Solutions Manual to

Numerical Linear Algebra

Original Text by

Lloyd N. Trefethen & David Bau, III

이명규 지음

Myeongkyu Lee

mgklee@kaist.ac.kr

Contents

Part I Fundamentals 1

Lecture 4 The Singular Value Decomposition . . . 2

Lecture 5 More on the SVD . . . 5

Part II QR Factorization and Least Squares 6

Lecture 6 Projectors . . . 7

Lecture 7 QR Factorization . . . 10

Lecture 10 Householder Triangularization . . . 12

Lecture 11 Least Squares Problems . . . 14

Part III Conditioning and Stability 15

Lecture 12 Conditioning and Condition Numbers . . . 16

Lecture 13 Floating Point Arithmetic . . . 17

Lecture 14 Stability . . . 18

Lecture 15 More on Stability . . . 20

Lecture 16 Stability of Householder Triangularization . . . 22

Lecture 17 Stability of Back Substitution . . . 24

Lecture 19 Stability of Least Squares Algorithms . . . 25

Part IV Systems of Equations 26

Lecture 20 Gaussian Elimination . . . 27

Lecture 21 Pivoting . . . 29

Lecture 22 Stability of Gaussian Elimination . . . 31

Lecture 23 Cholesky Factorization . . . 33

Part V Eigenvalues 34

Lecture 24 Eigenvalue Problems . . . 35

Lecture 25 Overview of Eigenvalue Algorithms 37
Lecture 27 Rayleigh Quotient, Inverse Iteration 39
Lecture 28 QR Algorithm without Shifts 40
Lecture 31 Computing the SVD 42

Part VI Iterative Methods **44**

Lecture 33 The Arnoldi Iteration 45
Lecture 35 GMRES 47
Lecture 36 The Lanczos Iteration 48
Lecture 38 Conjugate Gradients 49
Lecture 40 Preconditioning 51

Part I

Fundamentals

Lecture 4

The Singular Value Decomposition

4.1 Determine SVDs of the following matrices (by hand calculation):

(b) $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Solution. Let A be each matrix.

(b) Since $A = A^*$, by Theorem 5.5, the singular values of A are $\sigma_1 = 3$ and $\sigma_2 = 2$. The unit eigenvectors corresponding to $\lambda_1 = 3$ and $\lambda_2 = 2$ are

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ and } \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

respectively. Thus,

$$\mathbf{u}_1 = \frac{1}{\sigma_1} A\mathbf{v}_1 = \frac{1}{3} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and

$$\mathbf{u}_2 = \frac{1}{\sigma_2} A\mathbf{v}_2 = \frac{1}{2} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Therefore, the SVD of A is

$$A = U\Sigma V^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

(c) We first compute

$$A^*A = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}.$$

By Theorem 5.4, the nonzero singular value of A is $\sigma_1 = \sqrt{\lambda_1} = 2$. The unit eigenvector corresponding to $\lambda_1 = 4$ is $\mathbf{v}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}^*$. Thus,

$$\mathbf{u}_1 = \frac{1}{\sigma_1} A\mathbf{v}_1 = \frac{1}{2} \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

By the Gram–Schmidt process, we can extend $\{\mathbf{u}_1\}$ to an orthonormal basis for $\mathbb{C}^3$ and $\{\mathbf{v}_1\}$ to an orthonormal basis for $\mathbb{C}^2$. Therefore, an SVD of A is

$$A = U\Sigma V^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

(e) Since $A = A^*$, by Theorem 5.5, the singular values of A are the absolute values of the eigenvalues of A. It follows from

$$\det(\lambda I - A) = \begin{bmatrix} \lambda - 1 & -1 \\ -1 & \lambda - 1 \end{bmatrix} = (\lambda - 1)^2 - 1 = \lambda(\lambda - 2) = 0$$

that $\lambda_1 = 2$ and $\lambda_2 = 0$, so $\sigma_1 = 2$ and $\sigma_2 = 0$. The unit eigenvector corresponding to $\lambda_1 = 2$ is $\mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^*$. Thus,

$$\mathbf{u}_1 = \frac{1}{\sigma_1} A\mathbf{v}_1 = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

By the Gram–Schmidt process, we can extend $\{\mathbf{u}_1\}$ to an orthonormal basis for $\mathbb{C}^2$ and $\{\mathbf{v}_1\}$ to an orthonormal basis for $\mathbb{C}^2$. Therefore, an SVD of A is

$$A = U\Sigma V^* = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

■

4.2 Suppose A is an $m \times n$ matrix and B is the $n \times m$ matrix obtained by rotating A ninety degrees clockwise on paper (not exactly a standard mathematical transformation!). Do A and B have the same singular values? Prove that the answer is yes or give a counterexample.

Solution. Observe that

$$B = A^T \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Thus, we have

$$BB^* = A^T \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \left(A^T\right)^* = A^T I \overline{A} = A^T \overline{A} = \overline{A^* A},$$

where $\overline{A}$ denotes the matrix with complex conjugated entries. Since BB^* and $\overline{A^*A}$ are hermitian, their eigenvalues are real. Thus, A^*A and BB^* have the same eigenvalues. By Theorem 5.4, A and B have the same singular values. ■

4.4 Two matrices $A, B \in \mathbb{C}^{m\times m}$ are *unitarily equivalent* if $A = QBQ^*$ for some unitary $Q \in \mathbb{C}^{m\times m}$. Is it true or false that A and B are unitarily equivalent if and only if they have the same singular values?

Solution. Two square matrices A and B are unitarily equivalent only if they have the same singular values. The converse is not necessarily true.

($\Longrightarrow$) Suppose that $A = QBQ^*$ for some unitary $Q \in \mathbb{C}^{m\times m}$. Let $B = U\Sigma V^*$ be an SVD of B. Then $A = QU\Sigma V^*Q^* = (QU)\Sigma(QV)^*$ is an SVD of A because QU and QV are unitary. (Recall that the product of unitary matrices is unitary.) By Theorem 4.1, A and B have the same, uniquely determined singular values $\{\sigma_j\}$.

($\not\Longleftarrow$) Let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then

$$AA^* = BB^* = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}.$$

By Theorem 5.4, A and B have the same singular values, $\sigma_1 = \sqrt{2}$ and $\sigma_2 = 0$. Since B is a diagonal matrix, A and B are unitarily equivalent if and only if A is unitarily diagonalizable. By Theorem 24.8, a matrix is unitarily diagonalizable if and only if it is normal. However, A is not normal because

$$A^*A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

but

$$AA^* = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Therefore, A and B are not unitarily equivalent.

■

Remark.

($\Longrightarrow$) Suppose that $A = QBQ^*$ for some unitary $Q \in \mathbb{C}^{m\times m}$.

$$\begin{aligned} \lambda \text{ is an eigenvalue of } AA^* &\iff AA^*\mathbf{x} = QBB^*Q^*\mathbf{x} = \lambda\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{C}^m \setminus \{\mathbf{0}\} \\ &\iff BB^*(Q^*\mathbf{x}) = \lambda(Q^*\mathbf{x}) \text{ for some } Q^*\mathbf{x} \in \mathbb{C}^m \setminus \{\mathbf{0}\} \\ &\iff \lambda \text{ is an eigenvalue of } BB^* \end{aligned}$$

By Theorem 5.4, A and B have the same singular values.

($\not\Longleftarrow$) According to the elementary linear algebra, a matrix is orthogonally diagonalizable if and only if it is symmetric. Because A is not symmetric, it is not orthogonally diagonalizable.

Lecture 5

More on the SVD

5.4 Suppose $A \in \mathbb{C}^{m\times m}$ has an SVD $A = U\Sigma V^*$. Find an eigenvalue decomposition (5.1) of the $2m \times 2m$ hermitian matrix.

$$\begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}.$$

Solution. $A = U\Sigma V^*$ implies $AV = U\Sigma$, and $A^* = V\Sigma U^*$ implies $A^*U = V\Sigma$. Thus, we have

$$\begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}.$$

Therefore, an eigenvalue decomposition of the hermitian matrix is

$$\begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix}^{-1}.$$

■

Part II

QR Factorization and Least Squares

Lecture 6

Projectors

6.1 If P is an orthogonal projector, then $I-2P$ is unitary. Prove this algebraically, and give a geometric interpretation.

Solution. If P is an orthogonal projector, $P^2 = P$, and $P = P^*$ by Theorem 6.1. $I - 2P$ is Hermitian because it is the sum of two Hermitian matrices. It follows from

$$(I-2P)^2 = I - 4P + 4P^2 = I - 4P + 4P = I$$

that $(I-2P)^{-1} = I - 2P = (I-2P)^*$. Therefore, $I - 2P$ is unitary.

Note that $I - 2P = (I - P) - P$ is involutory; it is its own inverse. It represents a reflection across $\text{range}(I-P) = \text{null}(P)$. ■

6.2 Let E be the $m \times m$ matrix that extracts the "even part" of an m-vector: $E\mathbf{x} = (\mathbf{x} + F\mathbf{x})/2$, where F is the $m \times m$ matrix that flips $(x_1, \cdots, x_m)^*$ to $(x_m, \cdots, x_1)^*$. Is E an orthogonal projector, an oblique projector, or not a projector at all? What are its entries?

Solution. Let

$$F = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Then

$$F\mathbf{x} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{m-1} \\ x_m \end{bmatrix} = \begin{bmatrix} x_m \\ x_{m-1} \\ \vdots \\ x_2 \\ x_1 \end{bmatrix}$$

for all $\mathbf{x} \in \mathbb{C}^m$. Note that $E = \frac{1}{2}(I + F)$. E is a projector because

$$E^2 = \frac{(I+F)^2}{4} = \frac{I^2 + 2IF + F^2}{4} = \frac{I + 2F + I}{4} = \frac{I+F}{2} = E.$$

E is Hermitian because it is the sum of two Hermitian matrices. By Theorem 6.1, E is an orthogonal projector.

$$E = \begin{cases} \frac{1}{2}\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 1 & 0 \\ 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} & \text{if } m \text{ is even,} \\ \frac{1}{2}\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} & \text{if } m \text{ is odd.} \end{cases}$$

■

6.4 Consider the matrices

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Answer the following questions by hand calculation.

(a) What is the orthogonal projector P onto range(A), and what is the image under P of the vector $(1, 2, 3)^*$?

Solution. We first compute

$$A^*A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$(A^*A)^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}.$$

By Equation (6.13), the orthogonal projector onto range(A) is

$$P = A(A^*A)^{-1}A^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

Therefore, the image under P of the vector $\mathbf{x} = (1, 2, 3)^*$ is

$$P\mathbf{x} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}.$$

■

6.5 Let $P \in \mathbb{C}^{m\times m}$ be a nonzero projector. Show that $\|P\|_2 \geq 1$, with equality if and only if P is an orthogonal projector.

Solution. Since P is a projector, $P^2 = P$, so $\left\|P^2\right\|_2 = \|P\|_2$. By Inequality (3.14), $\|P\|_2^2 \geq \left\|P^2\right\|_2$. Hence, $\|P\|_2^2 \geq \|P\|_2$. Since P is nonzero, by (1) of Equations (3.15), $\|P\|_2 > 0$. Therefore, $\|P\|_2 \geq 1$. It suffices to show that $\|P\|_2 \leq 1$ if and only if P is an orthogonal projector.

