Solutions for Applied Numerical Linear Algebra

Zheng Jinchang

January 25, 2014

This solution is for Applied Numerical Linear Algebra, written by James. W. Demmel. It covers the majority of the questions from Chapter I to Chapter VI except the programming and few other questions . As it is done by myself, surely it will have errors. Thus, if there are some errors or something else, you can email me at jczheng1234@hotmail.com. I hope this solution is helpful :)

CONTENTS

1 Solutions for Chapter I: Introduction 2

2 Solutions for Chapter II: Linear Equation Solving 15

3 Solutions for Chapter III: Linear Least Squares Problems 29

4 Solutions for Chapter IV: Nonsymmetric Eigenvalue Problems 42

5 Solutions for Chapter V: The Symmetric Eigenproblem and Singular Value Decomposition 51

6 Solutions for Chapter VI: Iterative Methods for Linear Systems 67

1 SOLUTIONS FOR CHAPTER I: INTRODUCTION

Question 1.1. *Let A be an orthogonal matrix. Show that* $\det(A) = \pm 1$*. Show that if B is also orthogonal and* $\det(A) = -\det(B)$*, then $A+B$ is singular.*

Solution. According to the definition of orthogonal matrix, we have

$$AA^{\mathrm{T}} = I.$$

Take determinant on both sides. Since $\det(A) = \det(A^{\mathrm{T}})$, it results

$$\big(\det(A)\big)^2 = 1.$$

Consequently,

$$\det(A) = \pm 1.$$

To proof $A+B$ is singular, consider $\det(A+B) = \det(A^T + B^T)$, yielding

$$\det(I + BA^T)\det(A) = \det(A+B) = \det(A^T + B^T) = \det(B^T)\det(BA^T + I).$$

Since $\det(A) = -\det(B)$, it follows $\det(BA^T + I) = 0$. Thus

$$\det(A+B) = 0.$$

i.e, $A+B$ is singular. □

Question 1.2. *The **rank** of a matrix is the dimension of the space spanned by its columns. Show that A has rank one if and only if* $A = \boldsymbol{a}\boldsymbol{b}^T$ *for some column vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$*.*

Remark. For this question to hold, we have to assume that both $\boldsymbol{a}$ and $\boldsymbol{b}$ is not $\mathbf{0}$.
Proof. If $A = \boldsymbol{a}\boldsymbol{b}^T$, let $b = \{b_1, b_2, \dots, b_n\}$. Partition A by columns as $A = \big(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_n\big)$. We have

$$\boldsymbol{\alpha}_i = b_i \boldsymbol{a}.$$

Therefore, all column vectors of A are linear dependent on $\boldsymbol{a}$. As $\boldsymbol{a}, \boldsymbol{b} \neq \mathbf{0}$, resulting in $A \neq 0$. Then, $\dim\big(\mathrm{span}\big(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_n\big)\big) = 1$, i.e. $\mathrm{rank}(A) = 1$.
On the other hand, if $\mathrm{rank}(A) = 1$, partition A by columns as $A = \big(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_n\big)$. Without losing any generalities, suppose $\boldsymbol{\alpha}_1 \neq \mathbf{0}$. Because of $\mathrm{rank}(A) = 1$, there is only one linear independent column vector of A, which means

$$\boldsymbol{\alpha}_i = b_i \boldsymbol{\alpha}_1.$$

Denote $\boldsymbol{a} = \boldsymbol{\alpha}_1, \boldsymbol{b} = \big(b_1, b_2, \dots, b_n\big)$. Then $\boldsymbol{a}, \boldsymbol{b} \neq \mathbf{0}$, and

$$A = \boldsymbol{a}\boldsymbol{b}^T.$$

□

Question 1.3. *Show that if a matrix is orthogonal and triangular, then it is diagonal. What are its diagonal elements?*

Proof. Without losing any generalities, suppose $A = (a_{ij})$ is orthogonal and upper trangular. Because A is orthogonal, we get

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} \begin{pmatrix} a_{11} & & & \\ a_{12} & a_{22} & & \\ \vdots & \ddots & \ddots & \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & & & \\ a_{12} & a_{22} & & \\ \vdots & \ddots & \ddots & \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} = 1.$$

Equate the $(1,1)$ entry, yielding

$$a_{1i} = \pm\delta_{1i}.$$

Then, condition becomes

$$\begin{pmatrix} 1 & & & \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} \begin{pmatrix} 1 & & & \\ & a_{22} & & \\ & \ddots & \ddots & \\ & a_{2n} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & a_{22} & & \\ & \ddots & \ddots & \\ & a_{2n} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} 1 & & & \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} = 1,$$

which means $A(2:n, 2:n)$ is orthogonal and upper triangular. Thus, the result follows by induction, and the diagonal entries are ±1. □

Question 1.4. *A matirx is **stictly upper triangular** if it is upper triangluar with zero diagonal elements. Show that if A is strictly upper triangular and n-by-n then $A^n = 0$.*

Proof. Suppose B is a n-by-n matrix, and partition B by its columns as $B = \left(\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_n\right)$. Consider the ith column of BA, which is

$$(BA)(:,i) = \sum_{j=1}^{i-1} A(j,n)\boldsymbol{b}_j.$$

Thus, the ith column of BA is the linear combination of the first to the $(i-1)$th columns of B. Subsititue B by A, yielding the diagonal and first superdiagonal of A^2 becoming zero, because only the first $i-1$ entries of ith column of A are nonzeros. Consequently, the second superdiagonal of A^3 becomes zero. In this ananlogy, after n times, the $n-1$ superdiagonal of A^n becomes zero, i.e. $A^n = 0$. □

Question 1.5. *Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^m$ and assume that $C \in \mathbb{R}^{m\times n}$. Show that if rank$(C) = n$, then $\|\boldsymbol{x}\|_C \equiv \|C\boldsymbol{x}\|$ is a vector norm.*

Proof.

1. Positive definiteness.
   For any $\boldsymbol{x} \in \mathbb{R}^n$, since $\|\cdot\|$ is a vector norm, we have $\|C\boldsymbol{x}\| \geq 0$, and if and only if $C\boldsymbol{x} = 0$, $\|C\boldsymbol{x}\| = 0$. Consider the linear system of equations

   $$C\boldsymbol{x} = 0.$$

   Because C has full column rank, the above equations have only zero solution. This proves the positive definiteness property.

2. Homogeneity.
   For any $\boldsymbol{x} \in \mathbb{R}^n, \alpha \in \mathbb{R}$,

   $$\|\alpha \boldsymbol{x}\|_C = \|\alpha C\boldsymbol{x}\| = |\alpha|\|C\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|_C.$$

   This proves the homogeneity property.

3. The triangle inequality.
   For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

   $$\|\boldsymbol{x}+\boldsymbol{y}\|_C = \|C\boldsymbol{x}+C\boldsymbol{y}\| \le \|C\boldsymbol{x}\| + \|C\boldsymbol{y}\| = \|\boldsymbol{x}\|_C + \|\boldsymbol{y}\|_C.$$

   This proves the triangluar inequality.

Consequently, the $\|\boldsymbol{x}\|_C \equiv \|C\boldsymbol{x}\|$ is a vector norm. □

Question 1.6. *Show that if* $0 \neq s \in \mathbb{R}^n$ *and* $E \in \mathbb{R}^{m\times n}$*, then*

$$\left\|E\left(I - \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{s}}\right)\right\|_F^2 = \|E\|_F^2 - \frac{\|E\boldsymbol{s}\|_2^2}{\boldsymbol{s}^T\boldsymbol{s}}.$$

Proof. Since

$$E\left(I - \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{s}}\right)\left(E\left(I - \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{s}}\right)\right)^T = E\left(I - \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{s}}\right)E^T = EE^T - \frac{E\boldsymbol{s}(E\boldsymbol{s})^T}{\boldsymbol{s}^T\boldsymbol{s}},$$

it follows

$$\left\|E\left(I - \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{s}}\right)\right\|_F^2 = \operatorname{tr}(EE^T) - \frac{\operatorname{tr}(E\boldsymbol{s}(E\boldsymbol{s})^T)}{\boldsymbol{s}^T\boldsymbol{s}} = \|E\|_F^2 - \frac{\|E s\|_2^2}{\boldsymbol{s}^T\boldsymbol{s}}.$$

Therefore, the question is proved. □

Remark. It could also be proved by projection property, because for any $\boldsymbol{u} \in \mathbb{R}^n$, $P = \boldsymbol{u}\boldsymbol{u}^T/\boldsymbol{u}^T\boldsymbol{u}$ is the orthogonal projection of the subspace spanned by $\boldsymbol{u}$, and $I - P$ is its complement projection.

Question 1.7. *Verify that* $\|\boldsymbol{x}\boldsymbol{y}^*\|_F = \|\boldsymbol{x}\boldsymbol{y}^*\|_2 = \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$ *for any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$.

Proof. Because the Frobenius norm can be evaluated as $\|A\|_F^2 = \operatorname{tr}(AA^*)$, it follows

$$\|\boldsymbol{x}\boldsymbol{y}^*\|_F^2 = \operatorname{tr}(\boldsymbol{x}\boldsymbol{y}^*\boldsymbol{y}\boldsymbol{x}^*) = (\boldsymbol{y}^*\boldsymbol{y})\operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^*) = \|\boldsymbol{y}\|_2^2\|\boldsymbol{x}\|_2^2.$$

For two-norm, since we have the equation about two-norm: $\|A\|_2^2 = \lambda_{\max}(AA^*)$, subsititue A by $\boldsymbol{x}\boldsymbol{y}^*$. We have

$$\|\boldsymbol{x}\boldsymbol{y}^*\|_2^2 = \lambda_{\max}(\boldsymbol{x}\boldsymbol{y}^*\boldsymbol{y}\boldsymbol{x}^*) = (\boldsymbol{y}^*\boldsymbol{y})\lambda_{\max}(\boldsymbol{x}\boldsymbol{x}^*) = \|\boldsymbol{y}\|_2^2\lambda_{\max}(\boldsymbol{x}\boldsymbol{x}^*).$$

Now, what remains is to determine $\lambda_{\max}(\boldsymbol{x}\boldsymbol{x}^*)$. Because $\boldsymbol{x}\boldsymbol{x}^*$ is hermitian, we get

$$\lambda_{\max}^2(\boldsymbol{x}\boldsymbol{x}^*) = \max_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\boldsymbol{v}^*\boldsymbol{x}\boldsymbol{x}^*\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} = \max_{\boldsymbol{v}^*\boldsymbol{v}=1} (\boldsymbol{v}^*\boldsymbol{x})^2, \text{for any } \boldsymbol{v} \in \mathbb{C}.$$

According to Cauchy-Schwartz inequality, only when $\boldsymbol{v}$ is proportional to $\boldsymbol{x}$, $(\boldsymbol{v}^*\boldsymbol{x})$ get the greatest absolute value. Therefore, $\lambda^2_{\max}(\boldsymbol{x}\boldsymbol{x}^*) \le \|x\|_2^2$, and when $\boldsymbol{v} = \boldsymbol{x}$ the inequality becomes equality. Consequently,

$$\|\boldsymbol{x}\boldsymbol{y}^H\|_2 = \|\boldsymbol{y}\|_2^2\lambda_{\max}(\boldsymbol{x}\boldsymbol{x}^*) = \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2.$$

Combining the two eqations proves the question. □

Remark. If we regard vector $\boldsymbol{x}$ and $\boldsymbol{y}$ as n-by-1 matrices, by the consistancy of two-norm, $\|\boldsymbol{x}\boldsymbol{y}\|_2 \le \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$ holds, and this upper bound is attainable by applying vector $\boldsymbol{y}$. Since we have not proved the consistancy, I use the above method, instead.

Question 1.8. *One can identify the degree* $\boldsymbol{d}$ *polynomials* $p(x) = \sum_{i=0}^d a_i x^i$ *with* $\mathbb{R}^{d+1}$ *via the vector of coefficients. Let* $\boldsymbol{x}$ *be fixed. Let* $S_{\boldsymbol{x}}$ *be the set of polynomials with an infinite relative condition number with respect to evaluationg them at* $\boldsymbol{x}$ *(i.e., they are zero at* $\boldsymbol{x}$*). In a few words, describe* $S_{\boldsymbol{x}}$ *geometrically as a subset of* $\mathbb{R}^{d+1}$*. Let* $S_x(\kappa)$ *be the set of polynomails whose relative condition number is* κ *or greater. Describe* $S_{\boldsymbol{x}}(\kappa)$ *geometrically in a few words. Describe how* $S_{\boldsymbol{x}}(\kappa)$ *changes geometrically as* $\kappa \to \infty$*.*

Solution. Suppose the coefficients of the polynomials with degree d as $\boldsymbol{a} = \left(a_1, a_2, \dots, a_{d+1}\right)$, and the fixed $\boldsymbol{x} = \left(1, x, \dots, x^d\right)$. Since $S_{\boldsymbol{x}}$ are the polynomials that are zero at $\boldsymbol{x}$, we have

$$\boldsymbol{a}^T\boldsymbol{x} = 0.$$

Therefore, the set $S_{\boldsymbol{x}}$ forms the hyperplanes which are perpendicular to the fixed $\boldsymbol{x}$ (or their normal vectors are $\boldsymbol{x}$).

As for the $S_x(\kappa)$, set $|\boldsymbol{a}| = \left(|a_1|, |a_2|, \dots, |a_{d+1}|\right)$, and $|\boldsymbol{x}| = \left(|x_1|, |x_2|, \dots, |x_{d+1}|\right)$. Due to the definition of the relative condition number, we have

$$\frac{|\boldsymbol{a}^T||\boldsymbol{x}|}{|\boldsymbol{a}^T\boldsymbol{x}|} \ge \kappa.$$

Using Cauchy-Schwartz inequality, we can transform the inequality as

$$\frac{\|\boldsymbol{a}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{a}^T\boldsymbol{x}|} \ge \kappa.$$

Take reciprocal and arccos on both sides, the inequality does not change, resulting

$$\angle(\boldsymbol{a}, \boldsymbol{x}) \ge \arccos\frac{1}{\kappa}.$$

So, the set $S_x(\kappa)$ are hyperplanes whose normal vectors have greater angle than $\arccos\frac{1}{\kappa}$ with the fixed $\boldsymbol{x}$. Thus, as $\kappa \to \infty$, $\arccos\frac{1}{\kappa} \to \frac{\pi}{2}$, i.e, the hyperplanes become perpendicualr to the fixed $\boldsymbol{x}$. □

Question 1.9. *This question is too long, so I omit it. For details, please refer to the textbook.*

Proof. For the first algorithm, insert a roundoff term $(1+\delta_i)$ at each floating point operation, we get

$$\hat{y} = \frac{\log(1+x)(1+\delta_1)}{x}(1+\delta_2) = \frac{\log(1+x)+\log(1+\delta_1)}{x}(1+\delta_2).$$

Since $x, \delta_i \approx 0$, we get the relative error as

$$\left|\frac{\hat{y}-y}{y}\right| = \left|\frac{x}{\log(1+x)}\frac{\log(1+\delta_1)}{x}(1+\delta_2)\right| \approx \left|\frac{\delta_1}{x}(1+\delta_2)\right| \approx \left|\frac{\delta_1}{x}\right|$$

As δ is a bounded value, the relative error will grow when $x \to 0$.
Use the same technique to analyse the second algorithm, we have its relative error as

$$\left|\frac{\hat{y}-y}{y}\right| = \left|\frac{d-1}{\log d}\left(\frac{\log d}{(d-1)(1+\delta_1)}(1+\delta_2) - \frac{\log d}{d-1}\right)\right| = \left|\frac{1+\delta_2}{1+\delta_1}-1\right| \approx \delta,\ |\delta| \leq 2\epsilon.$$

Therefore, the second algorithm will correct near $x = 0$. □

Question 1.10. *Show that, barring overflow or underflow, fl$\big(\sum_{i=1}^d x_i y_i\big) = \sum_{i=1}^d x_i y_i(1+\delta_i)$, where $|\delta_i| \leq d\epsilon$. Use this to prove the following fact. Let $A^{m\times n}$ and $B^{n\times p}$ be matrices, and compute their produnct in the usual way. Baring overflow or underflow show that $|fl(A\cdot B) - A\cdot B| \leq n\cdot\epsilon\cdot|A|\cdot|B|$. Here the absolute value of a matrix $|A|$ means the matrix with entries $(|A|)_{ij} = |a_{ij}|$, and the inequality is meant componentwise.*

Proof. Denote δ_i as mutipilication roundoff error and $\hat{\delta}_i$ as addition roundoff error. Expanding the formula with the roundoff terms, we get

$$\mathrm{fl}\left(\sum_{i=1}^d x_i y_i\right) = \sum_{i=2}^d \left((1+\delta_i)\prod_{j=i-1}^{d-1}(1+\hat{\delta}_j)x_i y_j\right) + \prod_{j=1}^{d-1}(1+\hat{\delta}_j)x_1 y_1.$$

As

$$1-(d-i+1)\epsilon + O(\epsilon^2) \leq (1-\epsilon)^{d-i+1} \leq \prod_{j=i-1}^{d-1}(1+\hat{\delta}_j) \leq (1+\epsilon)^{d-i+1} \leq 1+(d-i+1)\epsilon + O(\epsilon^2),$$

we have the following two inequlities

$$\mathrm{fl}\left(\sum_{i=1}^d x_i y_i\right) \leq \sum_{i=2}^d (1+(d-i+2)\epsilon)x_i y_i + (1+d\epsilon)x_1 y_1 \leq \sum_{i=1}^d (1+d\epsilon)x_i y_i,$$

$$\sum_{i=1}^d (1-d\epsilon)x_i y_i \leq \sum_{i=2}^d (1-(d-i+2)\epsilon)x_i y_i + (1-d\epsilon)x_1 y_1 \leq \mathrm{fl}\left(\sum_{i=1}^d x_i y_i\right).$$

Therefore, $\mathrm{fl}\left(\sum_{i=1}^d x_i y_i\right) = \sum_{i=1}^d x_i y_i(1+\bar{\delta}_i)$, where $|\bar{\delta}_i| \leq d\epsilon$.
Now, turn to prove the matrix inequlity. Since the absolute value of a matrix is defined componentwise, we just need to prove it componentwise. Suppose $c_{ij} = (A\cdot B)(i,j) = \sum_{k=1}^n A(i,k)B(k,j)$. Applying what we have proved, we have

$$\left|\mathrm{fl}(c_{ij}) - c_{ij}\right| = \left|\sum_{k=1}^n A(i,k)B(k,j)\delta_k\right| \leq n\epsilon \sum_{k=1}^n |A(i,k)||B(k,j)| = n\epsilon(|A||B|)(i,j).$$

This proves the question. □

Question 1.11. *Let L be a lower triangular matrix and solve $L\boldsymbol{x} = b$ by forward substituion. Show that barring overflow or underflow, the computed solution $\hat{\boldsymbol{x}}$ satisfies $(L+\delta L)\hat{\boldsymbol{x}} = b$, where $|\delta l_{ij}| \le n\epsilon|l_{ij}|$, where ϵ is the machine precision. This means that forward substituion is backward stable. Argue that backward substitution for solving upper triangular systems satisfies the same bound.*

Proof. Insert a roundoff term $(1+\delta_i)$ at each floating point operation in the forward substituion formula, and use the results of last question. We get

$$x_i = \frac{b_i - \sum_{k=1}^{i-1} a_{ik}x_k(1+\delta_k)}{a_{ii}}(1+\hat{\delta}_i)(1+\bar{\delta}_i), \text{ where } |\delta_k| \le (i-1)\epsilon, |\hat{\delta}_i| \le \epsilon, |\bar{\delta}_i| \le \epsilon.$$

Solve b_i from the above equation, because of

$$1/(1+\hat{\delta}_i)(1+\bar{\delta}_i) = 1+\delta_i, \text{ where } |\delta_i| \le 2\epsilon.$$

We can combine the term x_i in the sum as follows:

$$b_i = \sum_{k=1}^{i} a_{ik}x_k(1+\delta_{ik}), \text{ where } |\delta_{ik}| \le \max\{2, i-1\}\epsilon.$$

Therefore, denote $(\delta L)_{ij} = a_{ij}\delta_{ij}$. We have proved $(L+\delta L)\hat{\boldsymbol{x}} = b$, where $|\delta l_{ij}| \le n\epsilon|l_{ij}|$. □

Remark. The result will be slightly different if we consider $b_i - \sum_{k=1}^{i-1} a_{ik}x_k$ as

$$(((b_i - a_{i1}x_1) - a_{i2}x_2) - a_{i3}x_3)\ldots.$$

In fact, it may be the evaluation sequence the question intended. My result is based on the following sequence

$$b_i - (\sum_{k=1}^{i-1} a_{ik}x_k).$$

For this question, there is not major difference between the two.

Question 1.12. *The question is too long, so I omit it. For details, please refer to the textbook.*

Remark. In my point of view, the question is intended to prove that complex arithmetics implemented in real arithmetics are backward stable. Thus, it may be sufficient to prove

$$|\mathrm{fl}(a \odot b) - ab| \le O(\epsilon)ab.$$

My proof tries to find out the expressions of the δ's.

