^{-1} such that $AA^{-1} = A^{-1}A = I$.
(ii) $rank(A) = n$
(iii) $range(A) = C^n$
(iv) $N(A) = \{\mathbf{0}\}$
(v) $\det(A) \neq 0$

(vi) 0 is not an eigenvalue of A

Definition: If $\mathbf{x} \in C^n$, then $\mathbf{x}^* = \overline{\mathbf{x}^T}$, $A^* = \overline{A^T}$ (adjoint).
If $\mathbf{x} \in Re^n$, then $\mathbf{x}^* = x^T, A^* = A^T$ (transpose).

Note: $(AB)^* = B^*A^*$ and $(A^*)^{-1} = (A^{-1})^* = A^{-*}$

	$\mathbf{x} \in C^n$, Complex	$\mathbf{x} \in Re^n$, Real
Inner product	$\mathbf{x}^*\mathbf{y} = \sum_{j=1}^{N} \overline{x_j} y_j$	$\mathbf{x}^T\mathbf{y} = \sum_{j=1}^{N} x_j y_j$

Inner product defines norms $||\mathbf{x}||^2 = \mathbf{x}^*\mathbf{x}$, and angles between vectors: $\cos\gamma = \frac{\mathbf{x}^*\mathbf{y}}{||\mathbf{x}||||\mathbf{y}||}$. As a result, two vectors are orthogonal (with respect to this inner product) if $\mathbf{x}^*\mathbf{y} = 0$.

Remember from calculus: $\frac{\mathbf{a}\cdot\mathbf{b}}{||\mathbf{b}||^2}\mathbf{b}$ is the component of $\mathbf{a}$ parallel to $\mathbf{b}$ (*vector projection*), and $\mathbf{a} - \frac{\mathbf{a}\cdot\mathbf{b}}{||\mathbf{b}||^2}\mathbf{b}$ is the component of $\mathbf{a}$ orthogonal to $\mathbf{b}$. If $\mathbf{b}$ is a unit vector $\mathbf{q}$ then $\mathbf{a} - \mathbf{a}\cdot\mathbf{q}\mathbf{q}$ is orthogonal to $\mathbf{q}$. Thus $\mathbf{a}$ can be written as a component parallel to $\mathbf{q}$ and one normal to $\mathbf{q}$. Exercise: show that if $\mathbf{q}_1, \ldots, \mathbf{q}_n$ is an orthonormal set, then $\mathbf{r} = \mathbf{a} - (\mathbf{a}^*\mathbf{q}_1)\mathbf{q}_1 - (\mathbf{a}^*\mathbf{q}_1)\mathbf{q}_2 - \ldots - (\mathbf{a}^*\mathbf{q}_n)\mathbf{q}_n$ is orthogonal to $\mathbf{q}_i$ for all $i = 1, \ldots, n$. Thus can write $\mathbf{a}$ as a sum of orthogonal components, $\mathbf{a} = \mathbf{r} + (\mathbf{a}^*\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{a}^*\mathbf{q}_1)\mathbf{q}_2 + \ldots + (\mathbf{a}^*\mathbf{q}_n)\mathbf{q}_n$.

Hermitian	$A^* = A$	$A^T = A$ (Symmetric)
Unitary	$A^* = A^{-1}$	$A^T = A^{-1}$ (Orthogonal)
	$A^*A = AA^* = I$	$A^TA = AA^T = I$

A matrix is unitary if and only if its columns $\mathbf{a}_i$ form an orthonormal basis for C^n with above inner product, since $(A^*A)_{ij} = \mathbf{a}_i^*\mathbf{a}_j = I_{ij} = \delta_{ij}$.

Normal	$A^*A = AA^*$	$A^TA = AA^T$

Notes:

1. $\mathbf{x}^*\mathbf{x} \geq 0$, $\mathbf{x}^*\mathbf{x} = 0 \Leftrightarrow \mathbf{x} = 0$

2. Let $A \in C^{n\times n}$ be hermitian. Then the eigenvalues λ_j, $j = 1, \ldots, n$ are real and the eigenvectors form an orthogonal set. By normalizing obtain an orthonormal basis for C^n, $\mathbf{u}_j, j = 1, \ldots, n$. The matrix
$$[\mathbf{u}_1\mathbf{u}_2 \ldots \mathbf{u}_n]$$
is unitary. From $A\mathbf{u}_j = \lambda\mathbf{u}_j$ we obtain that $AU = UD$, $D = diag(\lambda_1, \ldots, \lambda_n)$, or $A = UDU^{-l}$. Since U is unitary,
$$A = UDU^* \ .$$

3. If $A \in C^{m\times n}$ then A^*A is hermitian with $\lambda_j \geq 0$. Proof: (1) $(AB)^* = B^*A^* \Rightarrow (A^*A)^* = A^*A^{**} = A^*A$. (2) $A^*A\mathbf{x} = \lambda\mathbf{x} \Rightarrow \mathbf{x}^*A^*A\mathbf{x} = \mathbf{x}^*\lambda\mathbf{x} \Rightarrow (A\mathbf{x})^*(A\mathbf{x}) = \lambda\mathbf{x}^*\mathbf{x}$
$\Rightarrow \lambda \geq 0$ since $(A\mathbf{x})^*(A\mathbf{x}) \geq 0$ and $\mathbf{x}^*\mathbf{x} \geq 0$, by Note 1.