($\Longleftarrow$) If P is an orthogonal projector, $P = P^*$ by Theorem 6.1. Then

$$\|P\mathbf{x}\|_2^2 = |\mathbf{x}^* P^* P\mathbf{x}| = \left|\mathbf{x}^* P^2 \mathbf{x}\right| = |\mathbf{x}^* P\mathbf{x}| \leq \|\mathbf{x}\|_2 \|P\mathbf{x}\|_2$$

by the Cauchy–Schwarz inequality (3.12). Hence, $\|P\|_2 = \sup\{\|P\mathbf{x}\|_2 : \|\mathbf{x}\|_2 = 1\} \leq 1$.

($\Longrightarrow$) If $\|P\|_2 \leq 1$, then $\|P\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{C}^m$. Let $\mathbf{x} \in \text{null}(P)^\perp$, and let $\mathbf{y} = P\mathbf{x} - \mathbf{x}$. Note that $P\mathbf{y} = \mathbf{0}$, so $\mathbf{y} \in \text{null}(P)$. We have $\mathbf{x} \perp \mathbf{y}$, so $\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 = \|\mathbf{x} + \mathbf{y}\|_2^2$. It follows from

$$\|\mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 = \|\mathbf{x} + \mathbf{y}\|_2^2 = \|P\mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_2^2$$

that $\mathbf{y} = 0$, so $\mathbf{x} = P\mathbf{x} \in \text{range}(P)$. Hence, $\text{null}(P)^\perp \subset \text{range}(P)$. For arbitrary $\mathbf{v} \in \text{range}(P)$, let $\mathbf{v}_1 \in \text{null}(P)$ and $\mathbf{v}_2 \in \text{null}(P)^\perp$ such that $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$. Since $\mathbf{v}_2 \in \text{null}(P)^\perp \subset \text{range}(P)$, $\mathbf{v} = P\mathbf{v} = P\mathbf{v}_2 = \mathbf{v}_2$, so $\text{range}(P) \subset \text{null}(P)^\perp$. Therefore, $\text{range}(P) = \text{null}(P)^\perp$, so P is an orthogonal projector.

■

Lecture 7

QR Factorization

7.1 Consider again the matrices A and B of Exercise 6.4.

(a) Using any method you like, determine (on paper) a reduced QR factorization $A = \hat{Q}\hat{R}$ and a full QR factorization $A = QR$.

Solution. Let $\mathbf{a}_1 = (1, 0, 1)^*$ and let $\mathbf{a}_2 = (0, 1, 0)^*$. We use the Gram–Schmidt process.

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix},$$

$$\mathbf{q}_2 = \frac{\mathbf{a}_2 - (\mathbf{q}_1^*\mathbf{a}_1)\mathbf{q}_1}{\|\mathbf{a}_2 - (\mathbf{q}_1^*\mathbf{a}_1)\mathbf{q}_1\|} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Hence, a reduced QR factorization of A is

$$A = \hat{Q}\hat{R} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix}.$$

Let $\mathbf{a}_3 = (0, 0, 1)^*$ and continue the Gram–Schmidt process.

$$\mathbf{q}_3 = \frac{\mathbf{a}_3 - (\mathbf{q}_1^*\mathbf{a}_3)\mathbf{q}_1 - (\mathbf{q}_2^*\mathbf{a}_3)\mathbf{q}_2}{\|\mathbf{a}_3 - (\mathbf{q}_1^*\mathbf{a}_3)\mathbf{q}_1 - (\mathbf{q}_2^*\mathbf{a}_3)\mathbf{q}_2\|} = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore, a full QR factorization of A is

$$A = QR = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

■

7.4 Let $\mathbf{x}^{(1)}$, $\mathbf{y}^{(1)}$, $\mathbf{x}^{(2)}$, and $\mathbf{y}^{(2)}$ be nonzero vectors in $\mathbb{R}^3$ with the property that $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ are linearly independent and so are $\mathbf{x}^{(2)}$ and $\mathbf{y}^{(2)}$. Consider the two planes in $\mathbb{R}^3$,

$$P^{(1)} = \langle \mathbf{x}^{(1)}, \mathbf{y}^{(1)} \rangle, \quad P^{(2)} = \langle \mathbf{x}^{(2)}, \mathbf{y}^{(2)} \rangle.$$

Suppose we wish to find a nonzero vector $\mathbf{v} \in \mathbb{R}^3$ that lies in the intersection $P = P^{(1)} \cap P^{(2)}$. Devise a method for solving this problem by reducing it to the computation of QR factorizations of three 3×2 matrices.

Solution. Note that $\mathbf{0} \in P^{(1)} \cap P^{(2)}$. By a QR factorization of a 3×2 matrix $\begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix}$, we can find a nonzero "normal" vector $\mathbf{n} \in \mathbb{R}^3$ that is orthogonal to the plane spanned by $\mathbf{x}$ and $\mathbf{y}$. Find nonzero vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ in $\mathbb{R}^3$ such that $\mathbf{n}_1 \perp P^{(1)}$ and $\mathbf{n}_2 \perp P^{(2)}$ by QR factorizations of $\begin{bmatrix} \mathbf{x}^{(1)} & \mathbf{y}^{(1)} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{x}^{(2)} & \mathbf{y}^{(2)} \end{bmatrix}$, respectively. Then we can finally find such a nonzero vector $\mathbf{v} \in \mathbb{R}^3$ by a QR factorization of $\begin{bmatrix} \mathbf{n}_1 & \mathbf{n}_2 \end{bmatrix}$. $\mathbf{v}$ is orthogonal to the two normal vectors, so it lies in $P^{(1)} \cap P^{(2)}$. ■

7.5 Let A be an $m \times n$ matrix ($m \geq n$), and let $A = \hat{Q}\hat{R}$ be a reduced QR factorization.

(a) Show that A has rank n if and only if all the diagonal entries of $\hat{R}$ are nonzero.

Solution. It is equivalent to show that $\operatorname{rank}(A) < n$ if and only if $r_{jj} = 0$ for some j. $\operatorname{rank}(A) = \dim(\operatorname{col}(A)) < n$ if and only if there is a column vector

$$\mathbf{a}_j \in \operatorname{span}\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{j-1}\} = \operatorname{span}\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{j-1}\}$$

such that $\mathbf{v}_j = \mathbf{a}_j - \sum_{i=1}^{j-1} (\mathbf{q}_i^* \mathbf{a}_j)\mathbf{q}_i = \mathbf{0}$ and $r_{jj} = \|\mathbf{v}_j\|_2 = 0$ as mentioned in the proof of Theorem 7.1. ■

Lecture 10

Householder Triangularization

10.1 Determine the (a) eigenvalues, (b) determinant, and (c) singular values of a Householder reflector. For the eigenvalues, give a geometric argument as well as an algebraic proof.

Solution. Given a nonzero vector $\mathbf{v}$, by Equation (10.4), a Householder reflector F is

$$F = I - 2\frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}}.$$

(a) F has eigenvalues ± 1. For a nonzero vector $\mathbf{u}$ that is orthogonal to $\mathbf{v}$,

$$F\mathbf{u} = \mathbf{u} - 2\mathbf{v}\left(\frac{\mathbf{v}^*\mathbf{u}}{\mathbf{v}^*\mathbf{v}}\right) = \mathbf{u}.$$

Each vector in the "mirror" hyperplane H is mapped to itself.

$$F\mathbf{v} = \mathbf{v} - 2\mathbf{v}\left(\frac{\mathbf{v}^*\mathbf{v}}{\mathbf{v}^*\mathbf{v}}\right) = -\mathbf{v}.$$

A vector that is orthogonal to the hyperplane is mapped to the opposite side of the "mirror"; its mirror image is its negative.

(b) $\det F = -1$.

(c) F is Hermitian. By Theorem 5.5, the singular values of F are the absolute values of the eigenvalues of F. Hence, the singular values of F are all 1.

■

10.4 Consider the 2×2 orthogonal matrices

$$F = \begin{bmatrix} -c & s \\ s & c \end{bmatrix}, \qquad J = \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

where $s = \sin\theta$ and $c = \cos\theta$ for some θ. The first matrix has $\det F = -1$ and is a reflector—the special case of a Householder reflector in dimension 2. The second has $\det J = 1$ and effects a rotation instead of a reflection. Such a matrix is called a *Givens rotation*.

(a) Describe exactly what geometric effects left-multiplications by F and J have on the plane $\mathbb{R}^2$. (J rotates the plane by the angle θ, for example, but is the rotation clockwise or counterclockwise?)

Solution. Observe that

$$F\mathbf{e}_1 = \begin{bmatrix} -\cos\theta \\ \sin\theta \end{bmatrix}, \qquad F\mathbf{e}_2 = \begin{bmatrix} \sin\theta \\ \cos\theta \end{bmatrix}.$$

F represents the reflection about the line through the origin having an angle of $\dfrac{\pi - \theta}{2}$ with the positive x-axis. From

$$J\mathbf{e}_1 = \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix}, \qquad J\mathbf{e}_2 = \begin{bmatrix} \sin\theta \\ \cos\theta \end{bmatrix},$$

J rotates the plane by the angle θ clockwise. ■

Lecture 11

Least Squares Problems

11.1 Suppose the $m \times n$ matrix A has the form

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix},$$

where A_1 is a nonsingular matrix of dimension $n \times n$ and A_2 is an arbitrary matrix of dimension $(m-n) \times n$. Prove that $\|A^+\|_2 \leq \left\|{A_1}^{-1}\right\|_2$.

Solution. Recall that $\mathbf{x} \in \text{range}(A) \setminus \{\mathbf{0}\}$ if and only if $A\mathbf{y} = \mathbf{x}$ for some $\mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$. Therefore,

$$\begin{aligned}
\left\|A^+\right\|_2 &= \sup_{\mathbf{x} \in \mathbb{C}^m \setminus \{\mathbf{0}\}} \frac{\|A^+\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\
&= \sup_{\mathbf{x} \in \mathbb{C}^m \setminus \{\mathbf{0}\}} \frac{\|A^+\mathbf{x}\|_2}{\|AA^+\mathbf{x} + (I - AA^+)\mathbf{x}\|_2} \\
&\leq \sup_{\mathbf{x} \in \text{range}(A) \setminus \{\mathbf{0}\}} \frac{\|A^+\mathbf{x}\|_2}{\|AA^+\mathbf{x}\|_2} \\
&= \sup_{\mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}} \frac{\|\mathbf{y}\|_2}{\|A\mathbf{y}\|_2} \\
&= \sup_{\mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}} \frac{\|\mathbf{y}\|_2}{\sqrt{\|A_1\mathbf{y}\|_2^2 + \|A_2\mathbf{y}\|_2^2}} \\
&\leq \sup_{\mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}} \frac{\|\mathbf{y}\|_2}{\|A_1\mathbf{y}\|_2} \\
&= \sup_{\mathbf{z} \in \mathbb{C}^n \setminus \{\mathbf{0}\}} \frac{\|{A_1}^{-1}\mathbf{z}\|_2}{\|\mathbf{z}\|_2} \\
&= \left\|{A_1}^{-1}\right\|_2.
\end{aligned}$$

■

Remark. Let $A = \hat{U}\hat{\Sigma}V^*$ be the reduced SVD of an $m \times n$ matrix A with $m \geq n$. Since $A^+ = V\hat{\Sigma}^{-1}\hat{U}^*$, we have

$$\|A^+\|_2 = \left\|\hat{\Sigma}^{-1}\right\|_2 = \frac{1}{\sigma_{\min}}.$$

Part III

Conditioning and Stability

Lecture 12

Conditioning and Condition Numbers

12.1 Suppose A is a 202×202 matrix with $\|A\|_2 = 100$ and $\|A\|_F = 101$. Give the sharpest possible lower bound on the 2-norm condition number $\kappa(A)$.