Solution. First, consider the complex addition. Suppose there is such $\delta = \delta_1 + \delta_2\boldsymbol{i}$. Then, as complex addition is two real addition, we have two ways to compute the floating point operations as follows:

$$\mathrm{fl}\,(a_1 + b_1\boldsymbol{i} + a_2 + b_2\boldsymbol{i}) = (a_1+a_2)(1+\hat{\delta}) + (b_1+b_2)(1+\bar{\delta})\boldsymbol{i} = (a_1+a_2+b_1\boldsymbol{i}+b_2\boldsymbol{i})(1+\delta_1+\delta_2\boldsymbol{i}).$$

Solving δ_1, δ_2 from the above equations, we get

$$\delta_1 = \frac{(a_1+a_2)^2\hat{\delta} + (b_1+b_2)^2\bar{\delta}}{(a_1+a_2)^2 + (b_1+b_2)^2},$$
$$\delta_2 = \frac{(a_1+a_2)(b_1+b_2)(\bar{\delta} - \hat{\delta})}{(a_1+a_2)^2 + (b_1+b_2)^2}.$$

Since $|\delta_1|^2 + |\delta_2|^2 \le |\hat{\delta} + \bar{\delta}|^2 + |\bar{\delta} - \hat{\delta}|^2 \le 4\epsilon^2$, we have proved for the complex addition. As subtraction is essentially the same as addition, it is the same for complex subtraction .
As for mutiplication, use the same technique and the above result, we have the equation as

$$\begin{aligned} \mathrm{fl}\,((a_1 + b_1\boldsymbol{i})(a_2 + b_2\boldsymbol{i})) &= (a_1 + b_1\boldsymbol{i})(a_2 + b_2\boldsymbol{i})(1 + \delta_1 + \delta_2\boldsymbol{i}) \\ &= \left(a_1a_2(1+\bar{\delta}_1) - b_1b_2(1+\bar{\delta}_4) + a_2b_1(1+\bar{\delta}_2)\boldsymbol{i} + a_1b_2(1+\bar{\delta}_3)\boldsymbol{i}\right)(1 + \hat{\delta}_1 + \hat{\delta}_2\boldsymbol{i}). \end{aligned}$$

For simplicity, we ignore the higher order epsilons as they are not the major part of the error. Then solving δ_1, δ_2 from the above equations, we get

$$\begin{aligned} \delta_1 =& \frac{1}{(a_1a_2 - b_1b_2)^2 + (a_1b_2 + a_2b_1)^2}\Big((a_1^2a_2^2 - a_1a_2b_1b_2)\bar{\delta}_1 + (a_2^2b_1^2 + a_1a_2b_1b_2)\bar{\delta}_2 + (a_1^2b_2^2 + \\ & a_1a_2b_1b_2)\bar{\delta}_3 - (a_1a_2b_1b_2 + b_1^2b_2^2)\hat{\delta}_4 + (a_1^2a_2^2 - b_1^2b_2^2 + a_2^2b_1^2 + a_1^2b_2^2)\hat{\delta}_1 - (a_1b_1b_2^2 + a_1a_2^2b_1)\hat{\delta}_2\Big), \\ \delta_2 =& \frac{1}{(a_1a_2 - b_1b_2)^2 + (a_1b_2 + a_2b_1)^2}\Big(-(a_1^2a_2b_2 + a_1a_2^2b_1)\bar{\delta}_1 + (a_1a_2^2b_1 - a_2b_1^2b_2)\bar{\delta}_2 + (a_1^2a_2b_2 - \\ & a_1b_1b_2^2)\bar{\delta}_3 + (a_1b_1b_2^2 + a_2b_1^2b_2)\bar{\delta}_4 + ((a_1a_2 - b_1b_2)^2 + (a_1b_2 + a_2b_1)^2)\hat{\delta}_2\Big). \end{aligned}$$

We can claim that all the coefficient of each δ_i is bounded, because of the mean value inequality. Therefore, the question holds for the complex multiplication. Now, we can check whether both the real and imaginary parts of the complex product are always compute accurately. Since we have $\mathrm{fl}(a_1 + b_1\boldsymbol{i})(a_2 + b_2\boldsymbol{i}) = (a_1 + b_1\boldsymbol{i})(a_2 + b_2\boldsymbol{i})(1 + \delta_1 + \delta_2\boldsymbol{i})$, separate the real and imaginary parts as

$$\mathrm{Error_r} = \left|\frac{(a_1a_2 - b_1b_2)\delta_1 - (a_1b_2 + a_2b_1)\delta_2}{a_1a_2 - b_1b_2}\right|,$$
$$\mathrm{Error_i} = \left|\frac{(a_1a_2 - b_1b_2)\delta_2 - (a_1b_2 + a_2b_1)\delta_1}{a_1b_2 + a_2b_1}\right|.$$

Thus, if $a_1, b_1 \to 0$, the real part may go wrong; and if $b_1, b_2 \to 0$, the imaginary part may go wrong. So both parts of the complex product may not always be computed accurately.
At last, we check the complex division with the knowledge of complex mutiplication, and my algorithm for complex division is like the normal complex division just without actual complex mutiplication in realizing the denominator. Therefore, we have

$$\mathrm{fl}\left(\frac{a_1 + b_1\boldsymbol{i}}{a_2 + b_2\boldsymbol{i}}\right) = \frac{a_1 + b_1\boldsymbol{i}}{a_2 + b_2\boldsymbol{i}}(1 + \delta_1 + \delta_2\boldsymbol{i}) = \frac{(a_1 + b_1\boldsymbol{i})(a_2 - b_2\boldsymbol{i})(1 + \bar{\delta}_1 + \bar{\delta}_2\boldsymbol{i})}{\left(a_2^2(1+\hat{\delta}_1) + b_2^2(1+\hat{\delta}_2)\right)(1+\hat{\delta}_3)}(1 + \hat{\delta}_4).[1]$$

[1] In reality, there are two division δ's. For simplicity I suppose they are the same, becasue there are no difference after the following merging δ's. Otherwise, the procedure remains the same, but will be slightly more complicated.

Ignore the higher order epsilon. Then merge $(1+\bar{\delta}_1+\bar{\delta}_2\boldsymbol{i})(1+\hat{\delta}_4), (1+\hat{\delta}_1)(1+\hat{\delta}_3)$ and $(1+\hat{\delta}_2)(1+\hat{\delta}_3)$, and continue to use the old mark. We can solve δ_1, δ_2 as

$$\begin{aligned}\delta_1 = \bar{\delta}_1 - &\frac{a_2^2\hat{\delta}_1 + b_2^2\hat{\delta}_2}{a_2^2+b_2^2},\\ \delta_2 = &\bar{\delta}_2.\end{aligned}$$

As all the coefficient of each δ is smaller than one, we prove the result for complex division. What's more, not matter whether $|a|$ is large or small, a/a will approximate 1 by the above formulas. □

Question 1.13. *Prove lemma 1.3.*

Remark. Since there is not major difference between real and complex situations. For simplicity, I prove for the real case.
Proof. In one side, if A is s.p.d., for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we define an inner product as $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T A\boldsymbol{y}$. Verify four criteria of inner product, we get

1. $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T A\boldsymbol{y} = (\boldsymbol{y}^T A\boldsymbol{x})^T = \boldsymbol{y}^T A\boldsymbol{x} = \langle \boldsymbol{y}, \boldsymbol{x}\rangle$.
2. $\langle \boldsymbol{x}, \boldsymbol{y}+\boldsymbol{z}\rangle = \boldsymbol{x}^T A(\boldsymbol{y}+\boldsymbol{z}) = \boldsymbol{x}^T A\boldsymbol{y} + \boldsymbol{x}^T A\boldsymbol{z} = \langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{x}, \boldsymbol{z}\rangle$.
3. $\langle \alpha\boldsymbol{x}, \boldsymbol{y}\rangle = \alpha\boldsymbol{x}^T A\boldsymbol{y} = \alpha\langle \boldsymbol{x}, \boldsymbol{y}\rangle$.
4. $\langle \boldsymbol{x}, \boldsymbol{x}\rangle = \boldsymbol{x}^T A\boldsymbol{x} \geq 0$, and $\langle \boldsymbol{x}, \boldsymbol{x}\rangle = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.

Therefore, it is an inner product.
On the other side, if $\langle\cdot,\cdot\rangle$ is an inner product, consider its value on the canonical $\mathbb{R}^n$ bases $\boldsymbol{e}_i$. Set matrix $A(i,j) = \langle \boldsymbol{e}_i, \boldsymbol{e}_j\rangle$, then $A(i,j)$ is symmetric because of the symmetry property of inner product. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, expand them under the canonical bases as $\boldsymbol{x} = \sum_{i=1}^n \alpha_i\boldsymbol{e}_i, \boldsymbol{y} = \sum_{i=1}^n \beta_i\boldsymbol{e}_i$. We get

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = (\sum_{i=1}^n \alpha_i\boldsymbol{e}_i)^T(\sum_{i=1}^n \beta_i\boldsymbol{e}_i) = \boldsymbol{x}^T A\boldsymbol{y}.$$

This means A is symmetric positive definite, and is uniquely determined by the inner product (for the bases is fixed). Thus, the quetion is proved. □

Question 1.14. *Prove Lemma 1.5.*

Proof. First, since $\sqrt{a^2+b^2} \leq |a|+|b|$, we have

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \leq |x_1| + \sqrt{\sum_{i=2}^n x_i^2} \leq |x_1| + |x_2| + \sqrt{\sum_{i=3}^n x_i^2} \leq \ldots \leq \sum_{i=1}^n |x_i| = \|\boldsymbol{x}\|_1.$$

on the other hand,

$$\|\boldsymbol{x}\|_1^2 = (\sum_{i=1}^n |x_i|)^2 = \sum_{i=1}^n x_i^2 + 2\sum_{\substack{i=1\\j=i+1}}^n |x_i x_j| \leq \sum_{i=1}^n x_i^2 + \sum_{i\neq j}(x_i^2+x_j^2) = n\sum_{i=1}^n x_i^2 = n\|\boldsymbol{x}\|_2^2.$$

Therefore, we get the first inequality $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{n}\|\boldsymbol{x}\|_2$.
For the inequality about infinity-norm, since it is obvious that

$$\max(|x_i|) \leq \sqrt{\max(x_i{}^2) + \sum_{\text{else}} x_i^2} \leq \max(|x_i|) + \sum_{\text{else}} |x_i|,$$

and

$$n\max(|x_i|)^2 \geq \sum_{i=1}^{n} x_i^2, \quad n\max(|x_i|) \geq \sum_{i=1}^{n} |x_i|,$$

the second inequality $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty$ and third inequality $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_1 \leq n\|\boldsymbol{x}\|_\infty$ follows. □

Question 1.15. *Prove Lemma 1.6.*

Proof. Let $A \in \mathbb{R}^{m\times n}$, $\|\cdot\|_{mn}$ the operator norm and $\|\cdot\|_m, \|\cdot\|_n$ the corresponding vector norms. To prove the question, we just need to verify the three criteria.

1. Positive definiteness
   Since $\|A\|_{mn} = \max_{\|\boldsymbol{x}\|_n=1}(\|A\boldsymbol{x}\|_m)$, definitly $\|A\|_{mn} \geq 0$. If $\|A\|_{mn} = 0$, which means $A\boldsymbol{x} = \mathbf{0}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. i.e., $\mathbb{R}^n$ is the solution set of linear system of equations $A\boldsymbol{x} = 0$. As a result, we have $A = 0$.

2. Homogeneity
   Due to the homogeneity of corresponding vector norm, we have

   $$\|\alpha A\|_{mn} = \max_{\|\boldsymbol{x}\|_n=1}(\|\alpha A\boldsymbol{x}\|_m) = |\alpha| \max_{\|\boldsymbol{x}\|_n=1}(\|A\boldsymbol{x}\|_m) = |\alpha|\|A\|_{mn}.$$

3. The triangluar inequality
   Because the triangular inequality of corresponding vector norm, we have

   $$\|A+B\|_{mn} = \max_{\|\boldsymbol{x}\|_n=1}(\|(A+B)\boldsymbol{x}\|_m) \leq \max_{\|\boldsymbol{x}\|_n=1}(\|A\boldsymbol{x}\|_m) + \max_{\|\boldsymbol{x}\|_n=1}(\|B\boldsymbol{x}\|_m) = \|A\|_{mn} + \|B\|_{mn}.$$

In all, an operator norm is a matrix norm. □

Question 1.16. *Prove all parts except 7 of lemma 1.7.*

Remark. Part 14 of lemma 1.7 seems wrong. Thus, I do some modification there.
Proof.

1. $\|A\boldsymbol{x}\| \leq \|A\| \cdot \|\boldsymbol{x}\|$ *for a vector norm and its corresponding operator norm, or the vector two-norm and matrix Frobenius norm.*
   For a vector norm and its corresponding operator norm, the inequality is obvious, because operator norm is defined as smallest upbound for which $\|A\| \cdot \|\boldsymbol{x}\| \geq \|A\boldsymbol{x}\|$ holds. We just need to prove it for the two-norm and Frobenius norm.
   Partition matrix A by its rows as $A = \left(\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \ldots, \boldsymbol{\alpha}_m^T\right)^T$. We have

   $$\|A\boldsymbol{x}\|_2^2 = \sum_{i=1}^{m}(\boldsymbol{\alpha}_i^T\boldsymbol{x})^2 \leq \sum_{i=1}^{m}\|\boldsymbol{\alpha}_i\|_2^2\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}\|_2^2\sum_{i=1}^{m}\|\boldsymbol{\alpha}_i\|_2^2 = \|\boldsymbol{x}\|_2^2\|A\|_{\mathrm{F}}^2.$$

   Thus, it also holds for the two-norm and Frobenius norm.

2. $\|AB\| \le \|A\| \cdot \|B\|$ *for any operator norm or for the Frobenius norm.*
   Using the above result, for any operator norm, we have

$$\|AB\| = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|A(B\boldsymbol{x})\|}{\|\boldsymbol{x}\|} \le \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|A\| \cdot \|B\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \|A\| \cdot \|B\|.$$

   For Frobenius norm, parition A by its rows as $A = \left(\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \ldots, \boldsymbol{\alpha}_m^T\right)^T$, and split B by its columns as $B = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_s\right)$. We get

$$\|AB\|_F^2 = \sum_{\substack{i=1\\j=1}}^{m,s} (\boldsymbol{\alpha}_i^T \boldsymbol{\beta}_j)^2 \le \sum_{\substack{i=1\\j=1}}^{m,s} (\|\boldsymbol{\alpha}_i\|_2^2 \cdot \|\boldsymbol{\beta}_j\|_2^2) = \sum_{i=1}^{m} \|\boldsymbol{\alpha}_i\|_2^2 \cdot \sum_{j=1}^{s} \|\boldsymbol{\beta}_i\|_2^2 = \|A\|_F^2 \|B\|_F^2.$$

   Thus, it holds also for the two-norm and Frobenius norm.

3. *The max norm and Frobenius norm are not operator norms.*
   Consider a matrix $A = \boldsymbol{\alpha}\boldsymbol{\alpha}^T$. We have a lower bound of $\|A\|$ as follows,

$$\|A\| = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \ge \frac{\|A\boldsymbol{\alpha}\|}{\|\boldsymbol{\alpha}\|} = \frac{\|\boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\alpha}\|}{\|\boldsymbol{\alpha}\|} = \|\boldsymbol{\alpha}\|_2^2 \frac{\|\boldsymbol{\alpha}\|}{\|\boldsymbol{\alpha}\|} = \|\boldsymbol{\alpha}\|_2^2.$$

   Let $\boldsymbol{\alpha} = \left(1, 1, \ldots, 1\right)$, then $\|A\|_{\max} = 1$. However, according to the above inequality, if $\|A\|_{\max}$ is an operator norm, we have $\|A\|_{\max} \ge n$. Such conflict means that $\|\cdot\|_{\max}$ is not an operator norm.
   For Frobenius norm, consider any operator norm of identity matrix, we have

$$\|I\| = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|I\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = 1.$$

   While the Frobenius norm of I is n, it means that Frobenius norm is not an operator norm.

4. $\|QAZ\| = \|A\|$ *if Q and Z are orthogonal or unitary for the Frobenius norm and for the operator norm induced by* $\|\cdot\|_2$.[2]
   Separate the target equation as two equations: $\|QA\| = \|A\|, \|AZ\| = \|A\|$. Since $\|A^T\| = \|A\|$ holds for Frobenius and two-norm (it is obvious for Frobenius norm and later in this question we will prove this for two-norm), we just need to prove the first equation. For two-norm, since $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$, we get

$$\|QA\|_2 = \sqrt{\lambda_{\max}(A^TQ^TQA)} = \sqrt{\lambda_{\max}(A^TA)} = \|A\|_2.$$

   Next, for the Frobenius norm, we claim that for any vector $\boldsymbol{x}$, $\|Q\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$, because $\|Q\boldsymbol{x}\|_2^2 = \boldsymbol{x}^TQ^TQ\boldsymbol{x}$. Then, parition A by its columns as $A = \left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n\right)$. We have

$$\|QA\|_F^2 = \sum_{i=1}^{n} \|Q\boldsymbol{\alpha}_i\|_2^2 = \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_2^2 = \|A\|_F^2.$$

[2] This result can be extended into the case of the rectangular matrix. The proof is nearly the same, except that we use the equality $\|A\|_2 = \sigma_{\max}(A)$.

5. $\|A\|_\infty \equiv \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_i \sum_j |a_{ij}|$ = *maximum absolute row sum.*
Partition A by its rows as $A = \left(\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \ldots, \boldsymbol{\alpha}_n^T\right)^T$. We have

$$\|A\|_\infty \equiv \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{\|\boldsymbol{x}\|_\infty=1,i} |\boldsymbol{\alpha}_i^T \boldsymbol{x}| \leq \max_i \sum_{j=1}^{n} \left|a_{ij}\right|.$$

Denote i the number of row of the maximum absolute row sum, and set $\boldsymbol{y} = \big(\text{sign}(a_{i1}),$ $\text{sign}(a_{i2}), \ldots, \text{sign}(a_{in})\big)$. The up bound is attainable, as $\|\boldsymbol{y}\|_\infty = 1$ and $\|A\boldsymbol{y}\| = \max_i \sum_j |a_{ij}|$.

6. $\|A\|_1 \equiv \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|}_1 = \|A^T\|_\infty = \max_j \sum_i |a_{ij}|$ = *maximum absolute column sum.*
Partition A by its columns as $A = \left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n\right)$. We have

$$\|A\|_1 = \max_{\|\boldsymbol{x}\|_1=1} \|A\boldsymbol{x}\|_1 = \max_{\|\boldsymbol{x}\|_1=1} \left\| \sum_{i=1}^{n} \boldsymbol{\alpha}_1 x_1 \right\|_1 \leq \max_{\|\boldsymbol{x}\|_1=1} \sum_{i=1}^{n} |x_i| \, \|\boldsymbol{\alpha}_i\|_1 .$$

Denote j the number of column of the maximum absolute column sum. We have

$$\begin{aligned}\max_{\|\boldsymbol{x}\|_1=1} \sum_{i=1}^{n} |x_i| \, \|\boldsymbol{\alpha}_i\|_1 &= \max_{\|\boldsymbol{x}\|_1=1} \left(\sum_{i\neq j} |x_i| \, \|\boldsymbol{\alpha}_i\|_1 + \left(1 - \sum_{i\neq j} |x_i|\right) \|\boldsymbol{\alpha}_j\|_1 \right) \\ &= \max_{\|\boldsymbol{x}\|_1=1} \left(\|\boldsymbol{\alpha}_j\|_1 - \sum_{i\neq j} \left(\|\boldsymbol{\alpha}_j\|_1 - \|\boldsymbol{\alpha}_i\|_1\right) |x_i| \right) \\ &\leq \|\boldsymbol{\alpha}_j\|_1 .\end{aligned}$$

Set $\boldsymbol{x} = \boldsymbol{e}_j$, and the up bound is attainable.

7. $\|A\|_2 = \|A^T\|_2$.[3]
Since $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$, it suffice to prove that $\lambda_{\max}(A^T A) = \lambda_{\max}(AA^T)$. Denote λ is as a nonzero eigenvalue of $A^T A$ and $\boldsymbol{x}$ its corresponding eigenvector. We have

$$(AA^T)A\boldsymbol{x} = A(A^T A)\boldsymbol{x} = \lambda A\boldsymbol{x}.$$

Since $A^T A\boldsymbol{x} = \lambda\boldsymbol{x}$, and $\lambda \neq 0$, we have $A\boldsymbol{x} \neq \boldsymbol{0}$. Therefore, λ is an eigenvalue of AA^T and $A\boldsymbol{x}$ is its corresponding eigenvector. As both $A^T A$ and AA^T are nonnegative symmetric matrices, their eigenvalues must at least as large as zero. If one of the matrices has all its eigenvalues zero, it follows $A = 0$. As a result the other will has all its eigenvalues zero, too. Thus, $\lambda_{\max}(A^T A) = \lambda_{\max}(AA^T)$.

8. $\|A\|_2 = \max_i |\lambda_i(A)|$ *if* A *is normal, i.e.,* $AA^T = A^T A$.
Since A is normal, there exist an orthgonal matrix P and a diagonal matrix $\Lambda = \text{diag}\big(\lambda_1, \lambda_2, \ldots, \lambda_n\big)$ such that $P^{-1}AP = \Lambda$. Thus,

$$\|A\|_2 = \|P^{-1}AP\|_2 = \|\Lambda\|_2 = \sqrt{\lambda_{\max}(\Lambda^T\Lambda)} = \max_i |\lambda_i(\Lambda)| = \max_i |\lambda_i(A)|.$$

[3] There are several ways to prove that AB and BA have the same eigenvalues. Among them, I present an analytic version when A and B are square here. First assume A is invertable, then $\det(\lambda I - AB) = \det(\lambda I - BA)$ is easy to prove. For singular A, we perturb A as $\tilde{A} = A + tI$. $\tilde{A}$ is invertable when t is small enough, and operator det() is continuous. Thus, we have the result by taking limit. This method is inspired by my friend, Zhou Zeren. There is another method in Chapter IV.