4. Norms

Definition: norm

4.1 Vector Norms

Example: 1-norm, 2-norm, inf-norm, corresponding unit balls

Definition: inner product

Definition: vector norm and angles induced by an inner product

4.2 Matrix Norms

Definition: matrix norm induced by a vector norm, $||A|| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||}{||\mathbf{x}||}$.

Example: Compute induced matrix 1-norm,2-norm,inf-norm, by looking at image of unit ball.

Theorem: Induced norms satisfy the additional property that $||AB|| \leq ||A||||B||$. Proof: in class.

Theorem: $||A||_1 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||_1}{||\mathbf{x}||_1} = \max_j \sum_i |a_{ij}|$ = max absolute column sum. Proof: in class

Theorem: $||A||_\infty = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||_\infty}{||\mathbf{x}||_\infty} = \max_i \sum_j |a_{ij}|$ = max absolute row sum. Proof: next homework

Theorem: $||A||_2 = \sqrt{\rho(A^*A)}$. If A Hermitian : $||A||_2 = \rho(A)$. Proof: in class

Example: Compute induced matrix norms from theorems, for A as in previous example.

Definition: Frobenius norm (not an induced norm)

Theorem: The Frobenius norms also satisfies the additional property that $||AB|| \leq ||A||||B||$. Proof: in class.

II. SVD AND QR FACTORIZATIONS, LEAST SQUARES

1. Singular Value Decomposition (SVD)

Theorem: If A is an $m \times n$ matrix, then A has a singular value decomposition, $A = U\Sigma V^*$ where $U_{m\times m}$, $V_{n\times n}$ are unitary and $\Sigma_{m\times n}$ has only nonzero entries σ_i on the diagonal, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

Proof: in class (constructive).
Find SVD for an example.

Notes:

1. The singular values are unique, the matrices U and V are not.
2. Since V diagonalizes A^*A, columns $\mathbf{v}_j$ are eigenvectors of A^*A (right singular vectors of A) Similarly, U diagonalizes AA^* and its columns are eigenvectors of AA^* (left singular vectors)
3. $||A||_2 = \sigma_1, ||A||_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}$
4. If A hermitian, then $\sigma_i(A) = |\lambda_i(A)|$
5. $det(A) = det(\Sigma)$, $rank(A) = rank(\Sigma)$
6. A reduces to the diagonal matrix Σ when the domain is expressed in the basis $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and the range is expressed in the basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$.
7. SVD shows that $A_{m\times n}$ maps the unit sphere in $\Re^n$ into an ellipsoid in $\Re^m$ with principal axes $\sigma_1\mathbf{u}_1, \ldots, \sigma_r\mathbf{u}_r$.
8. Contrast singular value decomposition $U\Sigma V^*$ and Jordan form BJB^{-1}:
   (a) SVD applies to non-square matrices, Jordan doesn't
   (b) SVD uses orthonormal bases to achieve diagonal form, Jordan uses arbitrary bases to achieve bidiagonal form
   (c) Jordan used to copute A^k, e^A, SVD: we will see
9. Reduced, compressed SVD
10. Best approximation. SVD=sum of rank one matrices with special property :. Thus, SVD lets you find closest matrix in 2-norm with smaller rank. Applications to image compression, matrix compression, numerical rank.

2. QR Factorization

2.1 Classical Gram-Schmidt

2.2 Modified Gram-Schmidt

2.3 Householder Algorithm

3. Least Squares Problem

Problem statement: Given $A \in C^{m\times n}$, $\mathbf{b} \in C^m$, want to solve

$$A\mathbf{x} = \mathbf{b} \qquad (1)\ .$$

But if $\mathbf{b}$ not in $Range(A)$ then (1) has no solution. (For overdetermined systems, $m > n$, this will be the case for some $\mathbf{b}$.) In that case, find the **least squares solution** $\widehat{x}$ that minimizes the error or **residual** $\mathbf{r} = A\widehat{x} - \mathbf{b}$ in the 2-norm.