Solution. Let $\{\sigma_j\}_{j=1}^{202}$ be the non-increasing sequence of singular values of A. By Theorem 5.3,

$$\|A\|_2 = \sigma_1 = 100,$$
$$\|A\|_F = \left(\sum_{j=1}^{202} \sigma_j^2\right)^{1/2} = 101.$$

Hence,

$$201 = 101^2 - 100^2 = \sum_{j=2}^{202} \sigma_j^2 \geq \sum_{j=2}^{202} \sigma_{202}^2 = 201\sigma_{202}^2,$$
$$\therefore 0 \leq \sigma_{202} \leq 1.$$

By Equation (12.16),

$$\kappa(A) = \frac{\sigma_1}{\sigma_{202}} \geq 100.$$

The equality holds if $\sigma_2 = \cdots = \sigma_{202} = 1$. $\kappa(A)$ attains its minimum, so it is indeed the infimum. ■

Lecture 13

Floating Point Arithmetic

13.1 Between an adjacent pair of nonzero IEEE single precision real numbers, how many IEEE double precision numbers are there?

Solution. We count the number of x such that

$$\frac{m}{2^{24}} < \frac{x}{2^{53}} < \frac{m+1}{2^{24}} \iff 2^{29}m < x < 2^{29}m + 2^{29}.$$

Thus, there are $2^{29} - 1$ IEEE double precision numbers strictly between two adjacent nonzero IEEE single precision numbers. ■

Lecture 14

Stability

14.1 True or False?

(a) $\sin x = O(1)$ as $x \to \infty$.

(b) $\sin x = O(1)$ as $x \to 0$.

(f) $\mathrm{fl}(\pi) - \pi = O(\epsilon_{\text{machine}})$. (We do not mention that the limit is $\epsilon_{\text{machine}} \to 0$, since that is implicit for all expressions $O(\epsilon_{\text{machine}})$ in this book.)

(g) $\mathrm{fl}(n\pi) - n\pi = O(\epsilon_{\text{machine}})$, uniformly for all integers n. (Here $n\pi$ represents the exact mathematical quantity, not the result of a floating point calculation.)

Solution. $|\sin x| \le 1 \cdot 1$ for all $x \in \mathbb{R}$, so both (a) and (b) are true.

(f) True. By (13.5), there exists an ϵ with $|\epsilon| \le \epsilon_{\text{machine}}$ such that $\mathrm{fl}(\pi) = \pi(1+\epsilon)$. Hence,

$$|\mathrm{fl}(\pi) - \pi| = \pi|\epsilon| \le \pi\epsilon_{\text{machine}},$$

so $\mathrm{fl}(\pi) - \pi = O(\epsilon_{\text{machine}})$.

(g) False. By (13.5), there exists an ϵ with $|\epsilon| \le \epsilon_{\text{machine}}$ such that $\mathrm{fl}(n\pi) = n\pi(1+\epsilon)$. However, we can make $|\mathrm{fl}(n\pi) - n\pi| = |n| \cdot \pi|\epsilon|$ arbitrarily big by choosing some sufficiently large $|n|$. Therefore, the equality does not hold uniformly.

■

14.2 (a) Show that $(1 + O(\epsilon_{\text{machine}}))(1 + O(\epsilon_{\text{machine}})) = 1 + O(\epsilon_{\text{machine}})$. The precise meaning of this statement is that if f is a function satisfying $f(\epsilon_{\text{machine}}) = (1 + O(\epsilon_{\text{machine}}))(1 + O(\epsilon_{\text{machine}}))$ as $\epsilon_{\text{machine}} \to 0$, then f also satisfies $f(\epsilon_{\text{machine}}) = 1 + O(\epsilon_{\text{machine}})$ as $\epsilon_{\text{machine}} \to 0$.

(b) Show that $(1 + O(\epsilon_{\text{machine}}))^{-1} = 1 + O(\epsilon_{\text{machine}})$.

Solution.

(a) $f(\epsilon_{\text{machine}}) = O(\epsilon_{\text{machine}})$ if there exist some positive constants δ_1 and C_1 such that

$$0 \le \epsilon_{\text{machine}} < \delta_1 \implies \|f(\epsilon_{\text{machine}})\| \le C_1\epsilon_{\text{machine}}.$$

$g(\epsilon_{\text{machine}}) = O(\epsilon_{\text{machine}})$ if there exist some positive constants δ_2 and C_2 such that

$$0 \le \epsilon_{\text{machine}} < \delta_2 \implies \|g(\epsilon_{\text{machine}})\| \le C_2 \epsilon_{\text{machine}}.$$

Let $\delta = \min\{\delta_1, \delta_2, 1\}$. Then $0 \le \epsilon_{\text{machine}} < \delta$ implies

$$\begin{aligned}
&\|(1 + f(\epsilon_{\text{machine}}))(1 + g(\epsilon_{\text{machine}})) - 1\| \\
&= \|f(\epsilon_{\text{machine}}) + g(\epsilon_{\text{machine}}) + f(\epsilon_{\text{machine}})g(\epsilon_{\text{machine}})\| \\
&\le \|f(\epsilon_{\text{machine}})\| + \|g(\epsilon_{\text{machine}})\| + \|f(\epsilon_{\text{machine}})g(\epsilon_{\text{machine}})\| \\
&\le C_1\epsilon_{\text{machine}} + C_2\epsilon_{\text{machine}} + C_1C_2\epsilon_{\text{machine}}^2 \\
&\le (C_1 + C_2 + C_1C_2)\epsilon_{\text{machine}}.
\end{aligned}$$

$C_1 + C_2 + C_1C_2 > 0$, so $(1+f(\epsilon_{\text{machine}}))(1+g(\epsilon_{\text{machine}}))-1 = O(\epsilon_{\text{machine}})$. Therefore, $(1+O(\epsilon_{\text{machine}}))(1+O(\epsilon_{\text{machine}})) = 1 + O(\epsilon_{\text{machine}})$.

(b) $f(\epsilon_{\text{machine}}) = O(\epsilon_{\text{machine}})$ if there exist some positive constants δ_1 and C such that

$$0 \le \epsilon_{\text{machine}} < \delta_1 \implies \|f(\epsilon_{\text{machine}})\| \le C\epsilon_{\text{machine}}.$$

Let $\delta = \min\left\{\delta_1, \dfrac{1}{2C}\right\}$. Then $0 \le \epsilon_{\text{machine}} < \delta$ implies

$$\|f(\epsilon_{\text{machine}})\| \le C \cdot \frac{1}{2C} = \frac{1}{2},$$

and

$$\begin{aligned}
\left\| \frac{1}{1 + f(\epsilon_{\text{machine}})} - 1 \right\| &= \left\| \frac{f(\epsilon_{\text{machine}})}{1 + f(\epsilon_{\text{machine}})} \right\| \\
&\le \frac{1}{|1 - \|f(\epsilon_{\text{machine}})\||} \cdot \|f(\epsilon_{\text{machine}})\| \\
&\le 2 \cdot C\epsilon_{\text{machine}}.
\end{aligned}$$

Hence, $(1 + f(\epsilon_{\text{machine}}))^{-1} - 1 = O(\epsilon_{\text{machine}})$, so $(1 + O(\epsilon_{\text{machine}}))^{-1} = 1 + O(\epsilon_{\text{machine}})$.

■

Lecture 15

More on Stability

15.1 Each of the following problems describes an algorithm implemented on a computer satisfying the axioms (13.5) and (13.7). For each one, state whether the algorithm is *backward stable*, *stable but not backward stable*, or *unstable*, and prove it or at least give a reasonably convincing argument. Be sure to follow the definitions as given in the text.

(a) Data: $x \in \mathbb{C}$. Solution: $2x$, computed as $x \oplus x$.

(c) Data: $x \in \mathbb{C} \setminus \{0\}$. Solution: 1, computed as $x \oslash x$. (A machine satisfying (13.6) will give exactly the right answer, but our definitions are based on the weaker condition (13.7).)

Solution. By (13.5), we have $\mathrm{fl}(x) = x(1+\epsilon_1)$ for some $|\epsilon_1| \le \epsilon_{\text{machine}}$.

(a) By (13.7),

$$\begin{aligned}\tilde{f}(x) &= \mathrm{fl}(x) \oplus \mathrm{fl}(x) \\ &= (x(1+\epsilon_1) + x(1+\epsilon_1))(1+\epsilon_2) \\ &= 2x(1+\epsilon_1+\epsilon_2+\epsilon_1\epsilon_2)\end{aligned}$$

for some $|\epsilon_2| \le \epsilon_{\text{machine}}$. $f(\tilde{x}) = 2\tilde{x}$ where $\tilde{x} = x(1+\epsilon_1+\epsilon_2+\epsilon_1\epsilon_2)$ satisfies

$$\frac{|\tilde{x}-x|}{|x|} = |\epsilon_1+\epsilon_2+\epsilon_1\epsilon_2| \le |\epsilon_1| + |\epsilon_2| + |\epsilon_1\epsilon_2| \le 2\epsilon_{\text{machine}} + \epsilon_{\text{machine}}^2.$$

So, we can always find $\tilde{x}$ with

$$\frac{|\tilde{x}-x|}{|x|} = O(\epsilon_{\text{machine}}) \qquad (*)$$

(by Exercise 14.2) such that $\tilde{f}(x) = f(\tilde{x})$ for each $x \in \mathbb{C}$. The algorithm is backward stable.

(c) Since the solution is a constant, $f(\tilde{x}) = 1$. By (13.7),

$$\tilde{f}(x) = \mathrm{fl}(x) \oslash \mathrm{fl}(x) = \frac{x(1+\epsilon_1)}{x(1+\epsilon_1)}(1+\epsilon_2) = 1+\epsilon_2$$

for some $|\epsilon_2| \le \epsilon_{\text{machine}}$. For each $x \in \mathbb{C} \setminus \{0\}$, we cannot always find $\tilde{x}$ with $(*)$ such that $\tilde{f}(x) = f(\tilde{x})$,

so the algorithm is not backward stable. However, it is stable because, for each $x \in \mathbb{C} \setminus \{0\}$,

$$\frac{\left|\tilde{f}(x) - f(\tilde{x})\right|}{|f(\tilde{x})|} = O(\epsilon_{\text{machine}})$$

for some $\tilde{x}$ with $(*)$.

■

15.2 Consider an algorithm for the problem of computing the (full) SVD of a matrix. The data for this problem is a matrix A, and the solution is three matrices U (unitary), Σ (diagonal), and V (unitary) such that $A = U\Sigma V^*$. (We are speaking here of explicit matrices U and V, not implicit representations as products of reflectors.)

(a) Explain what it would mean for this algorithm to be backward stable.

(b) In fact, for a simple reason, this algorithm cannot be backward stable. Explain.

(c) Fortunately, the standard algorithms for computing the SVD (Lecture 31) are stable. Explain what stability means for such an algorithm.

Solution. Let $A \in \mathbb{C}^{m\times n}$. Let $\tilde{U}$, $\tilde{\Sigma}$, and $\tilde{V}$ denote the computed matrices.

(a) The algorithm is *backward stable* if, for each $A \in \mathbb{C}^{m\times n}$,

$$\tilde{U} = U', \quad \tilde{\Sigma} = \Sigma', \quad \tilde{V} = V',$$

where $\tilde{A} = U'\Sigma'V'^*$ is the SVD for some $\tilde{A} \in \mathbb{C}^{m\times n}$ with $\left\|\tilde{A} - A\right\|/\|A\| = O(\epsilon_{\text{machine}})$.