9. *If A is n-by-n, then* $n^{-1/2}\|A\|_2 \le \|A\|_1 \le n^{1/2}\|A\|_2$.
   Use the inequalities $\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{n}\|\boldsymbol{x}\|_2$. Suppose $\|\boldsymbol{x}\| \neq 0$. Invert the inequalities as

$$\frac{1}{\sqrt{n}\|\boldsymbol{x}\|_2} \le \frac{1}{\|\boldsymbol{x}\|_1} \le \frac{1}{\|\boldsymbol{x}\|_2}.$$

   As $\|A\|_p \equiv \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}$, combine the two inequalities. We have

$$\frac{\|A\boldsymbol{x}\|_2}{\sqrt{n}\|\boldsymbol{x}\|_2} \le \frac{\|A\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} \le \frac{\sqrt{n}\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

   Take max, it follows

$$n^{-1/2}\|A\|_2 \le \|A\|_1 \le n^{1/2}\|A\|_2.$$

10. *If A is n-by-n, then* $n^{-1/2}\|A\|_2 \le \|A\|_\infty \le n^{1/2}\|A\|_2$.
    Like the previous question, we use the inequalities $n^{-1/2}\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2$. Invert the inequalities as

$$\frac{1}{\|\boldsymbol{x}\|_2} \le \frac{1}{\|\boldsymbol{x}\|_\infty} \le \frac{n^{1/2}}{\|\boldsymbol{x}\|_2}.$$

    Combine the two inequalities. We have

$$\frac{\|A\boldsymbol{x}\|_2}{\sqrt{n}\|\boldsymbol{x}\|_2} \le \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \le \frac{\sqrt{n}\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

    Take max, we get

$$n^{-1/2}\|A\|_2 \le \|A\|_\infty \le n^{1/2}\|A\|_2.$$

11. *If A is n-by-n, then* $n^{-1}\|A\|_\infty \le \|A\|_1 \le n\|A\|_\infty$.
    Like the previous question, we use the inequalities $\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n\|\boldsymbol{x}\|_\infty$. Still, invert the inequalities as

$$\frac{1}{n\|\boldsymbol{x}\|_\infty} \le \frac{1}{\|\boldsymbol{x}\|_1} \le \frac{1}{\|\boldsymbol{x}\|_\infty}.$$

    Combine the two inequalities. We have

$$\frac{\|A\boldsymbol{x}\|_2}{n\|\boldsymbol{x}\|_\infty} \le \frac{\|A\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} \le \frac{n\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

    Take max, we get

$$n^{-1}\|A\|_\infty \le \|A\|_1 \le n\|A\|_\infty.$$

12. *If A is n-by-n, then* $\|A\|_2 \le \|A\|_F \le n^{1/2}\|A\|_2$.
    As A^TA is symmetric nonnegative. Therefore, the eigenvalues of A^TA are all nonnegative. As $\sum\lambda(A^TA) = \mathrm{tr}(A^TA)$, we get

$$\lambda_{\max}(A^TA) \le \mathrm{tr}(A^TA) \le n\lambda_{\max}(A^TA).$$

    i.e.,

$$\|A\|_2 \le \|A\|_F \le n^{1/2}\|A\|_2$$

□

Question 1.17. *We mentioned that on a Cray machine the expression* $\arccos(x/\sqrt{x^2+y^2})$ *caused an error, because roundoff caused* $x/\sqrt{x^2+y^2}$ *to exceed 1. Show that this is impossible using IEEE arithmetic, barring overflow or under flow. Extra credit: Prove the same result using correctly rounded **decimal** arithmetic.*

Remark. I am not sure about the correctness of my proof, since there is not major difference between **decimal** and **binary** arithmetic in my proof. It may due to that I use the different method to prove it instead of what the author is intended for. However, whether or not it is correct, my proof can serve as a reference.
Proof. If we have prove that $\text{fl}(\sqrt{x^2}) = x$ exactly on a machine using IEEE arithmetic, we will get

$$\text{fl}(x/\sqrt{x^2+y^2}) \geq \text{fl}(x/\sqrt{x^2}) = \text{fl}(x/x) = 1,$$

which means it is impossible for $\arccos(x/\sqrt{x^2+y^2})$ causing an error. Thus, we only need to prove that $\text{fl}(\sqrt{x^2}) = x$. Denote $\text{fl}(x^2) = x^2(1+\delta)$ with $|\delta| \leq \epsilon$, where x is already a floating point number. Then, to compute $\sqrt{\text{fl}(x^2)}$, the computer will compute $\sqrt{\text{fl}(x^2)} \approx x(1+\frac{1}{2}\delta)$, and round it. However, as $\left(\sqrt{\text{fl}(x^2)} - x\right)/x = \frac{1}{2}\delta$ and $\frac{1}{2}|\delta| \leq \frac{1}{2}\epsilon$, the $\frac{1}{2}\delta$ will be rounded off, resulting in $\text{fl}(\sqrt{x^2}) = x$. □

Question 1.18. *Suppose that a and b are normalized IEEE double precision floating point numbers, and consider the following algorithm, running with IEEE arithmetic:*

```
if(|a| < |b|), swap a and b
$s_1 = a + b$
$s_2 = (a - s_1) + b$
```

Prove the following facts:

1. *Barring overflow or underflow, the only roundoff error committed in running the algorithm is computing* $s_1 = fl(a+b)$. *In other words, both subtractions* $a - s_1$ *and* $(a-s_1)+b$ *are computed **exactly**.*

2. $s_1 + s_2 = a + b$, ***exactly**. This means that* s_2 *is actually the roundoff error committed when rounding the exact value of* $a+b$ *to get* s_1.

Proof. To do addition and subtraction, a computer will equalize the exponential parts of two operands. For instance, suppose

$$a = (1, a_1, a_2, a_3, a_4, \ldots, a_{50}, a_{51})_2 \times 2^{e_1},$$
$$b = (1, b_1, b_2, b_3, b_4, \ldots, b_{50}, b_{51})_2 \times 2^{e_2}.$$

To do addition, a computer will change e_1 and e_2 into e_3, then add their corresponding fraction parts.
Therefore, if a is too larger than b, that is to say, the bottomest bit of a is still greater than the

toppest bit of b, we will have $s_1 = a + b = a$, and $s_1 - a = a - a = 0$ is exact, since a and b are already floating point numbers. Then, $s_1 + b = 0 + b = b$ is also exact.
If a is not too larger than b, without losing any generality, we can suppose that the toppest bit of b is the 50th toppest bit of a. Remain to use the above marks. We have

$$s_1 = a + b = (1, a_1, a_2, a_3, a_4, \dots, a_{50} + 1, a_{51} + b_1) \times 2^{e_1}.$$

Then, compute $a - s_1$, we have

$$a - s_1 = (1-1, a_1 - a_1, \dots, a_{50} - a_{50} - 1, a_{51} - a_{51} - b_1) \times 2^{e_1} = (-1, -b_1) \times 2^{e_1 - 50} = (-1, -b_1) \times 2^{e_2}.$$

This is exact, because the toppest bit of s_1 is the same as the toppest bit of a. What's more, since $a - s_1$ has the same toppest bit of b, it is certain that $(a - s_1) + b$ is exact.
The situation of $a < b$ is the same. In all, we can see that under both situation $s_1 + s_2$ contain all the bits of a and b. Therefore, $s_1 + s_2 = a + b$ exactly. □

2 Solutions for Chapter II: Linear Equation Solving

Question 2.1. *Programming question. Skipped.*

Question 2.2. *Consider solving $AX = B$ for X, where A is n-by-n, and X and B are n-by-m. There are two obvious algorithms. The first algorithm factorizes $A = PLU$ using Gaussian elimination and then solves for each column of X by forward and back substitution. The second algorithm computes A^{-1} using Gaussian elimination and then multiplies $X = A^{-1}B$. Count the number of flops required by each algorithm, and show that the first one requires fewer flops.*

Solution. For the first algorithm, since permutation does not increase flops, we can suppose matrix A is already permuted. Then, denote

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

To determine the first column of L requires $n-1$ divisions. To update the rest of the matrix costs $n-1$ multiplications and subtractions for each row, which totally costs $2(n-1)^2$ flops. Therefore, the whole flops Gaussian elimination requires is

$$\sum_{i=n}^{2} (2(i-1)^2 + (i-1)) = \frac{2}{3}n^3 + O(n^2).$$

For a single forward substitution, it requires

$$\sum_{i=1}^{n} (2i-1) = n^2 + O(n).$$

It is the same for back substitution. Thus, the first method costs one step of LU decomposition, m steps of forward and back substitution. Totally

$$\frac{2}{3}n^3 + 2mn^2 + O(n^2) + O(mn)$$

flops.
For the second algorithm, Using $\sum_{i=n}^{1}(2(i-1)i+2(n-i+1)(i-1)) = n^3+O(n^2)$ flops, we reduce from

$$\left(\begin{array}{cccc|cccc} a_{11} & a_{12} & \dots & a_{1n} & 1 & & & \\ a_{21} & a_{22} & \dots & a_{2n} & & 1 & & \\ \vdots & \vdots & \ddots & \vdots & & & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{nn} & & & & 1 \end{array}\right)$$

to

$$\left(\begin{array}{cccc|cccc} a_{11} & a_{12} & \dots & a_{1n} & 1 & & & \\ & a'_{22} & \dots & a''_{2n} & b_{21} & 1 & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & \\ & & & a''_{nn} & b_{n1} & b_{n2} & \dots & 1 \end{array}\right).$$

Costing another $\sum_{i=n}^{1}((i-1)+2n(i-1)) = n^3 + O(n^2)$, we finish finding the inverse of A. As a matrix multiplication also costs $nm(2n-1)$, totally it costs

$$2n^3 + 2mn^2 + O(n^2) + O(mn).$$

Therefore, when n and m is large enough, the first algorithm costs fewer flops. □

Question 2.3. *Let $\|\cdot\|$ be the two-norm. Given a nonsingular matrix A and a vector $\boldsymbol{b}$, show that for sufficiently small $\|\delta A\|$, there are nonzero δA and $\boldsymbol{\delta b}$ such that inequality*

$$\|\boldsymbol{\delta x}\| \le \|A^{-1}\|(\|\delta A\| \cdot \|\hat{\boldsymbol{x}}\| + \|\boldsymbol{\delta b}\|)$$

is an equality. This justifies calling $\kappa(A) = \|A^{-1}\| \cdot \|A\|$ the condition number of A.

Proof. To find the suitable $\boldsymbol{\delta x}$ and δA, let us have a glance at how the inequality comes. The original equation is

$$\boldsymbol{\delta x} = A^{-1}(-\delta A \cdot \hat{\boldsymbol{x}} + \boldsymbol{\delta b}).$$

Using the inequality $\|B\boldsymbol{x}\| \le \|B\| \cdot \|\boldsymbol{x}\|$ twice, and $\|\delta A\hat{\boldsymbol{x}} + \boldsymbol{\delta b}\| \le \|\delta A\hat{\boldsymbol{x}}\| + \|\boldsymbol{\delta b}\|$, we get the target inequality. If we denote $\boldsymbol{y}$ as the vector which satisfy $\|A^{-1}\| \cdot \|\boldsymbol{y}\| = \|A^{-1}\boldsymbol{y}\|$, set δA satisfy $\|-\delta A\| \cdot \|\hat{\boldsymbol{x}}\| = \|-\delta A\hat{\boldsymbol{x}}\|$, and $\delta A\hat{\boldsymbol{x}}$ proportional to $\boldsymbol{\delta b}$. The three inequalities will both become equalities. This suggests how we can construct δA and $\boldsymbol{\delta b}$.
Assume the δA has the form $\delta A = c\boldsymbol{a}\boldsymbol{b}^T$, where c is a positive constant and $\boldsymbol{a}, \boldsymbol{b}$ are vectors to be determined. According to Question 1.7., $\|-\delta A\| = c\|\boldsymbol{b}\| \cdot \|\boldsymbol{a}\|$. Then, as the δA must satisfy $\|-\delta A\| \cdot \|\hat{\boldsymbol{x}}\| = \|-\delta A\hat{\boldsymbol{x}}\|$, we get

$$c\|\boldsymbol{b}\| \cdot \|\boldsymbol{a}\| \cdot \|\hat{\boldsymbol{x}}\| = c\|-\boldsymbol{a}\boldsymbol{b}^T\hat{\boldsymbol{x}}\| = c\|\boldsymbol{b}^T\hat{\boldsymbol{x}}\| \cdot \|\boldsymbol{a}\| \le c\|\boldsymbol{b}\| \cdot \|\boldsymbol{a}\| \cdot \|\hat{\boldsymbol{x}}\|.$$

Therefore, for the equation to hold, $\boldsymbol{b}$ must be proportional to $\hat{\boldsymbol{x}}$, while $\boldsymbol{a}$ can be free of choice. Then, for $\delta A\hat{\boldsymbol{x}}$ to be proportional to $\boldsymbol{\delta b}$, we can set $\boldsymbol{b} = \hat{\boldsymbol{x}}$ and $\boldsymbol{a} = d\boldsymbol{\delta b}$, where d is another constant to be determined. As the final condistion requires

$$-\delta A \cdot \hat{\boldsymbol{x}} + \boldsymbol{\delta b} = (cd\hat{\boldsymbol{x}}^T\hat{\boldsymbol{x}} + 1)\boldsymbol{\delta b} = \boldsymbol{y},$$

we can choose c, d such that $cd\hat{\boldsymbol{x}}^T\hat{\boldsymbol{x}} + 1 \neq 0$. Therefore, we have

$$\delta A = d\boldsymbol{\delta b}\hat{\boldsymbol{x}}^T, \boldsymbol{\delta b} = (cd\hat{\boldsymbol{x}}^T\hat{\boldsymbol{x}} + 1)^{-1}\boldsymbol{y}.$$

We can see that both δA and $\boldsymbol{\delta b}$ are determined by known value, and due to the conditions we use, the target inequality becomes equality under our choice of δA and $\boldsymbol{\delta b}$, whose norm can be modified by the choice of c, d and $\|\boldsymbol{y}\|$. □

Question 2.4. *Show that bounds*

$$\|\boldsymbol{\delta x}\| \leq \epsilon \||A^{-1}(|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|)\|,$$
$$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \leq \epsilon \||A^{-1}| \cdot |A|\|.$$

are attainable.

Remark. I could only prove that the bounds can be attainable for certain A, whose inverse satisfy the following condition:

$$\big(\mathrm{sgn}(b_{i1}), \mathrm{sgn}(b_{i2}), \ldots, \mathrm{sgn}(b_{in})\big) = \pm\big(\mathrm{sgn}(b_{j1}), \mathrm{sgn}(b_{j2}), \ldots, \mathrm{sgn}(b_{jn})\big), \text{for any } i, j.$$

Also, I believe that second inequality should be

$$\frac{\|\boldsymbol{\delta x}\|}{\|\hat{\boldsymbol{x}}\|} \leq \epsilon \||A^{-1}| \cdot |A|\|.$$

Proof. For the first inequality, since it comes from

$$|\boldsymbol{\delta x}| = |A^{-1}(-\delta A\hat{\boldsymbol{x}} + \boldsymbol{\delta b})| \leq |A^{-1}|(|\delta A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{\delta b}|) \leq |A^{-1}|(\epsilon|A| \cdot |\hat{\boldsymbol{x}}| + \epsilon|\boldsymbol{b}|) = \epsilon(|A^{-1}|(|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|)),$$

it is inevitable that

$$|\delta A| = \epsilon|A|, |\boldsymbol{\delta b}| = \epsilon|\boldsymbol{b}|,$$

which means we could only choose the signs of entries of δA and $\boldsymbol{\delta b}$. From inside, we have $|\delta A\hat{\boldsymbol{x}}| = |\delta A||\hat{\boldsymbol{x}}|$. To make the inequality become equality, for each row, $\delta A(i, j)$ and $\hat{\boldsymbol{x}}_j$ should have the same sign. Therefore,

$$\big(\mathrm{sgn}(\delta A(i,1)), \mathrm{sgn}(\delta A(i,2)), \ldots, \mathrm{sgn}(\delta A(i,n))\big) = \pm\big(\mathrm{sgn}(\hat{\boldsymbol{x}}_1), \mathrm{sgn}(\hat{\boldsymbol{x}}_2), \ldots, \mathrm{sgn}(\hat{\boldsymbol{x}}_n)\big), \text{for any } i,$$

which reduces the degree of freedom of δA to n. Next, to make $|-\delta A\hat{\boldsymbol{x}} + \boldsymbol{\delta b}| = |-\delta A\hat{\boldsymbol{x}}| + |\boldsymbol{\delta b}|$, we have

$$\mathrm{sgn}(\sum_{i=1}^{n} -\delta A\hat{\boldsymbol{x}}) = \mathrm{sgn}(\boldsymbol{\delta b}),$$

which reduces the degree of freedom of $\boldsymbol{\delta b}$ to 0. For simplicity, denote $\boldsymbol{z} = -\delta A\hat{\boldsymbol{x}} + \boldsymbol{\delta b}$. Use the last n degree of freedom of δA, we could determine the sign of each $\boldsymbol{z}_i$. According to the condition of A^{-1}, set

$$\operatorname{sgn}(\boldsymbol{z}_i) = \operatorname{sgn}(A^{-1}(i,j)), \text{for any } i, j.$$

We will get $|A^{-1}\boldsymbol{z}| = |A^{-1}||\boldsymbol{z}|$. In all, we get

$$|\boldsymbol{\delta x}| = \epsilon |A^{-1}|(|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|).$$

Take norm on both sides, and the result follows. For the second inequality, the procedure is the same. Since under the assumption of $\boldsymbol{\delta b} = \boldsymbol{0}$, the only difference is to divide $\hat{\boldsymbol{x}}$ in both sides. □

Question 2.5. *Prove Theorem 2.3. Given the residual* $\boldsymbol{r} = A\hat{\boldsymbol{x}} - \boldsymbol{b}$, *use Theorem 2.3 to show that bound (2.9) is no larger than bound (2.7). This explains why LAPACK computes a bound based on (2.9), as described in section 2.4.4.*

Proof. To prove Theorem 2.3. Rearrange the equation $(A + \delta A)\hat{\boldsymbol{x}} = \boldsymbol{b} + \boldsymbol{\delta b}$, yielding

$$|A\hat{\boldsymbol{x}} - \boldsymbol{b}| = |\boldsymbol{r}| = |\boldsymbol{\delta b} - \delta A\hat{\boldsymbol{x}}| \le |\boldsymbol{\delta b}| + |\delta A\hat{\boldsymbol{x}}| \le |\boldsymbol{\delta b}| + |\delta A| \cdot |\hat{\boldsymbol{x}}| \le \epsilon(|\boldsymbol{b}| + |A| \cdot |\hat{\boldsymbol{x}}|).$$

As the above inequality is componentwise, it follows

$$\epsilon \ge \max_i \frac{|\boldsymbol{r}_i|}{(|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|)_i}.$$

Use the same technique describe in the previous question, it can be proved that for certain δA and $\boldsymbol{\delta b}$

$$|A\hat{\boldsymbol{x}} - \boldsymbol{b}| = |\boldsymbol{\delta b} - \delta A\hat{\boldsymbol{x}}| = |\boldsymbol{\delta b}| + |\delta A| \cdot |\hat{\boldsymbol{x}}| \le \epsilon(|\boldsymbol{b}| + |A| \cdot |\hat{\boldsymbol{x}}|),$$

which implies that ϵ cannot be smaller than $\max_i \frac{|\boldsymbol{r}_i|}{(|A|\cdot|\hat{\boldsymbol{x}}|+|\boldsymbol{b}|)_i}$. There proves the question. Next, to prove $\||A^{-1}| \cdot |\boldsymbol{r}|\| \le \epsilon \||A^{-1}|(|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|)\|$. Because all entries of $|A^{-1}|, |\boldsymbol{r}|$, and $|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|$ are nonnegative, we just need to prove that $|\boldsymbol{r}_i| \le \epsilon(|A| \cdot |\hat{\boldsymbol{x}}| + |\boldsymbol{b}|)_i$. According to the smallest value of ϵ, the targeted inequality holds. Therefore, bound (2.9) is no larger than bound (2.7). □

Question 2.6. *Prove Lemma 2.2.*

Proof. First, consider the single permuted permutation matrix. That is, the permutation matrix with only one pair of rows permuted. It can be written as

$$P = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 0 & \cdots & 1 & & \\ & & & \ddots & & & \\ & & 1 & \cdots & 0 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}, \text{ where } i\text{th}, j\text{th rows are permuted.}$$

According to the form of the single permutation matrix, it follows $\det(P) = \pm 1$. Give a matrix $X = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_n) = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \dots, \boldsymbol{\beta}_n^T)^T$. The matrix product PX and XP have the form

$$PX = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \dots, \boldsymbol{\beta}_j^T, \dots, \boldsymbol{\beta}_i^T, \dots, \boldsymbol{\beta}_n^T)^T,$$
$$XP = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_j, \dots, \boldsymbol{\alpha}_i, \dots, \boldsymbol{\alpha}_n).$$

Mark this as property-one. Then, to transform P into identity matrix I, it just need to permute the permuated ith and jth rows again. Consequently, $P^{-1} = P$ for single permuted permutation matrix.

For any permutation matrix $\hat{P}$, decompose it into a series of single permutation matrices as P_i, $i = 1, 2, \dots, m$. i.e.,

$$\hat{P} = P_m P_{m-1} \cdots P_1.$$

According to what we have proved for the single permutation matrix, it is only necessary to verify the property about P^{-1}. Since $\hat{P}^T = P_1 P_2 \cdots P_m$, we have

$$\hat{P}\hat{P}^T = \hat{P}^T \hat{P} = I.$$

Thus, the question is proved. □

Question 2.7. *If A is a nonsingular symmetric matrix and has the facotization $A = LDM^T$, where L and M are unit lower triangluar matrices and D is a diagonal matrix, show that $L = M$.*

Proof. To prove the question, partition L by its columns and M by its rows. We have

$$L = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_n),$$
$$M = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \dots, \boldsymbol{\beta}_n^T)^T.$$

Then, we can decompose A as

$$A = LDM^T = \sum_{i=1}^{n} D_{ii} \boldsymbol{\alpha}_i \boldsymbol{\beta}_i^T.$$

Recall that both L and M is unit lower triangular matrices. Therefore, $D_{ii}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T$ only has nonzero entries in $i : n$ submatrix, and $\boldsymbol{\alpha}_i(i) = \boldsymbol{\beta}_i(i) = 1$. i.e., $D_{ii}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T$ should have form as

$$\begin{pmatrix} \mathbf{0} & \dots & \mathbf{0} & D_{ii}\boldsymbol{\alpha}_i & * \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ D_{ii}\boldsymbol{\beta}_i \\ * \end{pmatrix}.$$

As A is nonsingular symmetric, we know $D_{ii} \neq 0$ and $D_{11}\boldsymbol{\alpha}_1 = D_{11}\boldsymbol{\beta}_1$, which means $\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_1$. Then, denote $A_1 = A - D_{11}\boldsymbol{\alpha}_1\boldsymbol{\beta}_1$. Apply the same technique to $A(2:n, 2:n)$, which is also nonsingular and symmetric. As $D_{ii}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T$ only has nonzero entries in $i : n$ submatrix. We can prove that $\boldsymbol{\alpha}_2 = \boldsymbol{\beta}_2$. In this analogy, it proves $L = M$. □

Question 2.8. *Consider the following two ways of solving a 2-by-2 linear system of equations:*

$$A\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \boldsymbol{b}.$$

1. *Algorithm 1. Gauusian elimination with partial pivoting (GEPP).*
2. *Algorithm 2. Cramer's rule:*

$$\begin{aligned} \det &= a_{11} * a_{22} - a_{12} * a_{21}, \\ x_1 &= (a_{22} * b_1 - a_{12} * b_2)/\det, \\ x_2 &= (-a_{21} * b_1 + a_{11} * b_2)/\det. \end{aligned}$$

Show by means of a numerical example that Cramer's rule is not backward stable.