Definition: The *least squares solution* to $A\mathbf{x} = \mathbf{b}$ is the vector $\widehat{x}$ that minimizes the 2-norm of the residual $\mathbf{r} = A\mathbf{x} - b$,

$$||\mathbf{b} - A\widehat{x}||_2 = \min_x ||\mathbf{b} - A\mathbf{x}||_2$$

Theorem: Let $A \in C^{m\times n}$. The least squares solution $\widehat{x}$ is such that

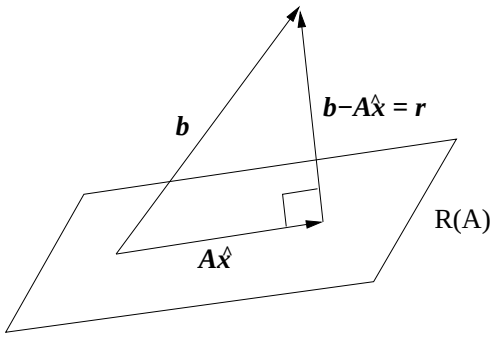
$$\mathbf{r} = A\widehat{x} - b \perp Range(A)\ .$$

That is $\mathbf{r}$ is perpendicular to all columns of A, $\mathbf{r} \perp \mathbf{a}_i$, that is

$$A^*\mathbf{r} = \mathbf{0}\ ,$$

that is

$$A^*A\widehat{x} = A^*\mathbf{b}\ . \qquad (2)$$



Proof : (in class)

Equations (2) are called the **normal equations**. These equations are equivalent to

$$A\mathbf{x} = P_A\mathbf{b} \qquad (3)$$

where P_A is the orthogonal projector onto the range of A.

Note: If $A_{m\times n}$, $m \geq n$, $rank(A) = n$, then
(1) A^*A is invertible (HW).
(2) $\widehat{x} = (A^*A)^{-1}A^*\mathbf{b}$ is unique
(3) $A^+ = (A^*A)^{-1}A^*$ is the **pseudoinverse**

There are 3 general methods for solving the LS problem:

(1) **Normal equations.**

$$A^*A\mathbf{x} = A^*\mathbf{b}$$

The matrix A^*A is positive definite. A positive definite system is solved with Choleskis LU factorization (later). Total number of ops: $mn^2 + n^3/3 + O(n^2)$ The first terms comes from forming A^*A, the second from the Choleski decomposition, the last from the solver.

(2) **QR Factorization.**

Let $\widehat{Q}\widehat{R} = A$ be the reduced QR factorization of A. Then $Q = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_r]$ where $rank(A) = r$. The $\{\mathbf{q}_i\}_{i=1}^r$ are and orthonormal basis for $Range(A)$. From our earlier notes,

$$P_A = \mathbf{q}_1\mathbf{q}_1^* + \mathbf{q}_2\mathbf{q}_2^* \ldots \mathbf{q}_r\mathbf{q}_r^* = \widehat{Q}\widehat{Q}^*$$

The last equality can be confirmed by looking at the components on the left and right hand side:

$$\Big(\sum_{k=1}^{r} \mathbf{q}_k\mathbf{q}_k^*\Big)_{ij} = \sum_{k=1}^{r} \mathbf{q}_{k,i}\mathbf{q}_{k,j} = \sum_{k=1}^{r} Q_{ik}Q_{jk} = \sum_{k=1}^{r} Q_{ik}Q_{kj}^* = (QQ^*)_{ij} \ .$$

Thus equation (3) states that $\widehat{Q}\widehat{R}\mathbf{x} = \widehat{Q}\widehat{Q}^*\mathbf{b}$. Premultiplying by $\widehat{Q}^*$ and using that $\widehat{Q}^*\widehat{Q} = I_{r\times r}$ yields

$$\widehat{R}\widehat{x} = \widehat{Q}^*\mathbf{b} \ . \tag{4}$$

$\widehat{R}$ is an upper triangular $r \times m$ matrix of rank r, so there are $n - r$ free variables. By construction, there exists a solution, unique if and only if $n = r$.

(3) **SVD Factorization.**

Let $\widehat{U}\widehat{\Sigma}\widehat{V}^* = A$ be the reduced SVD factorization of $A_{m\times n}$, where $r = rank(A)$, where $U_{m\times r} = [\mathbf{u}_1\mathbf{u}_2 \ldots \mathbf{u}_r], V_{m\times r} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]$, and $\widehat{\Sigma}_{r\times r} = diag\sigma_1, \ldots, \sigma_r$. These equations state that

$$A\mathbf{v}_j = \begin{cases} \sigma_j\mathbf{u}_j \text{ if } j \le r \\ 0 \ j \ge r+1 \end{cases}$$

Thus the $\{\mathbf{u}_j\}_{j=1}^r$ is an orthonormal basis for the range of A. Thus the orthogonal projector onto the range of A is

$$P_A = \mathbf{u}_1\mathbf{u}_1^* + \mathbf{u}_2\mathbf{u}_2^* \ldots \mathbf{u}_r\mathbf{u}_r^* = \widehat{U}\widehat{U}^*$$

Thus equation (3) states that $\widehat{U}\widehat{\Sigma}\widehat{V}^*\mathbf{x} = \widehat{U}\widehat{U}^*\mathbf{b}$. Premultiplying by $\widehat{U}^*$ and using that $\widehat{U}^*\widehat{U} = I_{r\times r}$ yields

$$\widehat{V}^*\widehat{x} = \widehat{\Sigma}^{-1}\widehat{U}^*\mathbf{b}. \tag{5}$$

$\widehat{V}^*$ has rank r so there are $n-r$ free variables. By construction, there exists a solution, unique if and only if $n = r$.