(b) Regardless of $\tilde{A}$, U' and V' are unitary, and Σ' is a diagonal matrix. However, due to floating point errors, the computed matrices $\tilde{U}$, $\tilde{\Sigma}$, and $\tilde{V}$ may not have such properties.

(c) The algorithm is *stable* if, for each $A \in \mathbb{C}^{m\times n}$,

$$\frac{\left\|\tilde{U} - U'\right\|}{\|U'\|} = O(\epsilon_{\text{machine}}),$$
$$\frac{\left\|\tilde{\Sigma} - \Sigma'\right\|}{\|\Sigma'\|} = O(\epsilon_{\text{machine}}),$$
$$\frac{\left\|\tilde{V} - V'\right\|}{\|V'\|} = O(\epsilon_{\text{machine}}),$$

where $\tilde{A} = U'\Sigma'V'^*$ is the SVD for some $\tilde{A} \in \mathbb{C}^{m\times n}$ with $\left\|\tilde{A} - A\right\|/\|A\| = O(\epsilon_{\text{machine}})$.

■

Lecture 16

Stability of Householder Triangularization

16.1 (a) Let unitary matrices $Q_1, \cdots, Q_k \in \mathbb{C}^{m\times m}$ be fixed and consider the problem of computing, for $A \in \mathbb{C}^{m\times m}$, the product $B = Q_k \cdots Q_1 A$. Let the computation be carried out from right to left by straightforward floating point operations on a computer satisfying (13.5) and (13.7). Show that this algorithm is backward stable. (Here A is thought of as data that can be perturbed; the matrices Q are fixed and not to be perturbed.)

Solution. The proof is by induction. We first consider the primitive problem of computing, for $A \in \mathbb{C}^{m\times m}$, the product $B = f(A) = QA$ where a unitary $Q \in \mathbb{C}^{m\times m}$ is fixed. Let $\tilde{B} = \tilde{f}(A)$ denote the computed product. For each $A \in \mathbb{C}^{m\times m}$, we shall find some $\delta A \in \mathbb{C}^{m\times m}$ with $\|\delta A\|/\|A\| = O(\epsilon_{\text{machine}})$ such that $\tilde{B} = Q(A + \delta A)$. Note that $\|\delta A\| = \left\|Q^*\tilde{B} - A\right\| = \left\|\tilde{B} - QA\right\| = \|A\|O(\epsilon_{\text{machine}})$. We need to show the last equality.

Recall from Exercise 14.2 that $(1 + O(\epsilon_{\text{machine}}))(1 + O(\epsilon_{\text{machine}})) = 1 + O(\epsilon_{\text{machine}})$. Denote $X = [x_{ij}]$ for $X \in \left\{A, B, \tilde{B}, Q\right\}$.

$$\begin{aligned}
\tilde{b}_{ij} &= \bigoplus_{k=1}^{m} \text{fl}(\text{fl}(q_{ik}) \otimes \text{fl}(a_{kj})) \\
&= \bigoplus_{k=1}^{m} \text{fl}\left(q_{ik}a_{kj}(1 + O(\epsilon_{\text{machine}}))\right) \\
&= \sum_{k=1}^{m} q_{ik}a_{kj}(1 + O(\epsilon_{\text{machine}})).
\end{aligned}$$

By Theorem 14.1, all norms on a finite-dimensional vector space are equivalent, so we can consider $\|\cdot\| = \|\cdot\|_F$.

$$\begin{aligned}
\left\|\tilde{B} - QA\right\|_F &= \left[\sum_{i=1}^{m}\sum_{j=1}^{m}\left|\tilde{b}_{ij} - \sum_{k=1}^{m} q_{ik}a_{kj}\right|^2\right]^{1/2} \\
&= \left[\sum_{i=1}^{m}\sum_{j=1}^{m}\left|\sum_{k=1}^{m} q_{ik}a_{kj}O(\epsilon_{\text{machine}})\right|^2\right]^{1/2} \\
&\le \left[\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}|q_{ik}|^2\sum_{k=1}^{m}|a_{kj}|^2\right]^{1/2} O(\epsilon_{\text{machine}}) \\
&= \left[\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}|a_{kj}|^2\right]^{1/2} O(\epsilon_{\text{machine}})
\end{aligned}$$

$$= \sqrt{m}\|A\|_F O(\epsilon_{\text{machine}}).$$

The inequality follows from the Cauchy–Schwarz inequality. $\sum_{k=1}^{m} |q_{ik}|^2 = 1$ because Q is unitary. Hence, $\|\delta A\| = \|A\| O(\epsilon_{\text{machine}})$, so the primitive algorithm that computes QA is backward stable.

Suppose that the algorithm that computes $B_k = Q_k \cdots Q_1 A$ is backward stable. Thus, $\tilde{B}_k = Q_k \cdots Q_1 (A + \delta A)$ for some $\delta A \in \mathbb{C}^{m \times m}$ with $\|\delta A\|/\|A\| = O(\epsilon_{\text{machine}})$. By the backward stable primitive algorithm, $\tilde{B}_{k+1} = Q_{k+1}\left(\tilde{B}_k + \delta \tilde{B}_k\right)$ for some $\delta \tilde{B}_k \in \mathbb{C}^{m \times m}$ with

$$\begin{aligned}
\left\|\delta \tilde{B}_k\right\| &= \left\|\tilde{B}_k\right\| O(\epsilon_{\text{machine}}) \\
&= \|Q_k \cdots Q_1 (A + \delta A)\| O(\epsilon_{\text{machine}}) \\
&= \|A + \delta A\| O(\epsilon_{\text{machine}}) \\
&\le \|A\| O(\epsilon_{\text{machine}}) + \|\delta A\| O(\epsilon_{\text{machine}}) \\
&= \|A\| O(\epsilon_{\text{machine}}).
\end{aligned}$$

Thus,

$$\tilde{B}_{k+1} = Q_{k+1}\left(Q_k \cdots Q_1 (A + \delta A) + \delta \tilde{B}_k\right) = Q_{k+1} \cdots Q_1 (A + \Delta A),$$

where $\Delta A = \delta A + (Q_1^* \cdots Q_k^*)\, \delta \tilde{B}_k$ with

$$\begin{aligned}
\|\Delta A\| &\le \|\delta A\| + \left\|(Q_1^* \cdots Q_k^*)\, \delta \tilde{B}_k\right\| \\
&= \|\delta A\| + \left\|\delta \tilde{B}_k\right\| \\
&= \|A\| O(\epsilon_{\text{machine}}).
\end{aligned}$$

Therefore, the algorithm is backward stable. ■

Lecture 17

Stability of Back Substitution

17.2 A triangular system (17.1) is solved by back substitution. Exactly what does Theorem 17.1 imply about the error $\|\tilde{x} - x\|$?

Solution. Theorem 17.1 states that back substitution (Algorithm 17.1) applied to solve the problem (17.1) consisting of floating point numbers on a computer satisfying (13.7) is backward stable. Assume that the computer also satisfies (13.5). By Theorem 15.1, $\|\tilde{x} - x\| = \|x\| \cdot O(\kappa(R)\epsilon_{\text{machine}})$. ■

Lecture 19

Stability of Least Squares Algorithms

19.2 Here is a stripped-down version of one of MATLAB's built-in m-files.

```
[U,S,V] = svd(A);
S = diag(S);
tol = max(size(A))*S(1)*eps;
r = sum(S > tol);
S = diag(ones(r,1)./S(1:r));
X = V(:,1:r)*S*U(:,1:r)';
```

What does this program compute?

Solution. Let $r = \operatorname{rank}(A)$ and let

$$A = U\Sigma V^* = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^*$$

be a full SVD of A where

$$\begin{aligned} U_1 &\in \mathbb{C}^{m\times r}, \\ U_2 &\in \mathbb{C}^{m\times(m-r)}, \\ \Sigma_1 &\in \mathbb{C}^{r\times r}, \\ V_1 &\in \mathbb{C}^{n\times r}, \\ V_2 &\in \mathbb{C}^{n\times(n-r)}. \end{aligned}$$

By Theorem 5.7, $A = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^* = U_1\Sigma_1 V_1^*$ is a compact SVD of A. `tol` is the tolerance to compute $\mathtt{r} = \operatorname{rank}(A)$. This program computes

$$X = V_1 \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_r \end{bmatrix} U_1^* = V_1 {\Sigma_1}^{-1} U_1^* = A^+,$$

the pseudoinverse of A. ■

Part IV

Systems of Equations

Lecture 20

Gaussian Elimination

20.3 Suppose an $m \times m$ matrix A is written in the block form $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where A_{11} is $n \times n$ and A_{22} is $(m-n) \times (m-n)$.
Assume that A satisfies the condition of Exercise 20.1.

(a) Verify the formula

$$\begin{bmatrix} I & \\ -A_{21}{A_{11}}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} - A_{21}{A_{11}}^{-1}A_{12} \end{bmatrix}$$

for "elimination" of the block A_{21}. The matrix $A_{22} - A_{21}{A_{11}}^{-1}A_{12}$ is known as the *Schur complement* of A_{11} in A.

(b) Suppose A_{21} is eliminated row by row by means of n steps of Gaussian elimination. Show that the bottom-right $(m-n) \times (m-n)$ block of the result is again $A_{22} - A_{21}{A_{11}}^{-1}A_{12}$.

Solution.

(a) A satisfies the condition of Exercise 20.1, so A_{11} is nonsingular. The partitions are conformable as all products and sums involved are defined. By direct computation,

$$\begin{aligned} &\begin{bmatrix} I_{n\times n} & \mathbf{0}_{n\times(m-n)} \\ -A_{21}{A_{11}}^{-1} & I_{(m-n)\times(m-n)} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \\ &= \begin{bmatrix} I_{n\times n}A_{11} + \mathbf{0}_{n\times(m-n)}A_{21} & I_{n\times n}A_{12} + \mathbf{0}_{n\times(m-n)}A_{22} \\ -A_{21}{A_{11}}^{-1}A_{11} + I_{(m-n)\times(m-n)}A_{21} & -A_{21}{A_{11}}^{-1}A_{12} + I_{(m-n)\times(m-n)}A_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0}_{(m-n)\times n} & A_{22} - A_{21}{A_{11}}^{-1}A_{12} \end{bmatrix}. \end{aligned}$$

(b) Suppose that

$$L_nL_{n-1}\cdots L_2L_1A = \begin{bmatrix} B & \mathbf{0}_{n\times(m-n)} \\ C & I_{(m-n)\times(m-n)} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I_{n\times n} & A_{12} \\ \mathbf{0}_{(m-n)\times n} & X \end{bmatrix}.$$

We have $CA_{11} + A_{21} = \mathbf{0}$. Since A_{11} is nonsingular, $C = -A_{21}{A_{11}}^{-1}$. Thus, the bottom-right $(m-n) \times (m-n)$ block of the result is $X = CA_{12} + A_{22} = A_{22} - A_{21}{A_{11}}^{-1}A_{12}$. ■

20.4 Like most of the algorithms in this book, Gaussian elimination involves a triply nested loop. In Algorithm 20.1, there are two explicit **for** loops, and the third loop is implicit in the vectors $u_{j,k:m}$ and $u_{k,k:m}$ Rewrite this algorithm with just one explicit **for** loop indexed by k. Inside this loop, U will be updated at each step by a certain rank-one outer product. This "outer product" form of Gaussian elimination may be a better starting point than Algorithm 20.1 if one wants to optimize computer performance.