Proof. To varify whether these two algorithms are backward stable or not, it requires to verify whether

$$\|\boldsymbol{\delta x}\|/\|\hat{\boldsymbol{x}}\| = O(\epsilon).$$

Consider the following example:

$$\begin{bmatrix} 1.112 & 1.999 \\ 3.398 & 6.123 \end{bmatrix} \cdot \boldsymbol{x} = \begin{bmatrix} 2.003 \\ 6.122 \end{bmatrix}.$$

The approximate answer should be $\boldsymbol{x}^T = \left(1.63788, 0.0908866\right)$. However, under four-decimal-digit floating point arthimetic, for the first algorithm, it yields

$$\begin{bmatrix} 1 & \\ 0.3273 & 1 \end{bmatrix} \begin{bmatrix} 3.398 & 6.123 \\ & -0.005 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 6.122 \\ 2.003 \end{bmatrix}.$$

Solving it, we have

$$\boldsymbol{x} = \begin{bmatrix} 1.441 \\ 0.2 \end{bmatrix}.$$

For the second algorithm, we get

$$\begin{aligned} \mathrm{fl}(\det) &= \mathrm{fl}(\mathrm{fl}(1.112 * 6.123) - \mathrm{fl}(1.999 * 3.398)) = \mathrm{fl}(6.809 - 6.793) = 0.016, \\ \mathrm{fl}(x_1) &= \mathrm{fl}(\mathrm{fl}(\mathrm{fl}(6.123 * 2.003) - \mathrm{fl}(1.999 * 6.122))/\det) = 1.875, \\ \mathrm{fl}(x_2) &= \mathrm{fl}(\mathrm{fl}(\mathrm{fl}(1.112 * 6.122) - \mathrm{fl}(3.398 * 2.003))/\det) = 0.125. \end{aligned}$$

Thus,

$$\boldsymbol{\delta x}_1 = \begin{bmatrix} 0.19688 \\ -0.1091134 \end{bmatrix}, \boldsymbol{\delta x}_2 = \begin{bmatrix} -0.23712 \\ -0.0341134 \end{bmatrix}.$$

It follows that $\|\boldsymbol{\delta x}_2\|_1/\|\boldsymbol{x}\|_1 \approx 0.156894 \approx 156\epsilon = O(1)$, which suggest that the second algorithm is not backward stable. □

Question 2.9. *Let B be an n-by-n upper bidiagonal matrix, i.e., nonzero only on the main diagonal and first superdiagonal. Derive an algorithm for computing $\kappa_\infty(B) \equiv \|B\|_\infty \|B^{-1}\|_\infty$* ***exactly*** *(ignoring roundoff). In other words, you should not use an iterative algothrim such as Hager's estimator. Your algorithm should be as cheap as possilbe; it should be possible to do using no more than $2n-2$ additions, n multiplications, n divisions, $4n-2$ absolute values, and $2n-2$ comparisions.*

Proof. For a bidiagonal matrix to have inverse, all the diagonal entries of B must be nonzero. Thus, we have the explicit form of B and B^{-1} as follow:

$$B = \begin{pmatrix} b_{1,1} & b_{1,2} & & & \\ & b_{2,2} & b_{2,3} & & \\ & & \ddots & \ddots & \\ & & & b_{n-1,n-1} & b_{n-1,n} \\ & & & & b_{n,n} \end{pmatrix},$$

$$B^{-1} = \begin{pmatrix} 1/b_{1,1} & -b_{1,2}/b_{2,2}b_{1,1} & b_{1,2}b_{2,3}/b_{1,1}b_{2,2}b_{3,3} & \cdots & (-1)^{n-1}\prod_{i=1}^{n-1} b_{i,1+1}/\prod_{i=1}^{n} b_{i,i} \\ & 1/b_{2,2} & -b_{2,3}/b_{2,2}b_{3,3} & \cdots & (-1)^{n-2}\prod_{i=2}^{n-1} b_{i,1+1}/\prod_{i=2}^{n} b_{i,i} \\ & & \ddots & \ddots & \vdots \\ & & & 1/b_{n-1,n-1} & -b_{n-1,n}/b_{n-1,n-1}b_{n,n} \\ & & & & 1/b_{n,n} \end{pmatrix}.$$

Thus, the absolute sum of ith row of B^{-1} can be compute from $(i+1)$th row by the following recursive formula:

$$s_i = s_{i+1}\left|b_{i,i+1}/b_{i,i}\right| + \left|1/b_{i,i}\right|.$$

Assume that we have done the $2n-1$ absolute value operations. The condition number can be computed as:

```
$MAX1 = B(n,n), MAX2 = 1/B(n,n), LASTSUM = MAX2$
for $i = 2$ to $n$ do
    $TEMP = B(i,i) + B(i,i+1)$
    if $TEMP > MAX1$ then
        $MAX1 = TEMP$
    end if
    $LASTSUM = (LASTSUM * B(i,i+1) + 1)/B(i,i)$
    if $LASTSUM > MAX2$ then
        $MAX2 = LASTSUM$
    end if
end for
```

Totally, it uses $2n-2$ additions, $n-1$ multiplications, n divisions, $2n-1$ absolute values and $2n-2$ comparisions, which meets the requirement. □

Question 2.10. *Let A be n-by-m with $n \geq m$. Show that $\|A^T A\|_2 = \|A\|_2^2$ and $\kappa_2(A^T A) = \kappa_2(A)^2$. Let M be n-by-n and positive definite and L be its Cholesky factor so that $M = LL^T$. Show that $\|M\|_2 = \|L\|_2^2$ and $\kappa_2(M) = \kappa_2(L)^2$.*

Proof. As A may be rectangular, denote $A = U\Sigma V^T$ as the SVD decomposition of A. Since

$$\|A^T A\|_2 = \|V\Sigma^2 V^T\|_2 = \|\Sigma^2\|_2 = \|\Sigma\|_2^2 = \|U\Sigma V^T\|_2^2 = \|A\|_2^2,$$

$\|A^T A\|_2 = \|A\|_2^2$ is proved. For equality $\kappa_2(A^T A) = \kappa_2(A)^2$, it also follows from $A^T A = V\Sigma^2 V^T$, which implies $\sigma_{\min}(A^T A) = \sigma_{\min}(A)^2$. Since the Cholesky factor matrix L is square and invertible, apply the above result and we prove the question. □

Question 2.11. *Let A be symmetic and positive definite. Show that* $|a_{ij}| \le (a_{ii}a_{jj})^{1/2}$.

Proof. Since A is s.p.d., all its diagonal entries are positive. Consider

$$\begin{aligned} \boldsymbol{x}_{ij} &= \{0, 0, \ldots, \sqrt{a_{jj}}, \ldots, -\sqrt{a_{ii}}, \ldots, 0\}, \\ \boldsymbol{y}_{ij} &= \{0, 0, \ldots, \sqrt{a_{jj}}, \ldots, \sqrt{a_{ii}}, \ldots, 0\}, \end{aligned}$$

yielding

$$\begin{aligned} \boldsymbol{x}_{ij}^T A\boldsymbol{x}_{ij} &= a_{jj}a_{ii} - \sqrt{a_{ii}a_{jj}}a_{ij} - \sqrt{a_{ii}a_{jj}}a_{ij} + a_{ii}a_{jj} = 2a_{ii}a_{jj} - 2\sqrt{a_{ii}a_{jj}}a_{ij} \ge 0, \\ \boldsymbol{y}_{ij}^T A\boldsymbol{y}_{ij} &= a_{jj}a_{ii} + \sqrt{a_{ii}a_{jj}}a_{ij} + \sqrt{a_{ii}a_{jj}}a_{ij} + a_{ii}a_{jj} = 2a_{ii}a_{jj} + 2\sqrt{a_{ii}a_{jj}}a_{ij} \ge 0, \end{aligned}$$

Combine them, and it follows

$$(a_{ii}a_{jj})^{1/2} \ge |a_{ij}|.$$

□

Question 2.12. *Show that if*

$$Y = \begin{pmatrix} I & Z \\ 0 & I \end{pmatrix},$$

where I is an n-by-n identity matrix, then $\kappa_F(Y) = \|Y\|_F\|Y^{-1}\|_F = (2n + \|Z\|_F^2)$.

Proof. Consider the following matrix

$$X = \begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix},$$

which satisfies $XY = YX = I$. Thus,

$$\kappa_F(Y) = \|Y\|_F\|Y^{-1}\|_F = \sqrt{(2n + \|Z\|_F^2)}\sqrt{(2n + \|Z\|_F^2)} = (2n + \|Z\|_F^2).$$

□

Question 2.13. *In this question we will ask how to solve* $B\boldsymbol{y} = \boldsymbol{c}$ *given a fast way to solve* $A\boldsymbol{x} = \boldsymbol{b}$, *where* $A - B$ *is "small" in some sense.*

1. *Prove the* ***Sherman-Morrison formula****: Let A be nonsingular,* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *be column vectors, and* $A + \boldsymbol{u}\boldsymbol{v}^T$ *be nonsingular. Then* $(A + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = A^{-1} - (A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1})/(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u})$.

   More generally, prove the ***Sherman-Morrison-Woodbury formula****: Let U and V be n-by-k rectangular matrices, where* $k \le n$ *and A is n-by-n. Then* $T = I + V^T A^{-1}U$ *is nonsingular if and only if* $A + UV^T$ *is nonsingular, in which case* $(A + UV^T)^{-1} = A^{-1} - A^{-1}UT^{-1}V^T A^{-1}$.

2. *If you have a fast algorithm to solve $A\boldsymbol{x} = \boldsymbol{b}$, show how to build a fast solver for $B\boldsymbol{y} = \boldsymbol{c}$, where $B = A + \boldsymbol{u}\boldsymbol{v}^T$.*

3. *Suppose that $\|A - B\|$ is "small" and you have a fast algorithm for solving $A\boldsymbol{x} = \boldsymbol{b}$. Describe an iterative scheme for solving $B\boldsymbol{y} = \boldsymbol{c}$. How fast do you expect your algorithm to converge?*

Proof.

1. First, it needs to prove that $\det(A + \boldsymbol{u}\boldsymbol{v}^T) = \det(A)\det(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u})$. Consider

$$\begin{pmatrix} I & \\ \boldsymbol{v}^T & 1 \end{pmatrix}\begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix}\begin{pmatrix} A + \boldsymbol{u}\boldsymbol{v}^T & \boldsymbol{u} \\ & 1 \end{pmatrix}\begin{pmatrix} I & \\ -\boldsymbol{v}^T & 1 \end{pmatrix}\begin{pmatrix} A & \\ & 1 \end{pmatrix} = \begin{pmatrix} A & A^{-1}\boldsymbol{u}A \\ & 1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u} \end{pmatrix}$$

Take determinant on both sides, and the result follows. Then, we can verify *Sherman-Morrison formula* as follows:

$$\begin{aligned}
&(A + \boldsymbol{u}\boldsymbol{v}^T)(A^{-1} - (A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1})/(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u})) \\
&\qquad = I + \boldsymbol{u}\boldsymbol{v}^T A^{-1} - (\boldsymbol{u}\boldsymbol{v}^T A^{-1} - \boldsymbol{u}\boldsymbol{v}^T A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1})/(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u}) \\
&\qquad = I + \boldsymbol{u}\boldsymbol{v}^T A^{-1} - (\boldsymbol{u}\boldsymbol{v}^T A^{-1})(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u})/(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u}) \\
&\qquad = I, \\
&(A^{-1} - (A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1})/(1+\boldsymbol{v}^T A^{-1}\boldsymbol{u}))(A + \boldsymbol{u}\boldsymbol{v}^T) \\
&\qquad = I + A^{-1}\boldsymbol{u}\boldsymbol{v}^T - (A^{-1}\boldsymbol{u}\boldsymbol{v}^T - A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1}\boldsymbol{u}\boldsymbol{v}^T)/(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u}) \\
&\qquad = I + A^{-1}\boldsymbol{u}\boldsymbol{v}^T - (A^{-1}\boldsymbol{u}\boldsymbol{v}^T)(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u})/(1 + \boldsymbol{v}^T A^{-1}\boldsymbol{u}) \\
&\qquad = I.
\end{aligned}$$

Thus, we prove the firs part of part I.

For the *Sherman-Morrison-Woodbury formula*, we first prove that $\det(A+UV^T) = \det(A)\det(I + V^T A^{-1}U)$. Consider

$$\begin{pmatrix} I & \\ V^T & I \end{pmatrix}\begin{pmatrix} A^{-1} & \\ & I \end{pmatrix}\begin{pmatrix} A + UV^T & U \\ & I \end{pmatrix}\begin{pmatrix} I & \\ -V^T & I \end{pmatrix}\begin{pmatrix} A & \\ & I \end{pmatrix} = \begin{pmatrix} A & A^{-1}UA \\ & I + V^T A^{-1}U \end{pmatrix}$$

Take determinant on both sides, and the result follows. As A is nonsingular, we prove they are mutually necessary and sufficient. Then, verify the matrix equation as follows:

$$\begin{aligned}
(A + UV^T)(A^{-1} - A^{-1}UT^{-1}V^T A^{-1}) &= I - UT^{-1}V^T A^{-1} + UV^T A^{-1} - UV^T A^{-1}UT^{-1}V^T A^{-1} \\
&= I + UV^T A^{-1} - (U + UV^T A^{-1}U)T^{-1}V^T A^{-1} \\
&= I + UV^T A^{-1} - U(I + V^T A^{-1}U)T^{-1}V^T A^{-1} \\
&= I, \\
(A^{-1} - A^{-1}UT^{-1}V^T A^{-1})(A + UV^T) &= I - A^{-1}UT^{-1}V^T + A^{-1}UV^T - A^{-1}UT^{-1}V^T A^{-1}UV^T \\
&= I + A^{-1}UV^T - A^{-1}UT^{-1}(V^T + V^T A^{-1}UV^T) \\
&= I + A^{-1}UV^T - A^{-1}UT^{-1}(I + V^T A^{-1}U)V^T \\
&= I.
\end{aligned}$$

In all, we prove part I.

2. Denote $\boldsymbol{x}_1, \boldsymbol{x}_2$, which satisfy $A\boldsymbol{x}_1 = \boldsymbol{u}, A\boldsymbol{x}_2 = \boldsymbol{c}$. According to what we have prove, to solve $B\boldsymbol{y} = \boldsymbol{c}$, we have

$$\begin{aligned}\boldsymbol{y} = B^{-1}\boldsymbol{c} &= (A^{-1} - (A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1})/(1+\boldsymbol{v}^T A^{-1}\boldsymbol{u}))\boldsymbol{c} \\ &= \boldsymbol{x}_2 - ((\boldsymbol{v}^T\boldsymbol{x}_2)\boldsymbol{x}_1)/(1+\boldsymbol{v}^T\boldsymbol{x}_1)\end{aligned}$$

Therefore, using the above formula, the solution can be computed from two solutions of $A\boldsymbol{x} = \boldsymbol{b}$ in addition to inner product, all of which are fast.

3. For the last question, since we have

$$B\boldsymbol{y} = (A+B-A)\boldsymbol{y} = A\boldsymbol{y} - (A-B)\boldsymbol{y} = \boldsymbol{c},$$

we get an equation as

$$\boldsymbol{y} = A^{-1}\boldsymbol{c} + A^{-1}(A-B)\boldsymbol{y}.$$

What we need to do is to find its root, thus we can use the fixed point iteration as

$$\boldsymbol{y}_{i+1} = A^{-1}\boldsymbol{c} + A^{-1}(A-B)\boldsymbol{y}_n,$$

or in pseudocode as:

```
while $\|\boldsymbol{r}\| > tol$ do
    solve $A\boldsymbol{x} = \boldsymbol{r}$
    $\boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$
    $\boldsymbol{r} = \boldsymbol{c} - B\boldsymbol{y}$
end while
```

For the fixed point iteration to converge, the norm of the coefficient matrix $A^{-1}(A-B)$ must smaller than one. Or, we can also get the same result from the iterative refinement prespective as:

$$\|\boldsymbol{r}_{i+1}\| = \|\boldsymbol{r}_i - B\boldsymbol{\delta y}\| = \|(A-B)\boldsymbol{\delta y}\| \le \|A-B\| \cdot \|\boldsymbol{\delta y}\| \le \|A-B\| \cdot \|A^{-1}\| \cdot \|\boldsymbol{r}_i\|.$$

Therefore, if $\|A-B\| \cdot \|A^{-1}\| < 1$, the algorithm will converge linearly.

□

Question 2.14. *Programming question. Skipped.*

Question 2.15. *Programming question. Skipped.*

Question 2.16. *Show how to reorganize the Cholesky algorithm (Algorithm 2.11.) to do most of its operations using Level 3 BLAS. Mimic Algorithm 2.10.*

Solution. As cholesky factorization does not need pivot, we first modify Algorithm 2.11 to fit rectangular matrices. In details, for a m-by-b rectangular matrix, use original Algorithm 2.11 to deal with the upper b-by-b submatrix, and use level 2 BLAS to update the lower $(m-n)$-by-n matrix, yielding

```
for $i = 1$ to $b$ do
```

```
    $A(i:m,i) = A(i:m,i)/\sqrt{A(i,i)}$
    for $j = i+1$ to $b$ do
        $A(j:b,j) = A(j:b,j) - A(j,i)A(j:b,i)$
    end for
    $A(b+1:m,i+1:b) = A(b+1:m,i:b) - A(b+1:m,i) \cdot A(i+1:b,i)^T$
end for
```

Then, by delaying the update of submatrix by b steps, we get level 3 BLAS implemented Cholesky. In details,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & \\ L_{12} & I \end{pmatrix} \begin{pmatrix} I & \\ & \bar{A} \end{pmatrix} \begin{pmatrix} L_{11}^T & L_{12}^T \\ & I \end{pmatrix} = \begin{pmatrix} L_{11}L_{11}^T & L_{11}L_{12}^T \\ L_{12}L_{11}^T & \bar{A} + L_{12}L_{12}^T \end{pmatrix}.$$

Thus, $\bar{A} = A_{22} - L_{12}L_{12}^T$, a symmetric rank-$b$ update which is level 3 BLAS.

```
for $i = 1$ to $n-1$ step $b$ do
    Use the above Algorithm to factorize $A(i:n,i:i+b-1) = \begin{pmatrix} L_{22} \\ L_{23} \end{pmatrix}$
    $A(i+b:n,i+b:n) = A(i+b:n,i+b:n) - L_{23}L_{23}^T$
end for
```

□

Question 2.17. *Suppose that, in Matlab, you have an n-by-n matrix A and an n-by-1 matrix b. What do A\b, b'/A, and A/b mean in Matlab? How does A\b differ from inv(A)*b?* [4]

Solution. A\b: equals to solve the linear systems of equations: $A\boldsymbol{x} = \boldsymbol{b}$;
b'/A: equals to solve $\boldsymbol{x}^T A = \boldsymbol{b}^T$;
A/b: ?;
Although both A\b and inv(A)*b get the solution of the linear systems of equations: $A\boldsymbol{x} = \boldsymbol{b}$, yet to solve $A\boldsymbol{x} = \boldsymbol{b}$ does not need to compute the inverse of A. It can be based on matrix factorization like LU, LL. □

Question 2.18. *Let*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

*where A_{11} is k-by-k and nonsingular. Then $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is called the **Schur complement** of A_{11} in A, or just Schur complement for short.*

1. *Show that after k steps of Gaussian elimination without pivoting, A_{22} has been overwritten by S.*

2. *Suppose $A = A^T$, A_{11} is positive definite, and A_{22} is negative definite. Show that A is nonsingular, that Gaussian elimination without pivoting will work in exact arithmetic, but that Gaussian elimination without pivoting may be numerically unstable.*

Proof.

[4] I don't use matlab much, so the answer may be modefied.