Solution.

$U = A$, $L = I$
for $k = 1$ **to** $m - 1$ **do**
  $\mathbf{l}_{k+1:m,k} \leftarrow \mathbf{u}_{k+1:m,k}/u_{kk}$
  $U_{k+1:m,k:m} \leftarrow U_{k+1:m,k:m} - \mathbf{l}_{k+1:m,k}\mathbf{u}_{k,k:m}$

■

Lecture 21

Pivoting

21.1 Let A be the 4×4 matrix (20.3) considered in this lecture and the previous one.

(a) Determine $\det A$ from (20.5).

(b) Determine $\det A$ from (21.3).

(c) Describe how Gaussian elimination with partial pivoting can be used to find the determinant of a general square matrix.

Solution. For $A, B \in \mathbb{C}^{m\times m}$, $\det(AB) = (\det A)(\det B)$. The determinant of a triangular matrix is the product of the diagonal entries.

(a) $\det A = \det(LU) = (\det L)(\det U) = (1 \cdot 1 \cdot 1 \cdot 1) \cdot (2 \cdot 1 \cdot 2 \cdot 2) = 8$.

(b) $\det P = (-1)^3$ because rows are interchanged three times. It follows from $PA = LU$ that

$$\begin{aligned} \det(PA) &= \det(LU) \\ (\det P)(\det A) &= (\det L)(\det U) \\ -\det A &= (1 \cdot 1 \cdot 1 \cdot 1) \cdot \left(8 \cdot \frac{7}{4} \cdot \left(-\frac{6}{7}\right) \cdot \frac{2}{3}\right) \\ \therefore \det A &= 8. \end{aligned}$$

(c) Compute an LU factorization of $A \in \mathbb{C}^{m\times m}$. If rows are interchanged k times, $\det P = (-1)^k$. Therefore,

$$\det A = \frac{(\det L)(\det U)}{\det P} = (-1)^k \prod_{j=1}^{m} l_{jj} u_{jj}.$$

■

21.3 Consider Gaussian elimination carried out with pivoting by columns instead of rows, leading to a factorization $AQ = LU$, where Q is a permutation matrix.

(b) Show that if A is singular, such a factorization does not always exist.

Solution. Let $A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$; then A is singular. Note that $AQ = A$ for any permutation matrix $Q \in \{0,1\}^{2\times 2}$. Suppose that there is a factorization:

$$AQ = LU = \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}.$$

$u_{11} = 0$ and $u_{12} = 0$. However, we must have $l_{21}u_{11} = 1$. This is a contradiction. Therefore, if A is singular, such a factorization does not always exist. ■

21.6 Suppose $A \in \mathbb{C}^{m\times m}$ is *strictly column diagonally dominant*, which means that for each k,

$$|a_{kk}| > \sum_{j\neq k} |a_{jk}|. \tag{21.11}$$

Show that if Gaussian elimination with partial pivoting is applied to A, no row interchanges take place.

Solution. $|a_{11}| = \max\{|a_{1k}| : 1 \le k \le m\}$, so no row interchange takes place for the first step. Let

$$L_1 A = \begin{bmatrix} 1 & & & \\ -l_{21} & 1 & & \\ \vdots & & \ddots & \\ -l_{m1} & & & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{1,2:m} \\ \mathbf{0} & A_2 \end{bmatrix},$$

where

$$A_2 = \begin{bmatrix} a_{22} - l_{21}a_{12} & a_{23} - l_{21}a_{13} & \cdots & a_{2m} - l_{21}a_{1m} \\ a_{32} - l_{31}a_{12} & a_{33} - l_{31}a_{13} & \cdots & a_{3m} - l_{31}a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m2} - l_{m1}a_{12} & a_{m3} - l_{m1}a_{13} & \cdots & a_{mm} - l_{m1}a_{1m} \end{bmatrix}.$$

For each $2 \le k \le m$,

$$\begin{aligned} \sum_{\substack{j\neq k \\ j\ge 2}} |a_{jk} - l_{j1}a_{1k}| &\le \sum_{\substack{j\neq k \\ j\ge 2}} |a_{jk}| + |a_{1k}| \sum_{\substack{j\neq k \\ j\ge 2}} |l_{j1}| \\ &< |a_{kk}| - |a_{1k}| + \frac{|a_{1k}|}{|a_{11}|} \sum_{\substack{j\neq k \\ j\ge 2}} |a_{j1}| \\ &< |a_{kk}| - |a_{1k}| + \frac{|a_{1k}|}{|a_{11}|}(|a_{11}| - |a_{k1}|) \\ &= |a_{kk}| - |l_{k1}a_{1k}| \\ &< |a_{kk} - l_{k1}a_{1k}|. \end{aligned}$$

Hence, $A_2 \in \mathbb{C}^{(m-1)\times(m-1)}$ is a also strictly column diagonally dominant matrix. No row interchange takes place for the second step. By induction, we conclude that no row interchanges take place during the $m-1$ steps of Gaussian elimination with partial pivoting. ■

Lecture 22

Stability of Gaussian Elimination

22.1 Show that for Gaussian elimination with partial pivoting applied to any matrix $A \in \mathbb{C}^{m\times m}$, the growth factor (22.2) satisfies $\rho \le 2^{m-1}$.

Solution. Gaussian elimination with partial pivoting is equivalent to the following procedure:

1. Permute the rows of A according to P.
2. Apply Gaussian elimination without pivoting to PA.

The first step preserves $\max_{i,j}|a_{ij}| = \max_{i,j}\left|a_{ij}^{(0)}\right|$. Let

$$PA = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1m}^{(0)} \\ a_{21}^{(0)} & a_{22}^{(0)} & a_{23}^{(0)} & \cdots & a_{2m}^{(0)} \\ a_{31}^{(0)} & a_{32}^{(0)} & a_{33}^{(0)} & \cdots & a_{3m}^{(0)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1}^{(0)} & a_{m2}^{(0)} & a_{m3}^{(0)} & \cdots & a_{mm}^{(0)} \end{bmatrix} \to L_1 PA = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1m}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2m}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3m}^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2}^{(1)} & a_{m3}^{(1)} & \cdots & a_{mm}^{(1)} \end{bmatrix},$$

where $a_{ij}^{(1)} = a_{ij}^{(0)} - \dfrac{a_{i1}^{(0)}}{a_{11}^{(0)}} a_{1j}^{(0)}$. By partial pivoting, we have $\left|\dfrac{a_{i1}^{(0)}}{a_{11}^{(0)}}\right| \le 1$, so

$$\left|a_{ij}^{(1)}\right| = \left|a_{ij}^{(0)} - \frac{a_{i1}^{(0)} a_{1j}^{(0)}}{a_{11}^{(0)}}\right| \le \left|a_{ij}^{(0)}\right| + \left|\frac{a_{i1}^{(0)}}{a_{11}^{(0)}}\right|\left|a_{1j}^{(0)}\right| \le \left|a_{ij}^{(0)}\right| + \left|a_{1j}^{(0)}\right| \le 2\max_{i,j}\left|a_{ij}^{(0)}\right| = 2\max_{i,j}|a_{ij}|.$$

Repeating the process, after the first k steps of Gaussian elimination,

$$\left|a_{ij}^{(k)}\right| \le \left|a_{ij}^{(k-1)}\right| + \left|a_{kj}^{(k-1)}\right| \le 2\max_{i,j}\left|a_{ij}^{(k-1)}\right|.$$

$m-1$ steps of Gaussian elimination are required to form U. Thus,

$$|u_{ij}| = \left|a_{ij}^{(m-1)}\right| \le 2\max_{i,j}\left|a_{ij}^{(k-2)}\right| \le 2^2 \max_{i,j}\left|a_{ij}^{(k-3)}\right| \le \cdots \le 2^{m-1}\max_{i,j}\left|a_{ij}^{(0)}\right| = 2^{m-1}\max_{i,j}|a_{ij}|.$$

Therefore, the growth factor ρ satisfies

$$\rho = \frac{\max_{i,j}|u_{ij}|}{\max_{i,j}|a_{ij}|} \le 2^{m-1}.$$

■

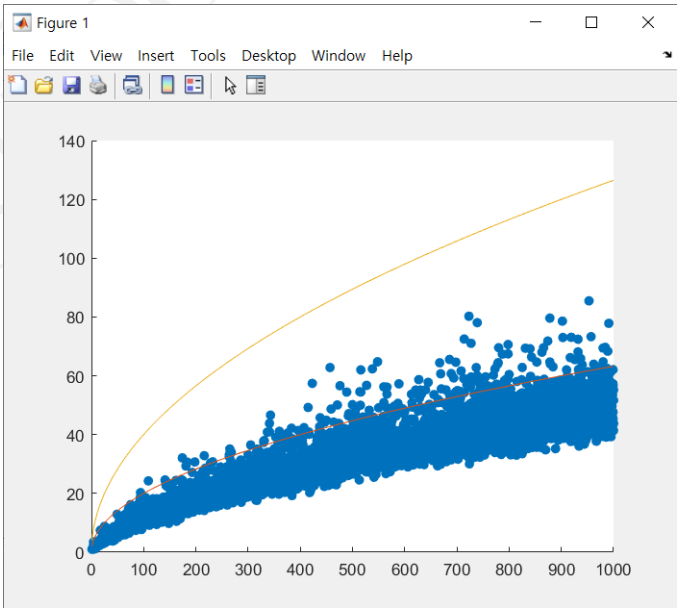
22.3 Reproduce the figures of this lecture, approximately if not in full detail, but based on random matrices with entries uniformly distributed in $[-1, 1]$ rather than normally distributed. Do you see any significant differences?

Solution. This MATLAB program computes LU factorizations of 5 random matrices with entries uniformly distributed in $[-1, 1]$ for each dimension m between 1 and 1000. In the figure below, the growth factors are presented as a scatter plot. The horizontal and vertical axes represent the dimension m and the growth factor ρ, respectively. The orange and yellow curves represent $2\sqrt{m}$ and $4\sqrt{m}$, respectively. We have obtained a similar result as Figure 22.1; the typical size of ρ is of order $\sqrt{m}$, which is much less than the maximal possible value 2^{m-1}. ■

```
X = [];
Y = [];
x = 1:1000;
y1 = 2*sqrt(x);
y2 = 4*sqrt(x);

for m=1:1000
    for i=1:5
        A = -1 + 2*rand(m);
        [L, U] = lu(A);
        gf = max(max(abs(U))) / max(max(abs(A)));
        X = [X m];
        Y = [Y gf];
    end
end

scatter(X, Y, 'filled')
hold on
plot(x, y1)
plot(x, y2)
```



Lecture 23

Cholesky Factorization

23.1 Let A be a nonsingular square matrix and let $A = QR$ and $A^*A = U^*U$ be QR and Cholesky factorizations, respectively, with the usual normalizations $r_{jj},\ u_{jj} > 0$. Is it true or false that $R = U$?

Solution. Since A is nonsingular, it has a unique QR factorization with $r_{jj} > 0$ by Theorem 7.2. Q is unitary, so we have a Cholesky factorization $A^*A = R^*Q^*QR = R^*R$ of A^*A, where R is upper-triangular.