1. As we know, one step of Gaussian elimination overwrite the $n-1$ lower submatrix by Schur complement. If it holds for $(k-1)$th step of Gaussian elimination, it suffices to prove it is also true for the k steps of Gaussian elimination. Denote

$$A = \begin{bmatrix} A_{11} & \boldsymbol{v}_{k-1} & A_{12} \\ \boldsymbol{a}_{k-1}^T & a_{kk} & \boldsymbol{b}_{n-k}^T \\ A_{21} & \boldsymbol{u}_{n-k} & A_{22} \end{bmatrix}, \text{ where } A_{11} \text{ is } (k-1)\text{-by-}(k-1).$$

As kth step of Gaussian elimination equals to $(k-1)$th step of Gaussian elimination followed by one more Gaussian elimination, and the first $k-1$ steps will update $A(k:n,k:n)$ as

$$\begin{bmatrix} a_{kk} & \boldsymbol{b}_{n-k}^T \\ \boldsymbol{u}_{n-k} & A_{22} \end{bmatrix} - \begin{bmatrix} \boldsymbol{a}_{k-1}^T \\ A_{21} \end{bmatrix} A_{11}^{-1} \begin{bmatrix} \boldsymbol{v}_{k-1} & A_{12} \end{bmatrix} = \begin{bmatrix} a_{kk} - \boldsymbol{a}_{k-1}^T A_{11}^{-1} \boldsymbol{v}_{k-1} & \boldsymbol{b}_{n-k}^T - \boldsymbol{a}_{k-1}^T A_{11}^{-1} A_{12} \\ \boldsymbol{u}_{n-k} - A_{21} A_{11}^{-1} \boldsymbol{v}_{k-1} & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{bmatrix},$$

the next step will update A_{22} as

$$A_{22}' = A_{22} - A_{21} A_{11}^{-1} A_{12} - (\boldsymbol{u}_{n-k} - A_{21} A_{11}^{-1} \boldsymbol{v}_{k-1})(\boldsymbol{b}_{n-k}^T - \boldsymbol{a}_{k-1}^T A_{11}^{-1} A_{12}) / (a_{kk} - \boldsymbol{a}_{k-1}^T A_{11}^{-1} \boldsymbol{v}_{k-1}).$$

On the other hand, since the inverse of $\bar{A} = \begin{pmatrix} A_{11} & \boldsymbol{v}_{k-1} \\ \boldsymbol{a}_{k-1}^T & a_{kk} \end{pmatrix}$ is

$$\bar{A}^{-1} = \frac{1}{a_{kk} - \boldsymbol{a}_{k-1}^T A_{11}^{-1} \boldsymbol{v}_{k-1}} \begin{bmatrix} (a_{kk} - \boldsymbol{a}_{k-1}^T A_{11}^{-1} \boldsymbol{v}_{k-1}) A_{11}^{-1} + A_{11}^{-1} \boldsymbol{v}_{k-1} \boldsymbol{a}_{k-1}^T A_{11}^{-1} & -A_{11}^{-1} \boldsymbol{v}_{k-1} \\ -\boldsymbol{a}_{k-1}^T A_{11}^{-1} & 1 \end{bmatrix},$$

the corresponding kth step Schur complement is

$$A_{22} - \begin{bmatrix} A_{21} & \boldsymbol{u}_{n-k} \end{bmatrix} A' \begin{bmatrix} A_{12} \\ \boldsymbol{b}_{n-k}^T \end{bmatrix} = A_{22} - A_{21} A_{11}^{-1} A_{12} - (A_{21} A_{11}^{-1} \boldsymbol{v}_{k-1} \boldsymbol{a}_{k-1}^T A_{11}^{-1} A_{12} -$$
$$\boldsymbol{u}_{n-k} \boldsymbol{a}_{k-1}^T A^{-1} A_{12} - A_{21} A_{11}^{-1} \boldsymbol{v}_{k-1} \boldsymbol{b}_{n-k}^T + \boldsymbol{u}_{n-k} \boldsymbol{b}_{n-k}^T) / (a_{kk} - \boldsymbol{a}_{k-1}^T A_{11}^{-1} \boldsymbol{v}_{k-1}) = A_{22}'.$$

Therefore, it also holds for kth step of Gaussian elimination.

2. According to the condition, A_{11}, A_{22} is invertible. Thus, to show that A is nonsingular, consider its determinant as

$$\det(A) = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = \begin{vmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{vmatrix} = \det(A_{11}) \det(A_{22} - A_{21} A_{11}^{-1} A_{12}).$$

For any vector $\boldsymbol{x} \in \mathbb{R}^{n-k}$, we have

$$\boldsymbol{x}^T (A_{22} - A_{21} A_{11}^{-1} A_{12}) \boldsymbol{x} = \boldsymbol{x}^T A_{22} \boldsymbol{x} - (A_{12} \boldsymbol{x})^T A_{11} (A_{12} \boldsymbol{x}) \geq 0.$$

The equality holds if and only if $\boldsymbol{x} = \boldsymbol{0}$. So, $A_{22} - A_{21} A_{11}^{-1} A_{12}$ is s.p.d., which means $\det(A_{22} - A_{21} A_{11}^{-1} A_{12}) > 0$. As a result, A is nonsingular.

According to Question 2.11., all the diagonal entries of A are nonzero. Consequently the Gaussian elimination will work in exact arthimetic. To show it is numerical unstable, consider the following counterexample:

$$A = \begin{bmatrix} 1 & 10^{10} \\ 10^{10} & -1 \end{bmatrix} = \begin{bmatrix} 1 & \\ 10^{10} & 1 \end{bmatrix} \begin{bmatrix} 1 & 10^{10} \\ & -1 - 10^{20} \end{bmatrix}.$$

□

Question 2.19. *Matrix A is called* ***strictly column diagonally dominant****, or diagonally dominant for short, if*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ji}|.$$

1. *Show that A is nonsingular.*
2. *Show that Gaussian elimination with partial pivoting does not actually permute any rows, i.e., that it is identical to Gaussian elimination without pivoting.*

Proof.

1. According to Gershgorin's theorem, the eigenvalues λ of A^T satisfy

$$|\lambda - a_{ii}| \leq \sum_{j\neq i} |a_{ji}|.$$

Thus, we have

$$a_{ii} - \sum_{j\neq i} |a_{ji}| \leq \lambda \leq a_{ii} + \sum_{j\neq i} |a_{ji}|.$$

When $a_{ii} > 0$, we have $\lambda > 0$. Otherwise, $a_{ii} < 0$ will result in $\lambda < 0$. Since A is strictly column diagonally dominant, $a_{ii} \neq 0$. Thus, we prove $\lambda \neq 0$, which means $\det(A) = \det(A^T) = \prod_{i=1}^{n} \lambda \neq 0$. i.e., A is nonsingular.

2. Since $|a_{11}| > |a_{j1}|$, it holds for the base case. If we could prove that after updating the submatrix still has the diagonally dominant property, we would prove the question. Denote

$$a'_{ij} = a_{ij} - a_{1j} * a_{i1}/a_{11}.$$

Sum them up, we get

$$\begin{aligned} \sum_{i=2, i\neq j}^{n} |a'_{ij}| &\leq \sum_{i=2, i\neq j}^{n} (|a_{ij}| + |a_{1j} * a_{i1}/a_{11}|) \\ &\leq |a_{ii}| - |a_{1i}| - |a_{1j}| + |a_{1j}/a_{11}| \sum_{i=1, i\neq j}^{n} |a_{i1}| \\ &\leq |a_{ii}| - |a_{1i}| \leq |a_{ii}| - |a_{1i} * a_{i1}/a_{11}| \leq |a'_{ii}|. \end{aligned}$$

Therefore, the diagonally dominant property holds. By mathematic induction, we prove the quesion.

□

Question 2.20. *Given an n-by-n nonsingular matrix A, how do you efficiently solve the following problems, using Gaussian elimination with partial pivoting?*

(a) Solve the linear system $A^k \boldsymbol{x} = \boldsymbol{b}$, where k is a positive integer.

(b) Compute $\alpha = \boldsymbol{c}^T A^{-1}\boldsymbol{b}$.

(c) Solve the matrix equation $AX = B$, where B is n-by-m.

You should (1) descirbe your algorithms, (2) present them in pseudocode and (3) give the require flops.

Solution.

(a) Since $A^k\boldsymbol{x} = \boldsymbol{b}$ equals $A^{k-1}\boldsymbol{x} = A^{-1}\boldsymbol{b}$, set $A^{-1}\boldsymbol{b} = \boldsymbol{b}_1$, and we have $A^{k-1}\boldsymbol{x} = \boldsymbol{b}_1$, which is the same kind of equation with smaller exponent. Thus, we have the algorithm

```
for $i = k$ to 1 do
    solve $A\boldsymbol{x} = \boldsymbol{b}$
    $\boldsymbol{b} = \boldsymbol{x}$
end for
```

Totally, it requires $\frac{2}{3}n^3 + 2kn^2 + O(n^2)$ flops.

(b) First solve $A\boldsymbol{x} = \boldsymbol{b}$, then do the inner product. The pseudocode is shown below.

```
solve $A\boldsymbol{x} = \boldsymbol{b}$
$\alpha = \boldsymbol{c}^T\boldsymbol{x}$
```

Totally, it require $\frac{2}{3}n^3 + O(n^2)$ flops.

(c) Solve $AX = B$ equals to solve m vectors.

```
for $i = 1$ to $m$ do
    solve $A\boldsymbol{x} = \boldsymbol{b}_i$
end for
```

Totally, it require $\frac{2}{3}n^3 + 2mn^2 + O(n^2)$ flops.

□

Question 2.21. *Prove that Strassen's algorithm (Algorithm 2.8) correctly multiplies n-by-n matrices, where n is a power of 2.*

Proof. Since n is a power of 2, all matrix partitions and matrix multiplication will be ok. To prove the result, it suffice to prove the correctness of the recursive step. According to Strassen's algorithm, we have

$$\begin{aligned}
&P_1 = A_{12}B_{21} + A_{12}B_{22} - A_{22}B_{21} - A_{22}B_{22},\ P_4 = A_{11}B_{22} + A_{12}B_{22},\\
&P_2 = A_{11}B_{11} + A_{11}B_{22} + A_{22}B_{11} + A_{22}B_{22},\ P_5 = A_{11}B_{12} - A_{11}B_{22},\\
&P_3 = A_{11}B_{11} + A_{11}B_{22} - A_{21}B_{11} - A_{21}B_{12},\ P_6 = A_{22}B_{21} - A_{22}B_{11},\\
&P_7 = A_{21}B_{11} + A_{22}B_{11},\ C_{11} = A_{12}B_{21} + A_{11}B_{11},\\
&C_{12} = A_{12}B_{22} + A_{11}B_{12},\ C_{21} = A_{22}B_{21} + A_{21}B_{11},\\
&C_{22} = A_{22}B_{22} + A_{21}B_{12}.
\end{aligned}$$

Therefore,

$$A = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

So the recursive yield the correct answer, and the basic case for the Strassen's algorithm is just the scalar arithmtic operations. So, Strassen's algorithm correctly multiplies n-by-n matrices, where n is a power of 2. □

3 Solutions for Chapter III: Linear Least Squares Problems

Question 3.1. *Show that the two variations of Algorithm 3.1, CGS and MGS, are mathematically equivalent by showing that the two formulas for r_{ij} yield the same results in exact arithmetic.*

Proof. The differences between CGS and MGS are the different ways to calculate coefficients r_{ij}. For CGS, it maintains to use the original vectors $\boldsymbol{a}_i$, while the MGS uses the updated vectors $\boldsymbol{q}_i$, which after j steps of updates becomes

$$\boldsymbol{q}_i = \boldsymbol{a}_i - \sum_{k=1}^{j} r_{ki}\boldsymbol{q}_k.$$

Thus, as for any $j < i$, $\boldsymbol{q}_j$ are mutually orthogonal, it is obvious that

$$r_{(j+1)i} = \boldsymbol{q}_{j+1}^T\boldsymbol{q}_i = \boldsymbol{q}_{j+1}^T(\boldsymbol{a}_i - \sum_{k=1}^{j} r_{ki}\boldsymbol{q}_j) = \boldsymbol{q}_{j+1}^T\boldsymbol{a}_i.$$

So, by induction, CGS and MGS are mathematically equivalent. □

Question 3.2. *Programming question. Skipped.*

Question 3.3. *Let A be m-by-n, $m \geq n$, and have full rank.*

1. *Show that $\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}$ has a solution where $\boldsymbol{x}$ minimizes $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$.*

2. *What is the conition number of the coefficient matrix in terms of the singular values of A?*

3. *Give an explicit expression for the inverse of the coefficient matrix, as a block 2-by-2 matrix.*

4. *Show how to use the QR decomposition of A to implement an iterative refinement algorithm to improve the accuracy of $\boldsymbol{x}$.*

Solution.

1. We can decompose the equation $\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}$ into the following two equations:

$$\begin{aligned} \boldsymbol{r} + A\boldsymbol{x} &= \boldsymbol{b}, \\ A^T\boldsymbol{r} &= \boldsymbol{0}. \end{aligned}$$

Thus, if $\boldsymbol{x}$ is the original solution, we have

$$A^T\boldsymbol{r} + A^TA\boldsymbol{x} = A^TA\boldsymbol{x} = A^T\boldsymbol{b},$$

which means $\boldsymbol{x}$ satisfies the normal equation. Therefore, $\boldsymbol{x}$ minimizes $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$.

2. Since A has full column rank, let $U^TAV = \Sigma$ the full SVD of A, where U, V are m-by-m, n-by-n orthgonal matirces and Σ is m-by-n diagonal matrix with nonzero entries. Applying orthgonal similarity transformation onto the coefficient matrix, we have

$$\begin{bmatrix} U^T & \\ & V^T \end{bmatrix} \begin{bmatrix} I_m & A \\ A^T & \end{bmatrix} \begin{bmatrix} U & \\ & V \end{bmatrix} = \begin{bmatrix} I & \Sigma \\ \Sigma^T & \end{bmatrix}.$$

Since Σ has full colum rank, we can compute the eigenvalues of the above matrix by applying elementary transformations with unit determinant as

$$\begin{aligned} \det\begin{bmatrix} (1-\lambda)I_m & \Sigma \\ \Sigma^T & -\lambda I_n \end{bmatrix} &= \det\begin{bmatrix} (1-\lambda)I_m & \Sigma \\ \Sigma^T & -\lambda I_n \end{bmatrix} \det\begin{bmatrix} I_m & \lambda\Sigma(\Sigma^T\Sigma)^{-1} \\ & I_n \end{bmatrix} \\ &= \det\begin{bmatrix} I_m & \\ -\Sigma^T & I_n \end{bmatrix} \det\begin{bmatrix} (1-\lambda)I_m & \lambda(1-\lambda)\Sigma(\Sigma^T\Sigma)^{-1} + \Sigma \\ \Sigma^T & \end{bmatrix} \\ &= \det\begin{bmatrix} I_m & \Sigma(\Sigma^T\Sigma)^{-1} \\ & I_n \end{bmatrix} \det\begin{bmatrix} (1-\lambda)I_m & \lambda(1-\lambda)\Sigma(\Sigma^T\Sigma)^{-1} + \Sigma \\ \lambda\Sigma^T & \lambda(\lambda-1)I_n - \Sigma^T\Sigma \end{bmatrix} \\ &= \det\begin{bmatrix} (1-\lambda)I_m + \lambda\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T & \\ \lambda\Sigma^T & \lambda(\lambda-1)I_n - \Sigma^T\Sigma \end{bmatrix} \\ &= (1-\lambda)^{m-n}\prod_{i=1}^{n}(\lambda^2 - \lambda - \sigma_i^2). \end{aligned}$$

Thus, the eigenvalues of the coefficient matrix are 1 and $\frac{1}{2}(1 \pm \sqrt{1+4\sigma_i^2})$. As $\sqrt{1+4\sigma_i^2} > 1$, we have $1 + \sqrt{1+4\sigma_i^2} > 2$. Consequently, the condition number is

$$|\lambda_{\max}/\lambda_{\min}| = \begin{cases} \frac{1}{2}\left(1 + \sqrt{1+4\sigma_{\max}^2}\right) / \min\left\{\frac{1}{2}\left(\sqrt{1+4\sigma_{\min}^2} - 1\right), 1\right\}, & \text{when } m > n; \\ \left(1 + \sqrt{1+4\sigma_{\max}^2}\right) / \left(\sqrt{1+4\sigma_{\min}^2} - 1\right), & \text{when } m = n; \end{cases}$$

3. To find its inverse, we use invertible matrices to transform the coefficient matrix into indentity matrix as follows

$$\begin{bmatrix} I_m & \\ & -(A^TA)^{-1} \end{bmatrix} \begin{bmatrix} I_m & A(A^TA)^{-1} \\ & I_n \end{bmatrix} \begin{bmatrix} I_m & \\ -A^T & I_n \end{bmatrix} \begin{bmatrix} I_m & A \\ A^T & \end{bmatrix} = \begin{bmatrix} I_m & \\ & I_n \end{bmatrix}.$$

Thus, we have the candidate matrix as

$$\begin{bmatrix} I_m & \\ & -(A^TA)^{-1} \end{bmatrix} \begin{bmatrix} I_m & A(A^TA)^{-1} \\ & I_n \end{bmatrix} \begin{bmatrix} I_m & \\ -A^T & I_n \end{bmatrix} = \begin{bmatrix} I_m - A(A^TA)^{-1}A^T & A(A^TA)^{-1} \\ (A^TA)^{-1}A^T & -(A^TA)^{-1} \end{bmatrix} = B.$$

It is easy to check that $B \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} = \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} B = I_{m+n}$. Consequently, B is the inverse of coefficient matrix.

4. Let $A = QR$ the QR factorization of A, where Q is m-by-m orthgonal matrix and R is m-by-n upper triangluar matrix with positive entries in $R(i,i)$. Since

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} = \begin{bmatrix} I & QR \\ R^TQ^T & \end{bmatrix} = \begin{bmatrix} Q & \\ & I \end{bmatrix} \begin{bmatrix} I & R \\ R^T & \end{bmatrix} \begin{bmatrix} Q^T & \\ & I \end{bmatrix},$$

the main work in solving the original linear equations lies in solving

$$\begin{bmatrix} I & R \\ R^T & \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 + R\boldsymbol{x}_2 \\ R^T\boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix},$$

which can be solved easily. Thus, we can apply iterative refinement onto the above equation to improve the answer.

□

Question 3.4. ***Weighted least squares:*** *If some components of $A\boldsymbol{x} - \boldsymbol{b}$ are more important than others, we can weight them with a scale factor d_i and solve the weighted least squares problem* $\min \|D(A\boldsymbol{x} - \boldsymbol{b})\|_2$ *instead, where D has diagnoal entries d_i. More generally, recall that if C is symmetric positive definite, then $\|\boldsymbol{x}\|_C \equiv (\boldsymbol{x}^TC\boldsymbol{x})^{1/2}$ is a norm, and we can consider minimizing $\|A\boldsymbol{x} - \boldsymbol{b}\|_C$. Derive the normal equations for this problem, as well as the formulation corresponding to the previous quesion.*

Solution. To derive the normal equations, we look for the $\boldsymbol{x}$, where the gradient of $\|A\boldsymbol{x} - \boldsymbol{b}\|_C^2 = (A\boldsymbol{x} - \boldsymbol{b})^TC(A\boldsymbol{x} - \boldsymbol{b})$ vanishes, i.e.,

$$\begin{aligned} 0 &= \lim_{\boldsymbol{e}\to\boldsymbol{0}} \left((A\boldsymbol{x} + A\boldsymbol{e} - \boldsymbol{b})^TC(A\boldsymbol{x} + A\boldsymbol{e} - \boldsymbol{b}) - (A\boldsymbol{x} - \boldsymbol{b})^TC(A\boldsymbol{x} - \boldsymbol{b})\right) / \boldsymbol{e}^TC\boldsymbol{e} \\ &= \lim_{\boldsymbol{e}\to\boldsymbol{0}} \left(2\boldsymbol{e}(A^TCA\boldsymbol{x} - A^TC\boldsymbol{b}) + \boldsymbol{e}^TA^TCA\boldsymbol{e}\right) / \boldsymbol{e}^TC\boldsymbol{e} \\ &= \lim_{\boldsymbol{e}\to\boldsymbol{0}} \left(2\boldsymbol{e}(A^TCA\boldsymbol{x} - A^TC\boldsymbol{b})\right) / \boldsymbol{e}^TC\boldsymbol{e}. \end{aligned}$$

Thus, we get its normal equations as

$$A^TCA\boldsymbol{x} = A^TC\boldsymbol{b}.$$

To derive the corresponding formulation like previous question, consider the following symmetric full rank linear systems of equations

$$\begin{bmatrix} I & C^{1/2}A \\ A^TC^{1/2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} C^{1/2}\boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}.$$

Thus, if we take $C^{1/2}A$ as $\tilde{A}$ and $C^{1/2}\boldsymbol{b}$ as $\tilde{\boldsymbol{b}}$, we have the same equation as before and the previous result also applies. □

Question 3.5. *Let $A \in \mathbb{R}^{n\times n}$ be positive definite. Two vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are called A-orthgonal if $\boldsymbol{u}_1^T A\boldsymbol{u}_2 = 0$. If $U \in \mathbb{R}^{n\times r}$ and $U^T AU = I$, then the columns of U are said to be A-orthgonal. Show that every subspace has an A-orthgonal basis.*

Proof. Given any subspace, there exists a set of linear independent vectors $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n)$, which can span the whole space. Then, we use the given vectors to create an A-orthonormal basis $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n)$ as follows:

```
for $i = 1$ to $n$ do
    $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_i - \sum_{j=1}^{i-1}(\boldsymbol{\beta}_j^T A\boldsymbol{\alpha}_i)\boldsymbol{\beta}_j$
    $\boldsymbol{\beta}_i = \boldsymbol{\beta}_i / \sqrt{\boldsymbol{\beta}_i^T A\boldsymbol{\beta}_i}$
end for
```

Since $\boldsymbol{\alpha}_i$ span the whole space, the above process terminates at nth step. It is easy to verify that the first step holds the A-orthonormal property as

$$\boldsymbol{\beta}_1^T A\boldsymbol{\beta}_2 = \left(\boldsymbol{\beta}_1^T A\boldsymbol{\alpha}_2 - (\boldsymbol{\beta}_1^T A\boldsymbol{\alpha}_2)\boldsymbol{\beta}_1^T A\boldsymbol{\beta}_1\right) / \boldsymbol{\beta}_2^T A\boldsymbol{\beta}_2 = \left(\boldsymbol{\beta}_1^T A\boldsymbol{\alpha}_2 - \boldsymbol{\beta}_1^T A\boldsymbol{\alpha}_2\right) / \boldsymbol{\beta}_2^T A\boldsymbol{\beta}_2 = 0.$$

Thus, by mathematic induction, for any $i > j$ we have

$$\boldsymbol{\beta}_j^T A\boldsymbol{\beta}_i = \left(\boldsymbol{\beta}_j^T A\boldsymbol{\alpha}_i - (\boldsymbol{\beta}_j^T A\boldsymbol{\alpha}_i)\boldsymbol{\beta}_j^T A\boldsymbol{\beta}_j\right) / \boldsymbol{\beta}_i^T A\boldsymbol{\beta}_i = \left(\boldsymbol{\beta}_j^T A\boldsymbol{\alpha}_i - \boldsymbol{\beta}_j^T A\boldsymbol{\alpha}_i\right) / \boldsymbol{\beta}_i^T A\boldsymbol{\beta}_i = 0,$$

which implies $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n)$ is an A-orthgonal basis. If there exist some $\boldsymbol{\beta}_{i_1}, \boldsymbol{\beta}_{i_2}, \ldots, \boldsymbol{\beta}_{i_s}$ which are linear dependent, there will exist $c_1, c_2, \ldots, c_s$ such that

$$c_1\boldsymbol{\beta}_{i_1} + c_2\boldsymbol{\beta}_{i_2} + \ldots + c_s\boldsymbol{\beta}_{i_s} = 0.$$

Apply $\boldsymbol{\beta}_{i_k}^T A$ on both side, and we have $c_k = 0$, which implies they are linear independent. As $\boldsymbol{\alpha}_i$ span the whole space, the vector set $\boldsymbol{\beta}_i$ also form a basis of the space and is mutually A-orthonormal. □

Question 3.6. *Let A have the form*

$$A = \begin{bmatrix} R \\ S \end{bmatrix},$$

where R is n-by-n and upper triangular, and S is m-by-n and dense. Describe an algorithm using Householder transformations for reducing A to upper triangular form. Your algorithm should not "fill in" the zeros in R and thus require fewer operations than would Algorithm 3.2 applied to A.