A^*A is obviously hermitian. Since A is nonsingular, $A\mathbf{x} \neq \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$. We have

$$\mathbf{x}^*A^*A\mathbf{x} = (A\mathbf{x})^*(A\mathbf{x}) = \|A\mathbf{x}\|_2^2 > 0$$

for all $\mathbf{x} \neq \mathbf{0}$. Thus, A^*A is positive definite. By Theorem 23.1, A^*A has a unique Cholesky factorization $A^*A = U^*U$ with $u_{jj} > 0$, where U is upper-triangular. Therefore, $R = U$. ■

Part V

Eigenvalues

Lecture 24

Eigenvalue Problems

24.2 Here is *Gerschgorin's theorem*, which holds for any $m \times m$ matrix A, symmetric or nonsymmetric. *Every eigenvalue of A lies in at least one of the m circular disks in the complex plane with centers a_{ii} and radii $\sum_{j \neq i} |a_{ij}|$. Moreover, if n of these disks form a connected domain that is disjoint from the other $m - n$ disks, then there are precisely n eigenvalues of A within this domain.*

(a) Prove the first part of Gerschgorin's theorem. (Hint: Let λ be any eigenvalue of A, and $\mathbf{x}$ a corresponding eigenvector with largest entry 1.)

(b) Prove the second part. (Hint: Deform A to a diagonal matrix and use the fact that the eigenvalues of a matrix are continuous functions of its entries.)

(c) Give estimates based on Gerschgorin's theorem for the eigenvalues of

$$A = \begin{bmatrix} 8 & 1 & 0 \\ 1 & 4 & \epsilon \\ 0 & \epsilon & 1 \end{bmatrix}, \quad |\epsilon| < 1.$$

(d) Find a way to establish the tighter bound $|\lambda_3 - 1| \leq \epsilon^2$ on the smallest eigenvalue of A. (Hint: Consider diagonal similarity transformations.)

Solution. The problem implicitly assumes that $A \in \mathbb{R}^{m \times m}$, but the theorem holds for any $A \in \mathbb{C}^{m \times m}$ in general.

(a) Let $\lambda \in \mathbb{C}$ be an eigenvalue of A. Let $\mathbf{x} \in \mathbb{C}^m \setminus \{\mathbf{0}\}$ be an eigenvector corresponding to λ with $\|\mathbf{x}\|_\infty = |x_k|$. Note that $x_k \neq 0$ because $\mathbf{x} \neq \mathbf{0}$. Since $A\mathbf{x} = \lambda\mathbf{x}$,

$$\begin{aligned} \lambda x_k &= \sum_{j=1}^{m} a_{kj} x_j \\ \lambda - a_{kk} &= \sum_{\substack{j=1 \\ j \neq k}}^{m} \frac{a_{kj} x_j}{x_k} \\ |\lambda - a_{kk}| &= \left| \sum_{\substack{j=1 \\ j \neq k}}^{m} \frac{a_{kj} x_j}{x_k} \right| \leq \sum_{\substack{j=1 \\ j \neq k}}^{m} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1 \\ j \neq k}}^{m} |a_{kj}|. \end{aligned}$$

λ lies in a circular disk in the complex plane centered at a_{kk} with radius $\sum_{j \neq k} |a_{kj}|$.

(b) Note: *There are two types of eigenvalue continuity: topological and functional. For rigorous proof of the second part, we should not just take eigenvalue continuity for granted. For a brief discussion and clarification, refer to* [1]. *The proof here is from* [2] *and does not require any type of eigenvalue continuity.*

Let $D = \text{diag}(a_{11}, \cdots, a_{mm})$ be the diagonal matrix that captures the main diagonal of A. Let $A(t) = D + t(A - D) = (1 - t)D + tA$. Let Γ be a simple closed rectifiable curve in the complex plane that surrounds the domain formed by n Gerschgorin disks. Γ is disjoint from the other $m - n$ disks and does not pass through any eigenvalue of any $A(t)$. Let $p_t(z)$ denote the characteristic polynomial of $A(t)$. For each $t \in [0, 1]$, $0 \notin p_t(\Gamma)$.

By the argument principle, the number of zeros—counted with algebraic multiplicities—of $p_t(z)$ inside Γ is

$$N(t) = \frac{1}{2\pi i} \oint_\Gamma \frac{p_t'(z)}{p_t(z)}\, dz.$$

The integrand is an analytic function of z in a neighborhood of Γ for each $t \in [0, 1]$. By Leibniz's rule, $N(t)$ is continuous on $[0, 1]$. Since $N(t)$ is an integer, it must be a constant function there. Thus, $n = N(0) = N(1)$ is the number of eigenvalues of A inside Γ. By the first part, there are precisely n eigenvalues of A within this domain.

(c) Note that A is symmetric, so all eigenvalues of A are real. (See Exercise 2.3(a).) Given that $|\epsilon| < 1$, $\{z \in \mathbb{C} : |z - 8| \le 1\}$, $\{z \in \mathbb{C} : |z - 4| \le 1 + |\epsilon|\}$, and $\{z \in \mathbb{C} : |z - 1| \le |\epsilon|\}$ are disjoint. By the theorem, there is exactly one eigenvalue of A within each disk. Hence, $|\lambda_1 - 8| \le 1$, $|\lambda_2 - 4| \le 1 + |\epsilon| < 2$, and $|\lambda_3 - 1| \le |\epsilon| < 1$.

(d) Consider the diagonal similarity transformation

$$DAD^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \epsilon \end{bmatrix} \begin{bmatrix} 8 & 1 & 0 \\ 1 & 4 & \epsilon \\ 0 & \epsilon & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/\epsilon \end{bmatrix} = \begin{bmatrix} 8 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & \epsilon^2 & 1 \end{bmatrix}.$$

By Theorem 24.3, similar matrices have the same eigenvalues. Therefore, we establish the tighter bound $|\lambda_3 - 1| \le \epsilon^2$ on the smallest eigenvalue of A.

■

[1] Li, C. K., & Zhang, F. (2019). Eigenvalue continuity and Geršgorin's theorem. *The Electronic Journal of Linear Algebra, 35*, 619-625.

[2] Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Lecture 25

Overview of Eigenvalue Algorithms

25.3 Suppose we have a 3×3 matrix and wish to introduce zeros by left- and/or right-multiplications by unitary matrices Q_j such as Householder reflectors or Givens rotations. Consider the following three matrix structures:

$$\text{(a)} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}, \qquad \text{(b)} \begin{bmatrix} \times & \times & 0 \\ \times & 0 & \times \\ 0 & \times & \times \end{bmatrix}, \qquad \text{(c)} \begin{bmatrix} \times & \times & 0 \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix}.$$

For each one, decide which of the following situations holds, and justify your claim.

(i) Can be obtained by a sequence of left-multiplications by matrices Q_j;

(ii) Not (i), but can be obtained by a sequence of left- and right-multiplications by matrices Q_j;

(iii) Cannot be obtained by any sequence of left- and right-multiplications by matrices Q_j.

Solution. Let $A \in \mathbb{C}^{3\times 3}$ be an arbitrary matrix. The product of finitely many unitary matrices is unitary. If a matrix structure can be obtained by a sequence of left-multiplications by unitary matrices Q_j, then it is Q^*A for some unitary matrix Q. Let

$$Q^*A = \begin{bmatrix} \mathbf{q}_1^* \\ \mathbf{q}_2^* \\ \mathbf{q}_3^* \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^*\mathbf{a}_1 & \mathbf{q}_1^*\mathbf{a}_2 & \mathbf{q}_1^*\mathbf{a}_3 \\ \mathbf{q}_2^*\mathbf{a}_1 & \mathbf{q}_2^*\mathbf{a}_2 & \mathbf{q}_2^*\mathbf{a}_3 \\ \mathbf{q}_3^*\mathbf{a}_1 & \mathbf{q}_3^*\mathbf{a}_2 & \mathbf{q}_3^*\mathbf{a}_3 \end{bmatrix}.$$

(a) We have $\mathbf{q}_3^*\mathbf{a}_1 = \mathbf{q}_3^*\mathbf{a}_2 = 0$, so $\mathbf{q_3} \in \operatorname{span}\{\mathbf{a}_1, \mathbf{a}_2\}^\perp$. Suppose that $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly independent. We also have $\mathbf{q}_2^*\mathbf{a}_1 = \mathbf{q}_2^*\mathbf{q}_3 = 0$, so $\mathbf{q_2} \in \operatorname{span}\{\mathbf{a}_1, \mathbf{q}_3\}^\perp$. $\mathbf{a}_1$ and $\mathbf{q}_3$ are linearly independent. Then the direction of $\mathbf{q_1} \in \operatorname{span}\{\mathbf{q}_2, \mathbf{q}_3\}^\perp$ is uniquely determined. Hence, we cannot guarantee $\mathbf{q}_1^*\mathbf{a}_3 = 0$ for an arbitrary matrix A. (i) does not hold.

Using Householder reflectors, we can introduce zeros in the specified positions.

$$A \xrightarrow{Q_1^*\cdot} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{Q_2^*\cdot} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{\cdot Q_3} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}.$$

Therefore, (ii) holds.

(b) Let $\mathbf{a} = \mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}_3 \neq \mathbf{0}$. We have linearly independent vectors $\mathbf{q}_1$, $\mathbf{q}_2$, and $\mathbf{q}_3$ such that $\mathbf{q}_1^*\mathbf{a} = \mathbf{q}_2^*\mathbf{a} = \mathbf{q}_3^*\mathbf{a} = 0$. All of them belong to $\text{span}\{\mathbf{a}\}^\perp$, but $\dim\left(\text{span}\{\mathbf{a}\}^\perp\right) = 2$. Thus, (i) is impossible.

Using a Householder reflector, we can introduce a zero in the $(3,1)$ position.

$$A \xrightarrow{Q_1^*\cdot} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Let $A_{12} = U\Sigma V^*$ be an SVD of A_{12}.

$$A = Q_1^* \begin{bmatrix} U^* & \mathbf{0} \\ \mathbf{0}^* & 1 \end{bmatrix} \begin{bmatrix} A_{11} & U\Sigma V^* \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & V \end{bmatrix} = \begin{bmatrix} \times & \sigma_1 & 0 \\ \times & 0 & \sigma_2 \\ 0 & \times & \times \end{bmatrix}.$$

Therefore, (ii) holds.

(c) The determinant for the matrix structure is zero. Multiplications by unitary matrices do not change the determinant, so (iii) holds for nonsingular matrices in general.

■

Lecture 27

Rayleigh Quotient, Inverse Iteration

27.1 Let $A \in \mathbb{C}^{m\times m}$ be given, not necessarily hermitian. Show that a number $z \in \mathbb{C}$ is a Rayleigh quotient of A if and only if it is a diagonal entry of Q^*AQ for some unitary matrix Q. Thus Rayleigh quotients are just diagonal entries of matrices, once you transform orthogonally to the right coordinate system.

Solution.

($\Longrightarrow$) Suppose that $z \in \mathbb{C}$ is a Rayleigh quotient of A. There is a vector $\mathbf{x} \in \mathbb{C}^m$ such that

$$r_A(\mathbf{x}) = \frac{\mathbf{x}^*A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = z.$$

Let $\mathbf{x}/\|\mathbf{x}\|_2$ be the first column of a unitary matrix Q. Then we have

$$(Q^*AQ)_{11} = \frac{\mathbf{x}^*}{\|\mathbf{x}\|_2} A \frac{\mathbf{x}}{\|\mathbf{x}\|_2} = \frac{\mathbf{x}^*A\mathbf{x}}{\|\mathbf{x}\|_2^2} = \frac{\mathbf{x}^*A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = z.$$

z is a diagonal entry of Q^*AQ.

($\Longleftarrow$) If $z \in \mathbb{C}$ is a diagonal entry of Q^*AQ for some unitary matrix Q, then $z = \mathbf{q}^*A\mathbf{q}$ for some unit vector $\mathbf{q} \in \mathbb{C}^m$. Hence,

$$z = \mathbf{q}^*A\mathbf{q} = \frac{\mathbf{q}^*A\mathbf{q}}{\|\mathbf{q}\|_2^2} = \frac{\mathbf{q}^*A\mathbf{q}}{\mathbf{q}^*\mathbf{q}} = r_A(\mathbf{q}).$$

z is a Rayleigh quotient of A.

■

Lecture 28

QR Algorithm without Shifts

28.2 The preliminary reduction to tridiagonal form would be of little use if the steps of the QR algorithm did not preserve this structure. Fortunately, they do.

(a) In the QR factorization $A = QR$ of a symmetric tridiagonal matrix A, which entries of R are in general nonzero? Which entries of Q? (In practice we do not form Q explicitly.)

(b) Show that the tridiagonal structure is recovered when the product RQ is formed.

(c) Explain how Givens rotations or 2×2 Householder reflections can be used in the computation of the QR factorization of a tridiagonal matrix, reducing the operation count far below what would be required for a full matrix.

Solution. Let $A = QR$ be the QR factorization of a symmetric tridiagonal matrix $A \in \mathbb{R}^{m \times m}$.

(a) Let $\mathbf{a}_j$ and $\mathbf{q}_j$ be the jth columns of A and Q, respectively. Throughout the QR factorization, we want the sequence $\{\mathbf{q}_i\}_{i=1}^{j}$ to have the property $\text{span}\{\mathbf{q}_1, \cdots, \mathbf{q}_j\} = \text{span}\{\mathbf{a}_1, \cdots, \mathbf{a}_j\}$ for each $1 \leq j \leq m$. Hence, Q is upper-Hessenberg.

R is at least upper-triangular. By Equation (7.7), $r_{ij} = \mathbf{q}_i^* \mathbf{a}_j$ for $i \neq j$. We have $q_{ij} = 0$ for $i > j + 1$ and $a_{ij} = 0$ for $i < j - 1$ or $i > j + 1$. Thus, for $j > i + 2$, $r_{ij} = 0$.

$$R = \begin{bmatrix} \times & \times & \times & & & & \\ & \times & \times & \times & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \times & \times & \times & \\ & & & & \times & \times & \\ & & & & & \times & \end{bmatrix}.$$

(b) RQ is symmetric because $(RQ)^T = \left(Q^T A Q\right)^T = Q^T A^T Q = Q^T A Q = RQ$. It suffices to show that RQ is upper-Hessenberg. Consider $(RQ)_{ij} = \sum_{k=1}^{m} r_{ik} q_{kj}$. $r_{ik} = 0$ if $i > k$ because R is upper-triangular. $q_{kj} = 0$ if $k > j + 1$ because Q is upper-Hessenberg. Thus, $(RQ)_{ij} \neq 0$ only if $i \leq j + 1$. In other words, $(RQ)_{ij} = 0$ if $i > j + 1$. A symmetric matrix RQ is upper-Hessenberg and therefore tridiagonal.

(c) The QR factorization of an $m \times m$ tridiagonal matrix only has to introduce zeros to the $m - 1$ subdiagonal

entries. Using 2×2 Givens rotations or Householder reflections, schematically, the process looks like this:

$$\begin{bmatrix} \times & \times & & & \\ \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{bmatrix} \rightarrow \begin{bmatrix} \times & \times & \times & & \\ \mathbf{0} & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{bmatrix} \rightarrow \begin{bmatrix} \times & \times & \times & & \\ 0 & \times & \times & \times & \\ & \mathbf{0} & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \times & \times & \times & & \\ 0 & \times & \times & \times & \\ & 0 & \times & \times & \times \\ & & \mathbf{0} & \times & \times \\ & & & \times & \times \end{bmatrix} \rightarrow \begin{bmatrix} \times & \times & \times & & \\ 0 & \times & \times & \times & \\ & 0 & \times & \times & \times \\ & & 0 & \times & \times \\ & & & \mathbf{0} & \times \end{bmatrix}.$$

The usual QR factorization for a full matrix would require $O\left(m^3\right)$ flops, but this algorithm drastically reduces the operation count to $O(m)$.

■

Lecture 31

Computing the SVD

31.4 Let A be the $m \times m$ upper-triangular matrix with 0.1 on the main diagonal and 1 everywhere above the diagonal. Write a program to compute the smallest singular value of A in two ways: by calling a standard SVD software, and by forming A^*A and computing the square root of its smallest eigenvalue. Run your program for $1 \leq m \leq 30$ and plot the results as two curves on a log scale. Do the results conform to our general discussion of these algorithms?

Solution. This MATLAB program computes the smallest singular value of A using the two methods.

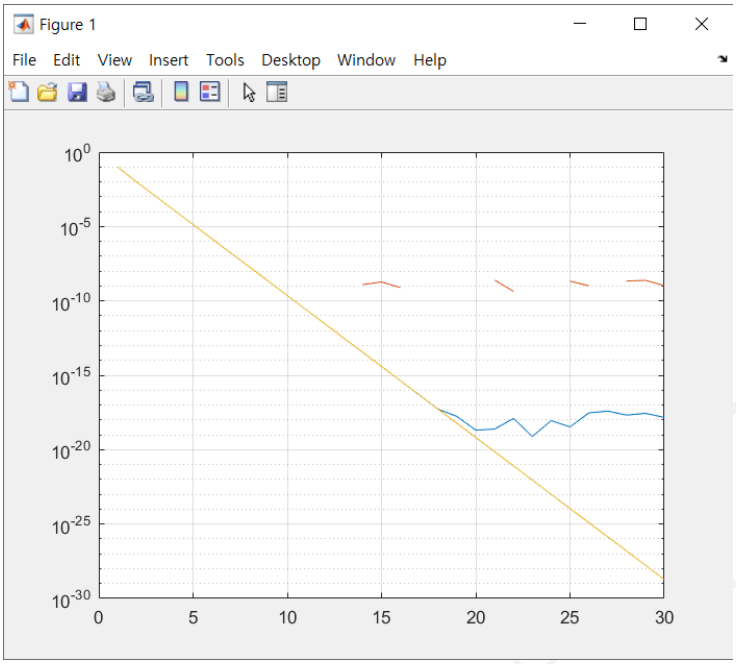
```
X = [];
Y1 = [];
Y2 = [];
Y3 = [];

for m=1:30
    A = triu(ones(m)) - 0.9*eye(m);
    X = [X, m];

    [U, S, V] = svd(A);
    Y1 = [Y1 S(m,m)];
    Y2 = [Y2 sqrt(min(eig(A'*A)))];
    Y3 = [Y3 sqrt(min(eig(sym(A)'*sym(A))))];
end

semilogy(X, Y1, X, Y2, X, Y3)
grid on
```

In the figure below, the results are presented as a semi-log plot. The horizontal and vertical axes represent the dimension m and the smallest singular value, respectively. The blue and orange curves represent the first and the second results, respectively.



We observe that some points of the orange curve are missing because the smallest eigenvalues of A^*A are computed to be negative for some m. Since A is nonsingular, A^*A is positive definite. Hence, the eigenvalues of A^*A must be positive real numbers. We conclude that the second method is numerically unstable. The first (standard) algorithm is stable as it is closer to the ideal yellow curve. Therefore, the results conform to our general discussion of these algorithms. ■

Part VI

Iterative Methods

Lecture 33

The Arnoldi Iteration

33.2 Suppose Algorithm 33.1 is executed for a particular A and $\mathbf{b}$ until at some step n, an entry $h_{n+1,n} = 0$ is encountered.

(a) Show how (33.13) can be simplified in this case. What does this imply about the structure of a full $m \times m$ Hessenberg reduction $A = QHQ^*$ of A?

(b) Show that $\mathcal{K}_n$ is an invariant subspace of A, i.e., $A\mathcal{K}_n \subseteq \mathcal{K}_n$.

(c) Show that if the Krylov subspaces of A generated by $\mathbf{b}$ are defined by $\mathcal{K}_k = \langle \mathbf{b}, A\mathbf{b}, \cdots, A^{k-1}\mathbf{b} \rangle$ as in (33.5), then $\mathcal{K}_n = \mathcal{K}_{n+1} = \mathcal{K}_{n+2} = \cdots$.

(d) Show that each eigenvalue of H_n is an eigenvalue of A.

(e) Show that if A is nonsingular, then the solution $\mathbf{x}$ to the system of equations $A\mathbf{x} = \mathbf{b}$ lies in $\mathcal{K}_n$.

The appearance of an entry $h_{n+1,n} = 0$ is called a *breakdown* of the Arnoldi iteration, but it is a breakdown of a benign sort. For applications in computing eigenvalues (Lecture 34) or solving systems of equations (Lecture 35), because of (d) and (e), a breakdown usually means that convergence has occurred and the iteration can be terminated. Alternatively, a new orthonormal vector $\mathbf{q}_{n+1}$ could be selected at random and the iteration then continued.

Solution. Let $A \in \mathbb{C}^{m \times m}$.

(a) Let H_n and $\tilde{H}_n$ be the $n \times n$ and $(n+1) \times n$ upper-left sections of H, respectively. Given $h_{n+1,n} = 0$,

$$\tilde{H}_n = \begin{bmatrix} H_n \\ \mathbf{0}^* \end{bmatrix}.$$

Then Equation (33.13) can be simplified as $AQ_n = Q_{n+1}\tilde{H}_n = Q_n H_n$. We also have

$$\text{span}\{\mathbf{q}_1, \cdots, \mathbf{q}_n\} \perp \text{span}\{\mathbf{q}_{n+1}, \cdots, \mathbf{q}_m\},$$

so H is of the form $\begin{bmatrix} H_n & O \\ O & B \end{bmatrix}$.

(b) By Equation (33.5), $\mathcal{K}_n = \operatorname{span}\left\{\mathbf{b}, A\mathbf{b}, \cdots, A^{n-1}\mathbf{b}\right\} = \operatorname{span}\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n\}$. Since

$$AQ_n = \begin{bmatrix} A\mathbf{q}_1 & A\mathbf{q}_2 & \cdots & A\mathbf{q}_n \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix} H_n = Q_n H_n,$$

we have $A\mathcal{K}_n = \operatorname{span}\{A\mathbf{q}_1, A\mathbf{q}_2, \cdots, A\mathbf{q}_n\} \subseteq \operatorname{span}\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n\} = \mathcal{K}_n$. Thus, $\mathcal{K}_n$ is an invariant subspace of A.

(c) We use strong induction. By definition, $\mathcal{K}_n \subseteq \mathcal{K}_{n+1} \subseteq \mathcal{K}_{n+2} \subseteq \cdots$.

Let $A^{n-1}\mathbf{b} = \sum_{j=1}^{n} c_j \mathbf{q}_j$; then

$$A^n\mathbf{b} = A\left(A^{n-1}\mathbf{b}\right) = \sum_{j=1}^{n} c_j A\mathbf{q}_j \in A\mathcal{K}_n \subseteq \mathcal{K}_n.$$

Hence, $\mathcal{K}_{n+1} \subseteq \mathcal{K}_n$. We first showed that $\mathcal{K}_n = \mathcal{K}_{n+1}$.

Suppose that $\mathcal{K}_n = \mathcal{K}_{n+1} = \cdots = \mathcal{K}_{n+l}$ for some $l \in \mathbb{Z}_+$. By the same argument, $A^{n+l}\mathbf{b} \in A\mathcal{K}_{n+l} = A\mathcal{K}_n \subseteq \mathcal{K}_n$, so $\mathcal{K}_{n+l+1} = \mathcal{K}_{n+l} = \mathcal{K}_n$ follows. Therefore, $\mathcal{K}_n = \mathcal{K}_{n+l}$ for each $l \in \mathbb{Z}_+$.