Solution. Adding $R(1,1)$ into the first column of S, we get $\boldsymbol{x}_1 = \big(R(1,1), S(1,1), S(2,1)\ldots, S(m,1)\big)$. If $\|\boldsymbol{x}_1\|_2 = 0$, we are done and move into submatrix $A(2 : m+n, 2 : n)$. Otherwise, we will have an $(m+1)$-by-$(m+1)$ Householder transformation P, which satisfies $P\boldsymbol{x}_1 = \big(\|\boldsymbol{x}_1\|_2, 0, \ldots, 0\big)$. Thus, we can construct a symmetric orthgonal matrix $\hat{P}$ as

$$\hat{P} = \begin{bmatrix} P(1,1) & \mathbf{0} & P(1, 2 : m+1) \\ \mathbf{0} & I_{n-1} & \mathbf{0} \\ P(2 : m+1, 1) & \mathbf{0} & P(2 : m+1, 2 : m+1) \end{bmatrix}.$$

Apply $\hat{P}$ into A, we have

$$\hat{P}A = A_1 = \begin{bmatrix} \|\boldsymbol{x}_1\|_2 & * \\ \mathbf{0} & R(2:n,2:n) \\ \mathbf{0} & * \end{bmatrix}.$$

Thus, we reduce the problem into smaller subproblem. Continue, and we will reduce A into upper triangluar. □

Question 3.7. *If $A = R + \boldsymbol{u}\boldsymbol{v}^T$, where R is an upper triangular matrix, and* **u** *and* **v** *are column vectors, describe an effiecient algorithm to compute the QR decomposition of A.*

Solution. When either $\boldsymbol{u}$ or $\boldsymbol{v}$ is $\mathbf{0}$, the answer is thrivial. Suppose $\boldsymbol{u}_n \neq 0$. Otherwise the last row of $\boldsymbol{u}\boldsymbol{v}^T$ will be $\mathbf{0}$ and we can start from the $(n-1)$th row of $\boldsymbol{u}\boldsymbol{v}^T$. As the rank of $\boldsymbol{u}\boldsymbol{v}^T$ is only one and by assumption the last row of $\boldsymbol{u}\boldsymbol{v}^T$ is not $\mathbf{0}$, it follows that the last row of $\boldsymbol{u}\boldsymbol{v}^T$ can linear represent the remaining rows. Thus, there exist $n-1$ Givens rotations $G_1, G_2, \ldots, G_{n-1}$ which rotate ith and nth rows to cancel ith row. Since R is upper triangular, $G_i R$ is also upper triangular, which means

$$\bar{A} = G_{n-1}G_{n-2}\ldots G_1 A = G_{n-1}G_{n-2}\ldots G_1 R + G_{n-1}G_{n-2}\ldots G_1 \boldsymbol{u}\boldsymbol{v}^T$$

is upper triangular except for the last row. Then, we proceed to cancel the last row. First, if $R(1,1)$ is not zero, we use the first row of $\bar{A}$ to cancel $\bar{A}(n,1)$. Otherwise, we swap the first row and the last row of $\bar{A}$. Continue like this, and we reduce $\bar{A}$ to upper triangular, and all the matrices we apply to A forms the orthgonal matrix Q. □

Question 3.8. *Let $\boldsymbol{x} \in \mathbb{R}^n$ and let P be a Householder matrix such that $P\boldsymbol{x} = \pm\|\boldsymbol{x}\|_2\boldsymbol{e}_1$. Let $Q_{1,2}, \ldots, Q_{n-1,n}$ be Givens rotations, and let $Q = G_{1,2}\ldots G_{n-1,n}$. Suppose $Q\boldsymbol{x} = \pm\|\boldsymbol{x}\|_2\boldsymbol{e}_1$. Must P equal Q?*

Solution. P does not need to equal Q. Consider the two dimension counterexample as

$$P\boldsymbol{x} = \frac{1}{5}\begin{bmatrix} -3 & -4 \\ -4 & 3 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix} = -\|\boldsymbol{x}\|_2\boldsymbol{e}_1.$$

The above matrix is a Householder matrix with $\boldsymbol{u} = \frac{1}{5}\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ as generating vector. In the same time, the following matrix is a Givens rotation, which satisfies the condition while does not equal to P,

$$Q\boldsymbol{x} = \frac{1}{5}\begin{bmatrix} -3 & -4 \\ 4 & -3 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix} = -\|\boldsymbol{x}\|_2\boldsymbol{e}_1.$$

□

Question 3.9. *Let A be m-by-n, with SVD $A = U\Sigma V^T$. Compute the SVDs of the following matrices in terms of U, Σ, and V:*

1. *$(A^TA)^{-1}$,*
2. *$(A^TA)^{-1}A^T$,*

3. $A(A^TA)^{-1}$,

4. $A(A^TA)^{-1}A^T$.

Remark. For the formulas to hold, we have to assume that A has full column rank. And for simplicity, later we will assume that $m \geq n$ if the contrary situation is not stated explicitly, since for the case $n \geq m$ can be treated similarly.
Proof.

1. for first formula, we have
$$(A^TA)^{-1} = (V\Sigma^TU^TU\Sigma V^T)^{-1} = (V\Sigma^2V^T)^{-1} = V\Sigma^{-2}V^T.$$

2. For second formula, we have
$$(A^TA)^{-1}A^T = V\Sigma^{-2}V^TV\Sigma U^T = V\Sigma^{-1}U^T.$$

3. For third formula, we have
$$A(A^TA)^{-1} = U\Sigma V^TV\Sigma^{-2}V^T = U\Sigma^{-1}V^T.$$

4. For the last formula, we have
$$A(A^TA)^{-1}A^T = U\Sigma^{-1}V^TV\Sigma U^T = I_n.$$

□

Question 3.10. *Let A_k be a best rank-k approximation of the matrix A, as defined in Part 9 of Theorem 3.3. Let σ_i be the ith singular value of A. Show that A_k is unique if $\sigma_k > \sigma_{k+1}$.*

Proof. I believe what the question claim is not correct. According to the textbook, the so-called best rank-k approximation of a matrix A under two-norm is a rank-k matrix which satisfies
$$\|A - \tilde{A}\|_2 = \sigma_{k+1}.$$

Thus, I present a counterexample to illustrate that it may not be unique under the assupmtion of the question. Consider the matrix A as

$$A = \operatorname{diag}(4,3,2,1) = \begin{bmatrix} 4 & & & \\ & 3 & & \\ & & 2 & \\ & & & 1 \end{bmatrix}.$$

Then any matrix $\tilde{A}$ like

$$\begin{bmatrix} 4-\epsilon & & & \\ & 3 & & \\ & & 2 & \\ & & & 0 \end{bmatrix},$$

where $3 < \epsilon < 4$, satisfies $\|A - \tilde{A}\|_2 = 1$. Therefore, the best rank-3 approximation of A is not unique. However, the result holds if we measure under Frobenius norm, and if the best rank-k approximation must have the form of $\sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$, the proof is trivial.

Remark. Under the assumption of the question, I can only prove that any matrix B satisfies $\|A - B\|_2 = \sigma_{k+1}$ has the condition

$$U^T BV = \begin{bmatrix} \boldsymbol{u}_1^T B\boldsymbol{v}_1 & \boldsymbol{u}_2^T B\boldsymbol{v}_1 & \dots & \boldsymbol{u}_k^T B\boldsymbol{v}_1 & \\ \boldsymbol{u}_1^T B\boldsymbol{v}_2 & \boldsymbol{u}_2^T B\boldsymbol{v}_2 & \dots & \boldsymbol{u}_k^T B\boldsymbol{v}_2 & \\ \vdots & \vdots & \ddots & \vdots & \\ \boldsymbol{u}_1^T B\boldsymbol{v}_k & \boldsymbol{u}_2^T B\boldsymbol{v}_k & \dots & \boldsymbol{u}_k^T B\boldsymbol{v}_k & \\ & & & & 0 \end{bmatrix},$$

where U and V are the orthgonoal matrices of the SVD of A. In some degree, the two norm equality $\|A - B\|_2 = \sigma_{k+1}$ does not provide enough information of the range of B, which results in its nonuniqueness. □

Question 3.11. *Let A be m-by-n. Show that $X = A^+$ (the Moore-Penrose pseudoinverse) minimizes $\|AX - I\|_F$ over all n-by-m matrices X. What is the value of this minimum?*

Proof. Denote the SVD decomposition of A as $A = U\Sigma V^T$. As there is an one-to-one correspondence from X to $\hat{X} = V^T XU^T$, we can substitue X by $\hat{X}$. If $\hat{X}$ is the minimizer, we can regain target X as $V\hat{X}U^T$ and vice verse. Consequently, the problem can be transformed into

$$\|AX - I\|_F^2 = \|U\Sigma V^T V\hat{X}U^T - I\|_F^2 = \|\Sigma\hat{X} - I\|_F^2 = \sum_{i=1}^{r} \left(\left(\sigma_i \hat{X}(i,i) - 1\right)^2 + \sum_{j \neq i}^{m} \sigma_i^2 (\hat{X}(i,j))^2 \right) + (n - r),$$

where r is the rank of A. Thus, to minimizes the above formula, we have

$$\hat{X}(i,i) = \sigma_i^{-1}, \hat{X}(i,j) = 0, \text{ where } i \in \{1,2,\dots,r\}, \text{ and } j \neq i \in \{1,2,\dots,m\}.$$

So, $\hat{X}$ may have many choices, and $\hat{X} = \Sigma^+$ always satisfies the condition. As a result, $X = A^+$ is the minimizer, and the minimum is $\sqrt{n - r}$. □

Question 3.12. *Let A, B, and C be matrices with dimensions such that the product $A^T CB^T$ is well defined. Let $\mathcal{X}$ be the set of matrices X minimizing $\|AXB - C\|_F$, and let X_0 be the unique member of $\mathcal{X}$ minimizing $\|X\|_F$. Show that $X_0 = A^+ CB^+$.*

Remark. This solution is unique only under the assumption that A and B have full rank.

Proof. Assume matrix A as m-by-s, B as t-by-n, C as m-by-n. Denote the SVD decomposition of A as $A = U_1\Sigma_1 V_1^T$, B as $B = U_2\Sigma_2 V_2^T$. As there is an one-to-one correspondence from X to $\hat{X} = V_1^T XU_2$, we can substitue X as $\hat{X}$. Thus, we have

$$\begin{aligned} \|AXB - C\|_F^2 &= \|U_1\Sigma_1 V_1^T V_1 \hat{X} U_2^T U_2 \Sigma_2 V_2^T - C\|_F^2 \\ &= \|\Sigma_1 \hat{X} \Sigma_2 - U_1^T CV_2\|_F^2 \\ &= \|(\Sigma_1 \hat{X} \Sigma_2 - U_1^T CV_2)(1 : r_1, 1 : r_2)\|_F^2 + \|(U_1^T CV_2)\left(((r_1 + 1) : m, 1 : n) + (1 : r_1, (r_2 + 1) : n)\right)\|_F^2, \end{aligned}$$

where r_1 is the rank of A and r_2 is the rank of B. As the matrix $U_1^T CV_2$ are set once A, B and C are given, the minimizer must and only have to minimize

$$\|(\Sigma_1 \hat{X}\Sigma_2 - U_1^T CV_2)(1:r_1, 1:r_2)\|_F^2.$$

Thus, any matrix $\hat{X}$, which satisfies $\Sigma_1\hat{X}\Sigma_2(1:r_1,1:r_2) = (U_1^T CV_2)(1:r_1,1:r_2)$ or equivalently for any matrix X which $\Sigma_1 V_1^T XU_2\Sigma_2(1:r_1,1:r_2) = (U_1^T CV_2)(1:r_1,1:r_2)$, is a minimizer, We claim that $X_0 = A^+CB^+$ is such a minimizer, since

$$\begin{aligned}\Sigma_1 V_1^T X_0 U_2\Sigma_2(1:r_1,1:r_2) &= \Sigma_1 V_1^T V_1\Sigma_1^+ U_1^T CV_2\Sigma_2^+ U_2^T U_2\Sigma_2(1:r_1,1:r_2)\\ &= \begin{bmatrix} I_{r_1} & 0\\ 0 & 0\end{bmatrix} U_1^T CV_2 \begin{bmatrix} I_{r_2} & 0\\ 0 & 0\end{bmatrix}(1:r_1,1:r_2)\\ &= (U_1^T CV_2)(1:r_1,1:r_2).\end{aligned}$$

□

Question 3.13. *Show that the Moore-Penrose pseudoinverse of A satisfies the following identities:*

$$AA^+A = A,\ A^+AA^+ = A^+,\ A^+A = (A^+A)^T,\ AA^+ = (AA^+)^T.$$

Proof. Denote the SVD decomposition of A as $U\Sigma V^T$, r as the rank of A. We prove these identities one by one.

1.
$$AA^+A = U\Sigma V^T V\Sigma^+U^T U\Sigma V^T = U\begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}\Sigma V^T = U\Sigma V^T = A.$$

2.
$$A^+AA^+ = V\Sigma^+U^T U\Sigma V^T V\Sigma^+U^T = V\begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}\Sigma^+U^T = V\Sigma^+U^T = A^+.$$

3.
$$A^+A = V\Sigma^+U^T U\Sigma V^T = V\begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}V^T = (V\begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}V^T)^T = (A^+A)^T.$$

4.
$$AA^+ = U\Sigma V^T V\Sigma^+U^T = U\begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}U^T = (U\begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}U^T)^T = (AA^+)^T.$$

Question 3.14. *Prove part 4 of Theorem 3.3: Let* $H = \begin{bmatrix} 0 & A^T\\ A & 0\end{bmatrix}$, *where A is square and* $A = U\Sigma V^T$ *is its SVD. Let* $\Sigma = diag(\sigma_1, \sigma_2 \ldots, \sigma_n)$, $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$, *and* $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$. *Prove that the* $2n$ *eigenvalues of H are* $\pm\sigma_i$, *with corresponding unit eigenvectors* $\frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{v}_i\\ \pm\boldsymbol{u}_i\end{bmatrix}$. *Extend to the case of retangular A.*

Proof. We directly prove the case of retangular A, thus the square case is proved without special treatment. First, denote the SVD decomposition of A as $U\Sigma V^T = \sum_{i=1}^n \sigma_i \boldsymbol{u}\boldsymbol{v}^T$, where $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$, $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ and $\Sigma = \text{diag}(\sigma_1, \sigma_2 \ldots, \sigma_n)$. We verify that $\pm\sigma_i$ are the $2n$ eigenvalues of H with corresponding unit eigenvectors $\frac{1}{\sqrt{2}}[\boldsymbol{v}_i, \pm\boldsymbol{u}_i]^T$ as follows:

$$\begin{aligned}
H\begin{bmatrix}\boldsymbol{v}_i\\ \boldsymbol{u}_i\end{bmatrix} &= \begin{bmatrix}0 & \sum_{i=1}^n \sigma_i \boldsymbol{v}\boldsymbol{u}^T\\ \sum_{i=1}^n \sigma_i \boldsymbol{u}\boldsymbol{v}^T & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_i\\ \boldsymbol{u}_i\end{bmatrix}\\
&= \begin{bmatrix}\sigma_i\boldsymbol{v}_i\\ \sigma_i\boldsymbol{u}_i\end{bmatrix} = \sigma_i\begin{bmatrix}\boldsymbol{v}_i\\ \boldsymbol{u}_i\end{bmatrix},\\
H\begin{bmatrix}\boldsymbol{v}_i\\ -\boldsymbol{u}_i\end{bmatrix} &= \begin{bmatrix}0 & \sum_{i=1}^n \sigma_i \boldsymbol{v}\boldsymbol{u}^T\\ \sum_{i=1}^n \sigma_i \boldsymbol{u}\boldsymbol{v}^T & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_i\\ -\boldsymbol{u}_i\end{bmatrix}\\
&= \begin{bmatrix}-\sigma_i\boldsymbol{v}_i\\ \sigma_i\boldsymbol{u}_i\end{bmatrix} = -\sigma_i\begin{bmatrix}\boldsymbol{v}_i\\ -\boldsymbol{u}_i\end{bmatrix}.
\end{aligned}$$

Then, as

$$\left\|\begin{bmatrix}\boldsymbol{v}_i\\ \pm\boldsymbol{u}_i\end{bmatrix}\right\|_2 = \sqrt{2},$$

and all $\frac{1}{\sqrt{2}}[\boldsymbol{v}_i, \pm\boldsymbol{u}_i]^T$ are linear indenpendent, $\pm\sigma_i$ are the $2n$ eigenvalues of H with corresponding unit eigenvectors $\frac{1}{\sqrt{2}}[\boldsymbol{v}_i, \pm\boldsymbol{u}_i]^T$. As for the remaining $m-n$ eigenvalues, let $\tilde{U} = [\boldsymbol{u}_{n+1}, \ldots, \boldsymbol{u}_m]$ be the orthgonal supplement of U. We can claim that the remaining $m-n$ eigenvalues of H are zeros, with corresponding unit eigenvectors $[\boldsymbol{0}, \boldsymbol{u}_i]^T$, since

$$H\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{u}_i\end{bmatrix} = \begin{bmatrix}0 & \sum_{i=1}^n \sigma_i \boldsymbol{v}\boldsymbol{u}^T\\ \sum_{i=1}^n \sigma_i \boldsymbol{u}\boldsymbol{v}^T & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{u}_i\end{bmatrix} = \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}.$$

□

Question 3.15. *Let A be m-by-n, $m < n$, and of full rank. Then* $\min\|A\boldsymbol{x} - \boldsymbol{b}\|_2$ *is called an* ***underdetermined least squares problem****. Show that the solution is an $(n-m)$-dimensional set. Show how to compute the unique minimum norm solution using appropriately modified normal equations, QR decomposition, and SVD.*

Solution. We explain these three techniques one by one, and prove the properties of the solution in the first technique.

1. For modified normal equation, partition any vector $\boldsymbol{x} \in \mathbb{R}^n$ into $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_2$ is in the null space of A and $\boldsymbol{x}_1$ is in the range of A^T. Thus,

   $$\|A\boldsymbol{x} - \boldsymbol{b}\|_2 = \|A\boldsymbol{x}_1 - \boldsymbol{b}\|_2 = \|AA^T\boldsymbol{y} - \boldsymbol{b}\|_2.$$

   As AA^T has full rank, the normal equation

   $$AA^T\boldsymbol{y} = \boldsymbol{b}$$

   has unique solution. Every vector in the form $A^T\boldsymbol{y} + \boldsymbol{x}_2$ is the minimizer, which forms an $(n-m)$-dimensional set. The unique minimum norm solution is $\boldsymbol{x} = A^T(AA^T)^{-1}\boldsymbol{b}$.

2. For QR decompostion, since A^T has the normal QR decomposition as $A^T = QR$ where Q is n-by-m and R is m-by-m, we have that $A = R^T Q^T$, and the least square problem equals the equation
$$R^T Q^T \boldsymbol{x} = \boldsymbol{b}.$$
Let $\boldsymbol{x} = Q(R^T)^{-1}\boldsymbol{b} + \tilde{\boldsymbol{x}}$, where $\tilde{\boldsymbol{x}}$ can be any vector. We can transform the above equation into
$$Q^T \tilde{\boldsymbol{x}} = \mathbf{0}.$$
Thus, $\tilde{\boldsymbol{x}}$ belong to the null space of Q^T. Evaluate the norm of our solution, we have
$$\|\boldsymbol{x}\|_2^2 = \|Q(R^T)^{-1}\boldsymbol{b} + \tilde{\boldsymbol{x}}\|_2^2 = \|Q(R^T)^{-1}\boldsymbol{b}\|_2^2 + \|\tilde{\boldsymbol{x}}\|_2^2 \geq \|Q(R^T)^{-1}\boldsymbol{b}\|_2^2.$$
Consequently, the unique minimum norm solution is $Q(R^T)^{-1}\boldsymbol{b}$.

3. For SVD, we can have the SVD decomposition of A as $A = V\Sigma U^T$, where V is m-by-m, Σ is m-by-m full rank diagonal matrix and U^T is m-by-n. Like QR decomposition, the least square problem equals to
$$U^T \boldsymbol{x} = \Sigma^{-1} V^T \boldsymbol{b}.$$
Let $\boldsymbol{x} = U\Sigma^{-1}V^T\boldsymbol{b} + \tilde{\boldsymbol{x}}$ and we have the solution of above linear equations as
$$U^T \tilde{\boldsymbol{x}} = \mathbf{0}.$$
Thus, $\tilde{\boldsymbol{x}}$ belong to the null space of U^T. Evaluate the norm of our solution, we have
$$\|\boldsymbol{x}\|_2^2 = \|U\Sigma^{-1}V^T\boldsymbol{b} + \tilde{\boldsymbol{x}}\|_2^2 = \|U\Sigma^{-1}V^T\boldsymbol{b}\|_2^2 + \|\tilde{\boldsymbol{x}}\|_2^2 \geq \|U\Sigma^{-1}V^T\boldsymbol{b}\|_2^2.$$
Consequently, the unique minimum norm solution is $U\Sigma^{-1}V^T\boldsymbol{b}$.

□

Question 3.16. *Prove Lemma 3.1:*
Let P be an exact Householder (or Givens) transformation, and $\tilde{P}$ be its floating point approximation. Then
$$\mathit{fl}(\tilde{P}A) = P(A+E) \qquad \|E\|_2 = O(\epsilon)\|A\|_2$$
and
$$\mathit{fl}(A\tilde{P}) = (A+F)P \qquad \|F\|_2 = O(\epsilon)\|A\|_2.$$

Proof. If $\mathrm{fl}(a \odot b) = (a \odot b)(1+\epsilon)$ holds, like Question 1.10., we can prove that
$$\mathrm{fl}(AB) - AB = O(\epsilon)AB.$$
Thus, for the first formula, we have
$$\begin{aligned}\|E\|_2 = \|P^{-1}\left(\mathrm{fl}(\tilde{P}A) - PA\right)\|_2 &= \|P^{-1}\left(\mathrm{fl}((1+O(\epsilon))PA) - PA\right)\|_2 \\ &\leq O(\epsilon)\|P^{-1}\|_2\|P\|_2\|A\|_2 \\ &= O(\epsilon)\|A\|_2.\end{aligned}$$
The second one can be deduced similarly. □

Question 3.17. *The question is too long, so I omit it. For details, please refer to the textbook.*

Solution.

1. We prove it by induction. First, consider the base case as $P = P_2P_1$. We have

$$\begin{aligned} P = P_2P_1 &= (I - \boldsymbol{u}_2\boldsymbol{u}_2^T)(I - \boldsymbol{u}_1\boldsymbol{u}_1^T) \\ &= I - 2\boldsymbol{u}_1\boldsymbol{u}_1^T - 2\boldsymbol{u}_2\boldsymbol{u}_2^T - 4\boldsymbol{u}_2^T\boldsymbol{u}_1\boldsymbol{u}_2\boldsymbol{u}_1^T \\ &= I - [\boldsymbol{u}_1, \boldsymbol{u}_2]\begin{bmatrix} 2 & \\ 4\boldsymbol{u}_2^T\boldsymbol{u}_1 & 2 \end{bmatrix}\begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \end{bmatrix}. \end{aligned}$$

Thus, the base case holds. Then, assume that $P' = P_kP_{k-1}\cdots P_1 = I - U_kT_kU_k^T$, then we have

$$\begin{aligned} P = P_{k+1}P' &= (I - 2\boldsymbol{u}_{k+1}\boldsymbol{u}_{k+1}^T)(I - U_kT_kU_k^T) \\ &= I - U_kT_kU_k^T - 2\boldsymbol{u}_{k+1}\boldsymbol{u}_{k+1}^T + 2\boldsymbol{u}_{k+1}\boldsymbol{u}_{k+1}^T U_kT_kU_k^T \\ &= I - [U_k, \boldsymbol{u}_{k+1}]\begin{bmatrix} T_k & \\ 2\boldsymbol{u}_{k+1}^T U_kT_k & 2 \end{bmatrix}\begin{bmatrix} U_k^T \\ \boldsymbol{u}_{k+1}^T \end{bmatrix} \\ &= I - U_{k+1}T_{k+1}U_{k+1}^T. \end{aligned}$$

Therefore, by induction, we have proved part I of the question. As for algorithm, we can use the recursive formula we get in induction as

$$T_{k+1} = \begin{bmatrix} T_k & \\ 2\boldsymbol{u}_{k+1}^T U_kT_k & 2 \end{bmatrix}.$$

And the base case is $T_1 = 2$. The pseudocode is shown below:

```
$T_1 = 1$
for $i = 2$ to $k$ do
    $\boldsymbol{\beta}_i = 2\boldsymbol{u}_i^T U_{i-1}T_{i-1}$
    $T_i = \begin{bmatrix} T_{i-1} & \\ \boldsymbol{\beta}_i & 2 \end{bmatrix}$
end for
```

2. The algorithm which the textbook introduces is just sufficient:

```
for $i = 1$ to $\min(m-1, n)$ do
    $\boldsymbol{u}_i =$ House($A(i:m, i)$)
    $\boldsymbol{v} = \boldsymbol{u}_i^T A(i:m, i:n)$
    $A(i:m, i:n) = A(i:m, i:n) - 2\boldsymbol{u}_i\boldsymbol{v}^T$
end for
```

Here, $\boldsymbol{u}_i^T A(i:m, i:n)$ is a matrix-vector multiplication and $A(i:m, i:n) = A(i:m, i:n) - 2\boldsymbol{u}_i\boldsymbol{v}^T$ is a rank-1 update. Since House($A(i:m, i)$) costs $2(m-i)$, the total flops are $\sum_{i=1}^{\min(m-1,n)}(2m-2i-1)(n-i) + 2(m-i)(n-i) + 2(m-i) \approx 2mn^2 - \frac{2}{3}n^3$.

3. Denote b as a small integer. Like Level 3 BLAS Gaussian elimination, we use the technique of delaying update. In details, assume we have already done the first $k-1$ columns, yielding

$$A \quad = \quad \begin{array}{ccc} & k-1 & n-k+1 \\ k-1 & Q_{11} & Q_{12} \\ m-k-1 & Q_{21} & Q_{22} \end{array} \cdot \left(\begin{array}{ccc} k-1 & b & n-k-b+1 \\ R_{11} & R_{12} & R_{13} \\ & \tilde{A}_{22} & \tilde{A}_{23} \\ & \tilde{A}_{32} & \tilde{A}_{33} \end{array} \right)$$

Then, apply the Level 2 BLAS QR implement to the submatrix $\begin{bmatrix} \tilde{A}_{22} \\ \tilde{A}_{23} \end{bmatrix}$ and get

$$\begin{bmatrix} \tilde{A}_{22} & \tilde{A}_{23} \\ \tilde{A}_{32} & \tilde{A}_{33} \end{bmatrix} = Q \begin{bmatrix} R_{22} & R_{23} \\ & \bar{A}_{33} \end{bmatrix},$$

where $Q\begin{bmatrix} R_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{A}_{22} \\ \tilde{A}_{32} \end{bmatrix}$ and $Q\begin{bmatrix} R_{23} \\ \bar{A}_{33} \end{bmatrix} = \begin{bmatrix} \tilde{A}_{23} \\ \tilde{A}_{33} \end{bmatrix}$. Thus,

$$\begin{aligned} A &= \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} I & \\ & Q \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ & R_{22} & R_{23} \\ & & \bar{A}_{33} \end{bmatrix} \\ &= \begin{bmatrix} Q_{11} & Q_{12}Q \\ Q_{21} & Q_{22}Q \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ & R_{22} & R_{23} \\ & & \bar{A}_{33} \end{bmatrix} \end{aligned}$$

The Q can be computed by part I, and $R_{23}, \bar{A}_{33}$ can be computed from matrix product, which is level 3 BLAS.

□

Question 3.18. *It is often of interest to solve* ***constrained least squares problems****, where the solution* $\boldsymbol{x}$ *must satisfy a linear or nonlinear constraint in addition to minimizing* $\|A\boldsymbol{x}-\boldsymbol{b}\|_2$*. We consider one such problem here. Suppose that we want to choose* $\boldsymbol{x}$ *to minimize* $\|A\boldsymbol{x}-\boldsymbol{b}\|_2$ *subject to the linear constraint* $C\boldsymbol{x}=\boldsymbol{d}$*. Suppose also that* A *is* m*-by-*n*,* C *is* p*-by-*n*, and* C *has full rank. We also assume that* $p \le n$ *(so* $C\boldsymbol{x}=\boldsymbol{d}$ *is guaranteed to be consisitent) and* $n \le m+p$ *(so the system is not underdetermined). Show that there is a unique solution under the assumption that* $\begin{bmatrix} A \\ C \end{bmatrix}$ *has full column rank. Show how to compute* $\boldsymbol{x}$ *using two QR decompositions and some matrix-vector multiplications and solving some triangular systems of equations.*

Solution. First, compute a full QR decomposition of C as $C = R^T Q^T$, where $Q^T = [Q_1, Q_2]^T$ is n-by-n and $R^T = [L, 0]$ is p-by-n. Then the constraint $C\boldsymbol{x} = \boldsymbol{d}$ can be transformed into

$$R^T Q^T \boldsymbol{x} = [L, 0] \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} \boldsymbol{x} = L Q_1^T \boldsymbol{x} = \boldsymbol{d}.$$

Like Question 3.15., the solutions of the above linear system are $Q_1L^{-1}\boldsymbol{d}+Q_2\boldsymbol{y}$, where $\boldsymbol{y}$ is any vector in $\mathbb{R}^{n-p}$. Therefore, the original least square problem equals to

$$\|A\boldsymbol{x}-\boldsymbol{b}\|_2 = \|A(Q_1L^{-1}\boldsymbol{d}+Q_2\boldsymbol{y})-\boldsymbol{b}\|_2 = \|AQ_2\boldsymbol{y}-(\boldsymbol{b}-AQ_1L^{-1}\boldsymbol{d})\|_2.$$

As AQ_2 is m-by-$(n-p)$ and $m+p\geq n$, we have $m\geq n-p$. If AQ_2 has full rank, we will get the unique solution. Since $\begin{bmatrix} A \\ C \end{bmatrix}$ has full column rank, we claim that the intersection of the null space of A and the null space of B is empty. If not, suppose $\boldsymbol{x}$ is one vector of the intersection, then $\begin{bmatrix} A \\ C \end{bmatrix}\boldsymbol{x}=\boldsymbol{0}$, which means the columns of $\begin{bmatrix} A \\ C \end{bmatrix}$ are linear dependent. Thus, the null spaces of A and C do not intersect. Consequently, AQ_2 has full rank. [5] If we denote the QR decomposition of AQ_2 as $AQ_2=\hat{Q}\hat{R}$, we have the unique solution of $\boldsymbol{y}$ as

$$\boldsymbol{y}=\hat{R}^{-1}\hat{Q}^T(\boldsymbol{b}-AQ_1L^{-1}\boldsymbol{d}).$$

Then the final answer should be

$$\boldsymbol{x}=AQ_2\boldsymbol{y}+Q_1L^{-1}\boldsymbol{d}.$$

□

Question 3.19. *Programming question. Skipped.*

Question 3.20. *Prove Theorem 3.4.*

Proof. First, since

$$(A^TA)^{-1}A^T(A+\delta A) = I+(A^TA)^{-1}A^T\delta A$$

and $\|A^TA^{-1}A^T\delta A\|_2 = 1/\sigma_n$, $I+(A^TA)^{-1}A^T\delta A$ is invertible accroding to Lemma 2.1. Consequently, $(A+\delta A)$ has full rank. As $\tilde{\boldsymbol{x}}$ minimizes the problem $\|(A+\delta A)\tilde{\boldsymbol{x}}-(\boldsymbol{b}+\boldsymbol{\delta b})\|_2$, it can be expressed as

$$\begin{aligned}
\tilde{\boldsymbol{x}} &= \left((A+\delta A)^T(A+\delta A)\right)^{-1}(A+\delta A)^T(\boldsymbol{b}+\boldsymbol{\delta b}) \\
&= \left(A^TA(I+(A^TA)^{-1}A^T\delta A+(A^TA)^{-1}(\delta A)^TA+(A^TA)^{-1}(\delta A)^T(\delta A)\right)^{-1}(A+\delta A)^T(\boldsymbol{b}+\boldsymbol{\delta b}) \\
&= \sum_{i=0}^{\infty}(-1)^i\left((A^TA)^{-1}A^T\delta A+(A^TA)^{-1}(\delta A)^TA+(A^TA)^{-1}(\delta A)^T(\delta A)\right)^i(A^TA)^{-1}(A+\delta A)^T(\boldsymbol{b}+\boldsymbol{\delta b}) \\
&\approx \left(I-(A^TA)^{-1}(\delta A)^TA-(A^TA)^{-1}A^T\delta A\right)(A^TA)^{-1}(A^T\boldsymbol{b}+(\delta A)^T\boldsymbol{b}+A^T\boldsymbol{\delta b}) \\
&\approx (I-(A^TA)^{-1}(\delta A)^TA-(A^TA)^{-1}A^T\delta A)\boldsymbol{x}+(A^TA)^{-1}(\delta A)^T\boldsymbol{b}+(A^TA)^{-1}A^T\boldsymbol{\delta b} \\
&= (I-(A^TA)^{-1}A^T\delta A)\boldsymbol{x}+(A^TA)^{-1}(\delta A)^T\boldsymbol{r}+(A^TA)^{-1}A^T\boldsymbol{\delta b}.
\end{aligned}$$

[5] Denote $Q_2=\left(\boldsymbol{q}_1,\boldsymbol{q}_2,\dots,\boldsymbol{q}_{n-p}\right)$. If there exist $b_1,b_2,\dots,b_{n-p}$ such that

$$b_1A\boldsymbol{q}_1+b_2A\boldsymbol{q}_2+\ldots+b_{n-p}A\boldsymbol{q}_{n-p}=A(b_1\boldsymbol{q}_1+b_2\boldsymbol{q}_2+\ldots+b_{n-p}\boldsymbol{q}_{n-p})=0,$$

it follows $b_1\boldsymbol{q}_1+b_2\boldsymbol{q}_2+\ldots+b_{n-p}\boldsymbol{q}_{n-p}$ is in the null space of A, which is also in the null space of C.

Also, since $\sin\theta = \frac{\|\boldsymbol{r}\|_2}{\|\boldsymbol{b}\|_2}$ and θ is acute, we can evaluate $\cos\theta$ and $\tan\theta$ as following:

$$\cos\theta = \sqrt{1-\sin^2\theta} = \sqrt{\frac{\|\boldsymbol{b}\|_2^2 - \|\boldsymbol{r}\|_2^2}{\|\boldsymbol{b}\|_2^2}} = \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{b}\|_2},$$
$$\tan\theta = \frac{\sin\theta}{\cos\theta} = (\|\boldsymbol{r}\|_2/\|\boldsymbol{b}\|_2)/(\|A\boldsymbol{x}\|_2/\|\boldsymbol{b}\|) = \frac{\|\boldsymbol{r}\|_2}{\|A\boldsymbol{x}\|_2}.$$

Thus,

$$\begin{aligned}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2 &\leq \epsilon\|(A^TA)^{-1}A^T\|_2\|A\|_2\|\boldsymbol{x}\|_2 + \epsilon\|(A^TA)^{-1}\|_2\|A\|_2\|\boldsymbol{r}\|_2 + \epsilon\|(A^TA)^{-1}A^T\|_2\|\boldsymbol{b}\|_2 \\ &\leq \epsilon\kappa_2(A)\|\boldsymbol{x}\|_2 + \epsilon\kappa_2^2(A)\|\boldsymbol{x}\|_2\tan\theta + \epsilon\kappa_2(A)\|\boldsymbol{x}\|_2/\cos\theta \\ &\leq \epsilon\left(\frac{2\kappa_2(A)}{\cos\theta} + \tan\theta\cdot\kappa_2^2(A)\right).\end{aligned}$$

□

4 Solutions for Chapter IV: Nonsymmetric Eigenvalue Problems

Question 4.1. *Let A be defined as in equation (4.1). Show that* $\det(A) = \prod_{i=1}^b \det(A_{ii})$ *and then that* $\det(A-\lambda I) = \prod_{i=1}^b \det(A_{ii}-\lambda I)$*. Conclude that the set of eigenvalues of A is the union of the sets of eigenvalues of A_{11} through A_{bb}.*

Proof. Denote A_{ii} as m_i-by-m_i and the whole matrix A as n-by-n. From the definition of determinant, we have $\det(A) = \sum(-1)^{\sigma_i} a_{i_1,1}a_{i_2,2},\dots,a_{i_n,n}$, where σ_i is the inverse number of the permutation $(i_1, i_2, \dots, i_n)$. Then we can group them as

$$\det(A) = \sum(-1)^{\sigma_i}(a_{i_1,1},\dots,a_{i_{m_1},m_1})\dots(a_{i_{m_{b-1}+1},m_{b-1}+1},\dots,a_{i_{m_b},m_b}).$$

For the first group, $a_{i_j,j} = 0$ when $i_j > m_1$. Thus, we can delete them from the sum, and get

$$\begin{aligned}\det(A) &= \sum(-1)^{\sigma_1}(a_{i_1,1},\dots,a_{i_{m_1},m_1})(-1)^{\sigma_{i'}}(a_{i_{m_1+1},m_1+1},\dots,a_{i_{m_2},m_2})\dots(a_{i_{m_{b-1}+1},m_{b-1}+1},\dots,a_{i_{m_b},m_b}) \\ &= \det(A_{11})\sum(-1)^{\sigma_{i'}}(a_{i_{m_1+1},m_1+1},\dots,a_{i_{m_2},m_2})\dots(a_{i_{m_{b-1}+1},m_{b-1}+1},\dots,a_{i_{m_b},m_b}).\end{aligned}$$

Therefore, we reduce our problem into smaller subproblem. Continue, and we have

$$\det(A) = \det(A_{11})\det(A_{22})\dots\det(A_{bb}) = \prod_{i=1}^b \det(A_{ii}).$$

Consequently,

$$\det(A-\lambda I) = \prod_{i=1}^b \det(A_{ii}-\lambda I).$$

Thus λ is an eigenvalue of A if and only if it is an eigenvalue of some A_{ii}. □

Question 4.2. *Suppose that A is normal; i.e., $AA^* = A^*A$. Show that if A is also triangular, it must be diagonal. Use this to show that an n-by-n matrix is normal if and only if it has n orthonormal eigenvectors.*

Proof. Since A is normal and triangular, we have

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{bmatrix} \begin{bmatrix} \bar{a}_{11} & & & \\ \bar{a}_{12} & \bar{a}_{22} & & \\ \vdots & \ddots & \ddots & \\ \bar{a}_{1n} & \bar{a}_{2n} & \dots & \bar{a}_{nn} \end{bmatrix} = \begin{bmatrix} \bar{a}_{11} & & & \\ \bar{a}_{12} & \bar{a}_{22} & & \\ \vdots & \ddots & \ddots & \\ \bar{a}_{1n} & \bar{a}_{2n} & \dots & \bar{a}_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{bmatrix}.$$

Equating $(1,1)$ entry, we have

$$|a_{11}|^2 = \sum_{i=1}^{n} |a_{1n}|^2.$$

Thus, for $i \neq 1$ we have $|a_{1i}| = 0$. Then we reduce the problem into smaller subproblem. Continue, and we will have A is diagonal. For any n-by-n normal matrix A, we have a unitary matrix U such that $U^* AU = T$, where T is triangular. Then, we claim that T is normal because

$$T^* T = U^* A^* UU^* AU = U^* A^* AU = U^* AA^* U = U^* AUU^* A^* U = TT^*.$$

Thus, T is diagonal and denote it as $\mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, we have

$$AU = U\mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

i.e., each column of U is the right eigenvector of A. On the other hand, if A has n orthonormal eigenvectors, denote them as U. We have $A = U\mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)U^*$, and it is easy to verify that A is normal. □

Question 4.3. *Let λ and μ be distinct eigenvalues of A, let $\boldsymbol{x}$ be a right eigenvector for λ, and let $\boldsymbol{y}$ be a left eigenvector for μ. Show that $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal.*

Proof. Consider $\boldsymbol{y}^* A\boldsymbol{x}$, and we have

$$\boldsymbol{y}^* A\boldsymbol{x} = \mu \boldsymbol{y}^* \boldsymbol{x} = \lambda \boldsymbol{y}^* \boldsymbol{x}.$$

Thus, $(\lambda - \mu)\boldsymbol{y}^* \boldsymbol{x} = 0$. As $\lambda \neq \mu$, we get $\boldsymbol{y}^* \boldsymbol{x} = 0$, i.e., $\boldsymbol{x}$ and $\boldsymbol{y}$ is orthogonal. □

Question 4.4. *Suppose A has distinct eigenvalues. Let $f(z) = \sum_{i=-\infty}^{\infty} a_i z^i$ be a function which is defined at the eigenvalues of A. Let $Q^* AQ = T$ be the Schur form of A (so Q is unitary and T upper triangular).*

1. *show that $f(A) = Qf(T)Q^*$. Thus to compute $f(A)$ it suffices to be able to compute $f(T)$. In the rest of the problem you will derive a simple recurrence formula for $f(T)$.*

2. *Show that $(f(T))_{ii} = f(T_{ii})$ so that the diagonal of $f(T)$ can be computed from the diagonal of T.*

3. *Show that $Tf(T) = f(T)T$.*

4. *From the last result, show that the ith superdiagonal of $f(T)$ can be computed from the (i-1)st and earlier subdiagonals. Thus, starting at the diagonal of $f(T)$, we can compute the first superdiagonal, second superdiagonal, and so on.*

Proof.