(d) Let θ be an eigenvalue of H_n. That is, $H_n\mathbf{x} = A\mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$.

$$H_n\mathbf{x} = Q_n^* A Q_n \mathbf{x} = \theta\mathbf{x} \iff A(Q_n\mathbf{x}) = \theta(Q_n\mathbf{x}).$$

Q_n has full rank, so $Q_n\mathbf{x} \neq \mathbf{0}$. Hence, θ is an eigenvalue of A.

(e) Let A be an $m \times m$ matrix. Let $p(z)$ be the characteristic polynomial of A. By the Cayley–Hamilton theorem,

$$p(A)\mathbf{x} = c_0\mathbf{x} + \sum_{j=1}^{m} c_j A^j\mathbf{x} = c_0\mathbf{x} + \sum_{j=1}^{m} c_j A^{j-1}\mathbf{b} = \mathbf{0}.$$

Since A is nonsingular, $c_0 = (-1)^m \det A \neq 0$.

$$\mathbf{x} = -\frac{1}{c_0}\sum_{j=1}^{m} c_j A^{j-1}\mathbf{b} \in \mathcal{K}_m = \mathcal{K}_n.$$

■

Lecture 35

GMRES

35.5 Our statement of the GMRES algorithm (Algorithm 35.1) begins with the initial guess $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$. (The same applies to CG and BCG, Algorithms 38.1 and 39.1.) Show that if one wishes to start with an arbitrary initial guess $\mathbf{x}_0$, this can be accomplished by an easy modification of the right-hand side $\mathbf{b}$.

Solution. $A\mathbf{x} = \mathbf{b} \iff A(\mathbf{x} - \mathbf{x}_0) = \mathbf{b} - A\mathbf{x}_0$. Let $\mathbf{y} = \mathbf{x} - \mathbf{x}_0$ and $\mathbf{c} = \mathbf{b} - A\mathbf{x}_0$; then use GMRES to solve $A\mathbf{y} = \mathbf{c}$ with the usual initial guess $\mathbf{y}_0 = \mathbf{0}$ and $\mathbf{r}_0 = \mathbf{c}$. We have $\mathbf{x}_n = \mathbf{y}_n + \mathbf{x}_0$. ■

Lecture 36

The Lanczos Iteration

36.1 In Lecture 27 it was pointed out that the eigenvalues of a symmetric matrix $A \in \mathbb{R}^{m\times m}$ are the stationary values of the Rayleigh quotient $r(\mathbf{x}) = (\mathbf{x}^T A\mathbf{x})/(\mathbf{x}^T\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^m$. Show that the Ritz values at step n of the Lanczos iteration are the stationary values of $r(\mathbf{x})$ if $\mathbf{x}$ is restricted to $\mathcal{K}_n$.

Solution. By (27.2),

$$\nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^T\mathbf{x}}(A\mathbf{x} - r(\mathbf{x})\mathbf{x}) = \mathbf{0} \iff A\mathbf{x} = r(\mathbf{x})\mathbf{x} \text{ with } \mathbf{x} \neq \mathbf{0}.$$

The Ritz values at step n of the Lanczos iteration are the eigenvalues of $T_n = Q_n^* A Q_n$. In Exercise 33.2, we showed that each eigenvalue of T_n is an eigenvalue of A with a corresponding eigenvector in $\mathcal{K}_n$. Thus, $\{\text{Ritz values at step } n\} = \Lambda(T_n) \subset \Lambda(A) = \{\text{stationary values of } r(\mathbf{x})\}$. ■

Lecture 38

Conjugate Gradients

38.3 The conjugate gradient is applied to a symmetric positive definite matrix A with the result $\|\mathbf{e}_0\|_A = 1$, $\|\mathbf{e}_{10}\|_A = 2 \times 2^{-10}$. Based solely on this data,

(a) What bound can you give on $\kappa(A)$?

(b) What bound can you give on $\|\mathbf{e}_{20}\|_A$?

Solution. Let $\kappa = \kappa(A)$.

(a) We know that $\kappa \geq 1$, so $\dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \geq 0$. By Theorem 38.5,

$$\begin{aligned} 2 \times 2^{-10} = \frac{\|\mathbf{e}_{10}\|_A}{\|\mathbf{e}_0\|_A} &\leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{10} = 2\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{10} \\ \frac{1}{2} &\leq 1 - \frac{2}{\sqrt{\kappa}+1} \\ \frac{2}{\sqrt{\kappa}+1} &\leq \frac{1}{2} \\ 4 &\leq \sqrt{\kappa}+1 \\ \sqrt{\kappa} &\geq 3 \\ \therefore \kappa &\geq 9. \end{aligned}$$

(b) By Theorem 38.5,

$$\|\mathbf{e}_{20}\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{20} \|\mathbf{e}_0\|_A = 2\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{20} < 2$$

because $f(x) = 1 - \dfrac{2}{x+1} < 1$ is strictly increasing on $[1, \infty)$. $\kappa \geq 9$ does not give any tighter bound on $\|\mathbf{e}_{20}\|_A$. Instead, by Theorem 38.2, $\|\mathbf{e}_{20}\|_A \leq \|\mathbf{e}_{10}\|_A = 2 \times 2^{-10}$.

■

38.5 We have described CG as an iterative minimization of the function $\varphi(\mathbf{x})$ of (38.7). Another way to minimize the same function—far slower, in general—is by the method of *steepest descent*.

(a) Derive the formula $\nabla\varphi(\mathbf{x}) = -\mathbf{r}$ for the gradient of $\varphi(\mathbf{x})$. Thus the steepest descent iteration corresponds to the choice $\mathbf{p}_n = \mathbf{r}_n$ instead of $\mathbf{p}_n = \mathbf{r}_n + \beta_n \mathbf{p}_{n-1}$ in Algorithm 38.1.

(b) Determine the formula for the optimal step length α_n of the steepest descent iteration.

(c) Write down the full steepest descent iteration. There are three operations inside the main loop.

Solution. Let $A = [a_{ij}] \in \mathbb{R}^{m\times m}$ be a symmetric matrix. Recall that $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}$.

(a) By direct computation,

$$\mathbf{x}^T A\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}^T \begin{bmatrix} \sum_{j=1}^m a_{1j}x_j \\ \sum_{j=1}^m a_{2j}x_j \\ \vdots \\ \sum_{j=1}^m a_{mj}x_j \end{bmatrix} = \sum_{i=1}^m \sum_{j=1}^m x_i a_{ij} x_j.$$

Hence,

$$\frac{\partial\left(\mathbf{x}^T A\mathbf{x}\right)}{\partial x_k} = \sum_{i=1}^m x_i a_{ik} + \sum_{j=1}^m a_{kj}x_j = \left[\left(A^T + A\right)\mathbf{x}\right]_k = 2(A\mathbf{x})_k.$$

Therefore,

$$\nabla\varphi(\mathbf{x}) = \frac{1}{2}\cdot 2A\mathbf{x} - \mathbf{b} = -\mathbf{r}.$$

(b) $\mathbf{p}_n = \mathbf{r}_n$, so we apply the formula $\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n A\mathbf{r}_{n-1}$ to compute

$$\mathbf{r}_n^T\mathbf{r}_j = \mathbf{r}_{n-1}^T\mathbf{r}_j - \alpha_n \mathbf{r}_{n-1}^T A\mathbf{r}_j.$$

We want to find α_n such that the residuals are orthogonal. If $j < n-1$, both terms on the right are zero by induction. If $j = n-1$, the right-hand side is zero provided

$$\alpha_n = \frac{\mathbf{r}_{n-1}^T\mathbf{r}_{n-1}}{\mathbf{r}_{n-1}^T A\mathbf{r}_{n-1}}.$$

(c) **Steepest Descent Iteration**

$\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$
for $n = 1, 2, 3, \cdots$ **do**
  $\alpha_n = \left(\mathbf{r}_{n-1}^T\mathbf{r}_{n-1}\right) / \left(\mathbf{r}_{n-1}^T A\mathbf{r}_{n-1}\right)$
  $\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_n\mathbf{r}_{n-1}$
  $\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n A\mathbf{r}_{n-1}$

■

Lecture 40

Preconditioning

40.1 Suppose $A = M - N$, where M is nonsingular. Suppose $\left\|I - M^{-1}A\right\|_2 \leq 1/2$, and M is used as a preconditioner as in (40.2).

(a) Show that if GMRES is applied to this preconditioned problem, then the residual norm is guaranteed to be six orders of magnitude smaller, or better, after twenty steps.

(b) How many steps of CGN are needed for the same guarantee?

Solution. Let $\|\cdot\| = \|\cdot\|_2$. If M is a left preconditioner, then we solve the system $M^{-1}A\mathbf{x} = M^{-1}\mathbf{b}$. Let $\mathbf{r} = M^{-1}\mathbf{b} - M^{-1}A\mathbf{x}$.

(a) Note that $1024 = 2^{10} > 10^3$. By (35.13),

$$\begin{aligned}
\|\mathbf{r}_{20}\| &\leq \left\|M^{-1}\mathbf{b}\right\| \inf_{p_{20}\in\mathbb{P}_{20}} \left\|p_{20}\left(M^{-1}A\right)\right\| \\
&\leq \left\|M^{-1}\right\|\|\mathbf{b}\|\left\|\left(I - M^{-1}A\right)^{20}\right\| \\
&\leq \left\|M^{-1}\right\|\|\mathbf{b}\|\left\|I - M^{-1}A\right\|^{20} \\
&\leq \frac{\left\|M^{-1}\right\|\|\mathbf{b}\|}{2^{20}} \\
&< \frac{\left\|M^{-1}\right\|\|\mathbf{b}\|}{10^6}.
\end{aligned}$$

(b) We apply the CG iteration to the normal equations

$$\left(M^{-1}A\right)^* M^{-1}A\mathbf{x} = \left(M^{-1}A\right)^* M^{-1}\mathbf{b}.$$

Then we have

$$\|\mathbf{e}_n\|^2_{(M^{-1}A)^*M^{-1}A} = \mathbf{e}_n^*\left(M^{-1}A\right)^* M^{-1}A\mathbf{e}_n = \left\|M^{-1}A\mathbf{e}_n\right\|^2 = \|\mathbf{r}_n\|^2.$$

By (39.4),

$$\|\mathbf{r}_n\| \leq 2\|\mathbf{r}_0\|\left(\frac{\kappa-1}{\kappa+1}\right)^n = 2\|\mathbf{b}\|\left(\frac{\kappa-1}{\kappa+1}\right)^n.$$

Note that $\kappa = \kappa\left(M^{-1}A\right) \geq 1$. We solve

$$\begin{aligned}
\left(\frac{\kappa-1}{\kappa+1}\right)^n \leq \frac{1}{10^6} &\iff n \ln\left(\frac{\kappa-1}{\kappa+1}\right) \leq -6\ln 10, \\
\therefore n &\geq -\frac{6\ln 10}{\ln(\kappa-1) - \ln(\kappa+1)} \\
&= \frac{6\ln 10}{\ln(\kappa+1) - \ln(\kappa-1)} \\
&= \frac{6\ln 10}{2/\xi} \qquad \text{for some } \xi \in (\kappa-1, \kappa+1) \\
&> 3(\kappa-1)\ln 10.
\end{aligned}$$

■