1. It is suffices to prove that $A^i = QT^iQ^*$, and it can be proved as

$$A^i = QTQ^*QTQ^* \dots QTQ^* = QT^iQ^*.$$

   Thus,

$$f(A) = \sum_{i=-\infty}^{\infty} a_i A^i = \sum_{i=-\infty}^{\infty} a_i QT^iQ^* = Qf(T)Q^*.$$

2. It is suffices to prove that $(T^i)_{ii} = (T_{ii})^i$, and it can be proved by mathematic induction. As the multiplication of upper triangular is also an upper triangular matrix, the diagonal entries are the products of the corresponding diagonal entries. For the base case, we have $(T^2)_{ii} = T_{ii}^2$. Then, suppose it is correct for k, we can deduce for $k+1$ as

$$(T^{k+1})_{ii} = (T^kT)_{ii} = T_{ii}^kT_{ii} = T_{ii}^{k+1}.$$

   Thus,

$$(f(T))_{jj} = \sum_{i=-\infty}^{\infty} a_i(T^i)_{jj} = \sum_{i=-\infty}^{\infty} a_i T_{jj}^i = f(T_{jj}).$$

3. It is self evident that any matrix is commutative with itself. Thus

$$Tf(T) = \sum_{i=-\infty}^{\infty} a_i TT^i = \sum_{i=-\infty}^{\infty} a_i T^iT = f(T)T.$$

4. It is suffices to deduce the result for T^i, and it can be prove by mathematic induction. First, for T^2, we have

$$T^2(i, i+k) = \sum_{t=i}^{i+k} T(i,t)T(t,i+k).$$

   Thus, the kth superdiagonal can be computed from 0th to $(k-1)$th superdiagonals. Then, suppose it is correct for T^n, so we can express $T^n(i,j) = \sum_{t=1}^{j-i+1} c_{i,j,t}T(i,i+t-1)$, where $c_{i,j,t}$ are consts. Therefore, we can prove for T^{n+1} as

$$T^{k+1}(i,i+k) = \sum_{s=i}^{i+k} T^k(i,s)T(s,i+k) = \sum_{s=i}^{i+k} T(s,i+k) \sum_{t=1}^{s-i+1} c_{i,s,t}T(i,i+t-1).$$

   Thus, the kth superdiagonal of T^{n+1} can still be computed from 0th to $(k-1)$th superdiagonals.

□

Question 4.5. *Let A be a square matrix. Apply either Question 4.4 to the Schur form of A or equation (4.6) to the Jordan form of A to conclude that the eigenvalues of $f(A)$ are $f(\lambda_i)$, where the λ_i are the eigenvalues of A. This result is called the* ***spectral mapping theorem****.*

Proof. Let UTU^* be the Schur form of A. According to part II of the previous question, we know that the eigenvalues of $f(T)$ and $f(A)$ are the same. Since $(f(T))_{ii} = f(T_{ii}) = f(\lambda_i)$, the eigenvalues of $f(T)$ are $f(\lambda_i)$. Thus, the eigenvalues of $f(A)$ are $f(\lambda_i)$ with the same multiplicity as λ_i. □

Question 4.6. *In this problem we will show how to solve the* ***Sylvester*** *or* ***Lyapunov*** *equation $AX - XB = C$, where X and C are m-by-n, A is m-by-m, and B is n-by-n. This is a system of mn linear equations for the entries of X.*

1. *Given the Schur decompositions of A and B, show how $AX - XB = C$ can be transformed into a similar system $A'Y - YB' = C'$, where A' and B' are upper triangular.*

2. *Show how to solve for the entries of Y one at a time by a process analogous to back substitution. what condition on the eigenvalues of A and B guarantees that the system of equations is nonsingular?*

3. *Show how to transform Y to get the solution X.*

Solution.

1. Suppose $U_1T_1U_1^*$ and $U_2T_2U_2^*$ are the Schur form of A and B repectively. Then the equation can be expressed as

$$T_1U_1^*XU_2 - U_1^*XU_2T_2 = U_1^*CU_2.$$

Denote $Y = U_1^*XU_2$ and $C' = U_1^*CU_2$, we have

$$T_1Y - YT_2 = C',$$

which satisfies the condition.

2. For illustration, we rewrite the left side of equation $A'Y - YB' = C'$ in matrix form as

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{bmatrix} - \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ & b_{22} & \dots & b_{2n} \\ & & \ddots & \vdots \\ & & & b_{nn} \end{bmatrix}.$$

Equating its corresponding entry of $C'(n,:)$ one by one, we have

$$\begin{aligned} C'(n,1) &= a_{nn}y_{n1} - b_{11}y_{n1}, \\ C'(n,2) &= a_{nn}y_{n2} - \sum_{i=1}^{2} b_{i2}y_{ni}, \\ &\dots\dots \\ C'(n,k) &= a_{nn}y_{nk} - \sum_{i=1}^{k} b_{ik}y_{ni}, \\ &\dots\dots \\ C'(n,n) &= a_{nn}y_{nk} - \sum_{i=1}^{n} b_{in}y_{ni}. \end{aligned}$$

Thus, if $a_{nn} \neq b_{ii}$, we can solve the last row of Y one at a time. Then continue like this with the requisition that all the eigenvalues of A and B do not coincide. We can solve Y.

3. As what we have deduced, $Y = U_1^* X U_2$, where U_1, U_2 are the unitary matrices which transform A and B into Schur form separately. We can get

$$X = U_1 Y U_2^*.$$

□

Question 4.7. *Suppose that* $T = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$ *is in Schur form. We want to find a matrix S so that* $S^{-1}TS = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$. *It turns out we can choose S of the form* $\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}$. *Show how to solve for R.*

Solution. As the question does not tell us the dimension of each block and matrix, we can suppose that the matrix multiplication which the question requires is compatible. Since $S^{-1} = \begin{bmatrix} I & -R \\ 0 & I \end{bmatrix}$, we have

$$S^{-1}TS = \begin{bmatrix} I & -R \\ 0 & I \end{bmatrix} \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} \begin{bmatrix} I & R \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & AR + C - RB \\ 0 & B \end{bmatrix}$$

Thus, the question is equal to solve

$$AR - RB = -C,$$

which is the type of equation discussed in the previous question. □

Question 4.8. *Let A be m-by-n and B be n-by-m. show that the matrices*

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$$

are similar. conclude that the nonzero eigenvalues of AB are the same as those of BA.

Proof. Consider the nonsingular matrix $S = \begin{bmatrix} I & -A \\ 0 & I \end{bmatrix}$. We have

$$S \begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} S^{-1} = \begin{bmatrix} I & -A \\ 0 & I \end{bmatrix} \begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix}.$$

Thus, $\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$ are similar. Consequently, they have the same eigenvalues, and according to Question 4.1., we have

$$\lambda^n \det(\lambda I - AB) = \lambda^m \det(\lambda I - BA)$$

Thus, AB and BA have the same nonzero eigenvalues. □

Question 4.9. *Let A be n-by-n with eigenvalues $\lambda_1,\ldots,\lambda_n$. Show that*

$$\sum_{i=1}^{n} |\lambda_i|^2 = \min_{\det(S)\neq 0} \|S^{-1}AS\|_F^2.$$

Remark. The question is not correct unless we change the operation from "min" into "inf". I present a counterexample below to illustrate why. Consider A as

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

For any nonsingular matrix $S = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we have $ad \neq bc$ and

$$\begin{aligned}\|S^{-1}AS\|_F^2 &= \frac{1}{(ad-bc)^2} \left\| \begin{bmatrix} ad-ac-bc & a^2 \\ -c^2 & ac+ad-bc \end{bmatrix} \right\|_F^2 \\ &= \frac{c^4+a^4}{(ad-bc)^2} + 2.\end{aligned}$$

Since a and c cannot equals zero simultaneously, we have

$$\|S^{-1}AS\|_F^2 > \sum_{i=1}^{n} |\lambda_i|^2.$$

Thus, the lower bound cannot be reached in this situation.

Proof. First, denote the Schur form of A as $U^*AU = T$. As there is an one to one coresspondent from S to US, it means that the set is not changed. Thus, we we can transform the minimum problem into

$$\inf_{\det(US)\neq 0} \|S^{-1}U^*AUS\|_F^2 = \inf_{\det(S)\neq 0} \|S^{-1}TS\|_F^2.$$

Then, we can transform the set from nonsingular matrices to nonsingular upper triangular matrices, because every nonsingular matrix S^{-1} has a QR decomposition as $S^{-1} = QR$. What's more, since unitary matrices do not change Frobenius norm, we have

$$\inf_{\det(S)\neq 0} \|S^{-1}TS\|_F^2 = \inf_{\det(QR)\neq 0} \|QRTR^{-1}Q^{-1}\|_F^2 = \inf_{\det(R)\neq 0} \|RTR^{-1}\|_F^2.$$

As the inverse of an upper triangular matrix is also upper triangular, we have

$$RTR^{-1} = \begin{bmatrix} r_{11} & & & * \\ & r_{22} & & \\ & & \ddots & \\ & & & r_{nn} \end{bmatrix} \begin{bmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} r_{11}^{-1} & & & * \\ & r_{22}^{-1} & & \\ & & \ddots & \\ & & & r_{nn}^{-1} \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

Thus,

$$\sum_{i=1}^{n} |\lambda_i|^2 \leq \inf_{\det(R)\neq 0} \|RTR^{-1}\|_F^2 = \inf_{\det(S)\neq 0} \|S^{-1}AS\|_F^2.$$

Then, we should that the lower bound can be approximated to any precision. Denote $c = \max_{i\neq j}|T(i,j)|$. For any $\epsilon < 1$, set $a = \epsilon/\max(c,1)$. Consider $S = \operatorname{diag}\big(a^{n-1}, a^{n-2}, \dots, a, 1\big)$

$$STS^T = \begin{bmatrix} \lambda_1 & aT_{12} & a^2T_{13} & \dots & a^{n-1}T_{1n} \\ & \lambda_2 & aT_{23} & \dots & a^{n-2}T_{2n} \\ & & \lambda_3 & \dots & a^{n-3}T_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix}.$$

Since $|a^iT_{kj}| < \epsilon$, the square sum of the off-diagnoal entries is smaller than $\frac{n(n-1)}{2}\epsilon$, which proves the question. □

Question 4.10. *Let A be an n-by-n matrix with eigenvalues $\lambda_1, \dots, \lambda_n$.*

1. *Show that A can be written $A = H + S$, where $H = H^*$ is Hermitian and $S = -S^*$ is skew-Hermitian. Give explicit formulas for H and S in terms of A.*
2. *Show that $\sum_{i=1}^n |\Re\lambda_i|^2 \leq \|H\|_F^2$.*
3. *Show that $\sum_{i=1}^n |\Im\lambda_i|^2 \leq \|S\|_F^2$.*
4. *Show that A is normal ($AA^* = A^*A$) if and only if $\sum_{i=1}^n |\lambda_i|^2 = \|A\|_F^2$.*

Solution.

1. Suppose $A = H + S$, then we have $A^* = H - S$. Thus, we can solve H and S as

$$H = \frac{1}{2}(A + A^*),$$
$$S = \frac{1}{2}(A - A^*),$$

   where $A = H + S$, H is Hermitian and S is skew-Hermitian.

2. Let $U^*AU = T$ be the Schur form of A. Then, for H, we have

$$\|H\|_F^2 = \|U^*HU\|_F^2 = \|\frac{1}{2}(T + T^*)\|_F^2 = \left\| \begin{bmatrix} \Re\lambda_1 & & & * \\ & \Re\lambda_2 & & \\ & * & \ddots & \\ & & & \Re\lambda_n \end{bmatrix} \right\|_F^2 \geq \sum_{i=1}^n |\Re\lambda_i|^2.$$

3. Again, let $U^*AU = T$ be the Schur form of A. Then, for S, we have

$$\|S\|_F^2 = \|U^*SU\|_F^2 = \|\frac{1}{2}(T - T^*)\|_F^2 = \left\| \begin{bmatrix} \Im\lambda_1 & & & * \\ & \Im\lambda_2 & & \\ & * & \ddots & \\ & & & \Im\lambda_n \end{bmatrix} \right\|_F^2 \geq \sum_{i=1}^n |\Im\lambda_i|^2.$$

4. If A is normal, according to Question 4.2., there exists unitary matrix U such that $UAU^* = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Thus, we have

$$\|A\|_F^2 = \|UAU^*\|_F^2 = \sum_{i=1}^n |\lambda_i|^2.$$

On the other hand, if A is not normal, denote its Schur form as $UAU^* = T$. We can claim that T is not diagonal, otherwise T will be normal and consequently A is normal, which contradicts our assumption. So, we get

$$\|A\|_F^2 = \|UAU^*\|_F^2 = \|T\|_F^2 > \sum_{i=1}^n \lambda_i.$$

Thus, if $\sum_{i=1}^n |\lambda_i|^2 = \|A\|_F^2$, A must be normal.

Question 4.11. *Let λ be a simple eigenvalue, and let $\boldsymbol{x}$ and $\boldsymbol{y}$ be right and left eigenvectors. We define the **spectral projection** P corresponding to λ as $P = \boldsymbol{x}\boldsymbol{y}^*/(\boldsymbol{y}^*\boldsymbol{x})$. Prove that P has the following properties.*

1. *P is uniquely defined, even though we could use any nonzero scalar multiples of $\boldsymbol{x}$ and $\boldsymbol{y}$ in its definition.*

2. *$P^2 = P$. (Any matrix satisfying $P^2 = P$ is called a **projection matrix**.)*

3. *$AP = PA = \lambda P$. (These properties motivate the name **spectral projection**, since P "contains" the left and right invariant subspaces of λ.)*

4. *$\|P\|_2$ is the condition number of λ.*

Remark. Part IV of the question follows directly from Question 1.7. However, I present a similar proof as shown below to remain us of the result.
Proof.

1. Since λ is a simple eigenvalue, any other right and left eigenvectors can be expressed as $\hat{\boldsymbol{x}} = c\boldsymbol{x}$ and $\hat{\boldsymbol{y}} = d\boldsymbol{y}$. Thus, for the spectral projection defined by $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, we have

$$\hat{P} = \hat{\boldsymbol{x}}\hat{\boldsymbol{y}}^*/(\hat{\boldsymbol{y}}^*\hat{\boldsymbol{x}}) = cd\boldsymbol{x}\boldsymbol{y}^*/(cd\boldsymbol{y}^*\boldsymbol{x}) = P,$$

which suggests that P is uniquely defined.

2. We can verify $P^2 = P$ as follows:

$$P^2 = \boldsymbol{x}\boldsymbol{y}^*\boldsymbol{x}\boldsymbol{y}^*/(\boldsymbol{y}^*\boldsymbol{x})^2 = \boldsymbol{x}\boldsymbol{y}^*/(\boldsymbol{y}^*\boldsymbol{x}) = P.$$

3. We can verify $AP = PA = \lambda P$ as follows:

$$AP = A\boldsymbol{x}\boldsymbol{y}^*/(\boldsymbol{y}^*\boldsymbol{x}) = \lambda\boldsymbol{x}\boldsymbol{y}^*/(\boldsymbol{y}^*\boldsymbol{x}) = \lambda P = \boldsymbol{x}\boldsymbol{y}^*A/(\boldsymbol{y}^*\boldsymbol{x}) = PA.$$

4. According to the result of part I, we can assume that $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$. For any nonzero vector $\boldsymbol{\alpha}$ in $\mathbb{C}^n$, we can decompose it into $\boldsymbol{\alpha} = c\boldsymbol{y} + d\boldsymbol{z}$, where $\boldsymbol{y}^*\boldsymbol{z} = 0$ and $\|\boldsymbol{z}\|_2 = 1$. Then we have

$$\|P\|_2 = \max_{\boldsymbol{\alpha}\neq 0} \frac{\|P\boldsymbol{\alpha}\|_2}{\|\boldsymbol{\alpha}\|_2} = \max_{\boldsymbol{\alpha}\neq 0} \frac{\|P(\boldsymbol{y}+\boldsymbol{z})\|_2}{\|\boldsymbol{\alpha}\|_2} = \max_{\boldsymbol{\alpha}\neq 0} \frac{|c|}{\|\boldsymbol{\alpha}\|_2 |\boldsymbol{y}^*\boldsymbol{x}|} = \max_{|c|+|d|\neq 0} \frac{|c|}{|\boldsymbol{y}^*\boldsymbol{x}|\sqrt{|c|+|d|}}.$$

It is easy to find that $c = 0$ cannot be the maximizer. Thus, we can divide $|c|$ on the fraction and have

$$\|P\|_2 = \max_{|c|+|d|\neq 0} \frac{|c|}{|\boldsymbol{y}^*\boldsymbol{x}|\sqrt{|c|+|d|}} = \max_{|c|\neq 0} \frac{1}{|\boldsymbol{y}^*\boldsymbol{x}|\sqrt{1+\frac{|d|}{|c|}}} \leq \frac{1}{|\boldsymbol{y}^*\boldsymbol{x}|}.$$

The maximum is obtained when $d = 0$, and consequently $\|P\|_2$ is the condition number of λ.

□

Question 4.12. *Let* $A = \begin{bmatrix} a & c \\ 0 & b \end{bmatrix}$. *Show that the condition numbers of the eigenvalues of A are both equal to* $(1+\left(\frac{c}{a-b}\right)^2)^{1/2}$. *Thus, the condition number is large if the difference* $a-b$ *between the eigenvalues is small compared to c, the offdiagonal part of the matrix.*

Proof. As what the question suggests, we assume that all the variables appearing are real. It is easy to verify that

$$\boldsymbol{x}_1 = [1,0]^T,$$
$$\boldsymbol{y}_1 = [1, \tfrac{c}{a-b}]^T$$

are the right and left eigenvectors of eigenvalue $\lambda_1 = a$, and

$$\boldsymbol{x}_2 = [1, \tfrac{a-b}{c}]^T,$$
$$\boldsymbol{y}_2 = [0,1]^T$$

are the right and left eigenvectors of eigenvalue $\lambda_2 = b$. Thus, the condition number of λ_1 is

$$\frac{\|\boldsymbol{x}_1\|_2\|\boldsymbol{y}_1\|_2}{|\boldsymbol{y}_1^T\boldsymbol{x}_1|} = (1+\left(\frac{c}{a-b}\right)^2)^{1/2},$$

and that of λ_2 is

$$\frac{\|\boldsymbol{x}_2\|_2\|\boldsymbol{y}_2\|_2}{|\boldsymbol{y}_2^T\boldsymbol{x}_2|} = (1+\left(\frac{c}{a-b}\right)^2)^{1/2}.$$

Thus, the result is proved. □

Question 4.13. *Let A be a matrix,* $\boldsymbol{x}$ *be a unit vector* ($\|\boldsymbol{x}\|_2 = 1$)*,* μ *be a scalar, and* $\boldsymbol{r} = A\boldsymbol{x} - \mu\boldsymbol{x}$. *Show that there is a matrix E with* $\|E\|_F = \|\boldsymbol{r}\|_2$ *such that* $A+E$ *has eigenvalue* μ *and eigenvector* $\boldsymbol{x}$.

Proof. Suppose such E exists. Since we have the condition

$$(A+E)\boldsymbol{x} = \mu \boldsymbol{x},$$

we have the equation for E as

$$-\boldsymbol{r} = E\boldsymbol{x}.$$

Set $E = -\boldsymbol{r}\boldsymbol{x}^*$, which satisfies the condition. What's more, we can evaluate the Frobenius norm of E as

$$\|E\|_F^2 = \mathrm{tr}(\boldsymbol{x}\boldsymbol{r}^*\boldsymbol{r}\boldsymbol{x}^*) = \|\boldsymbol{r}\|_2^2 \mathrm{tr}(\boldsymbol{x}\boldsymbol{x}^*) = \|\boldsymbol{r}\|_2^2.$$

So, there is a matrix E satisfies the conditions of the question. □

5 Solutions for Chapter V: The Symmetric Eigenproblem and Singular Value Decomposition

Question 5.1. *Show that $A = B + iC$ is hermitian if and only if*

$$M = \begin{bmatrix} B & -C \\ C & B \end{bmatrix}$$

is symmetric. Express the eigenvalues and eigenvectors of M in terms of those of A.

Remark. I believe the question is correct under the assumption of both B and C are real. Thus, it is what my proof will assume.

Solution. Denote a_{ij}^* asthe (i,j) entry of matrix A^*. As B and C are both real and

$$a_{ij}^* = \overline{a_{ji}} = \overline{b_{ji} + \boldsymbol{i}c_{ji}} = \overline{b_{ji}} - \boldsymbol{i}\overline{c_{ji}} = b_{ji} - \boldsymbol{i}c_{ji}.$$

A is hermitian if and only if $A^* = B - \boldsymbol{i}C = B + \boldsymbol{i}C = A$, which is equivalent to the symmetry of M.

As for eigenvalues and eigenvectors, suppose $\alpha = \beta + \boldsymbol{i}\gamma$ is the eigenvector with corresponding eigenvalue $\lambda \in \mathbb{R}$ of A. Because of the definition of eigenvalues and eigenvectors, $A\alpha = (B + \boldsymbol{i}C)(\beta + \boldsymbol{i}\gamma) = \lambda(\beta + \boldsymbol{i}\gamma)$. Thus, we have the equation

$$\begin{aligned} B\beta - C\gamma &= \lambda\beta, \\ B\gamma + C\beta &= \lambda\gamma. \end{aligned}$$

Then, considering

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} B\beta - C\gamma \\ C\beta + B\gamma \end{bmatrix} = \begin{bmatrix} \lambda\beta \\ \lambda\gamma \end{bmatrix} = \lambda \begin{bmatrix} \beta \\ \gamma \end{bmatrix},$$

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} -\gamma \\ \beta \end{bmatrix} = \begin{bmatrix} -B\gamma - C\beta \\ -C\gamma + B\beta \end{bmatrix} = \begin{bmatrix} -\lambda\gamma \\ \lambda\beta \end{bmatrix} = \lambda \begin{bmatrix} -\gamma \\ \beta \end{bmatrix}.$$

Thus, as we can see, from the above equation, we know that $[\beta, \gamma]^T, [-\gamma, \beta]^T$ are the eigenvectors with corresponding eigenvalue λ of M. □

Question 5.2. *Prove Corollary 5.1, using Weyl's theorem and part 4 of Theorem 3.3.*

Proof. According to the hint provided, denote matrices H_1, H_2, H as

$$H_1 = \begin{bmatrix} 0 & G^T \\ G & 0 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 0 & G^T + F^T \\ G+F & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & F^T \\ F & 0 \end{bmatrix}.$$

We know that $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_n \geq -\sigma_n \geq -\sigma_{n-1} \geq \ldots \geq -\sigma_1$ are the eigenvalues of H_1, and $\sigma'_1 \geq \sigma'_2 \ldots \geq \sigma'_n \geq -\sigma'_n \geq -\sigma'_{n-1} \geq \ldots \geq -\sigma'_1$ are the eigenvalues of H_2. Using Weyl's theorem, we have

$$|\sigma_i - \sigma'_i| \leq \|H\|_2.$$

Thus, what remains to prove is $\|H\|_2 = \|F\|_2$. To do this, consider

$$H^T H = \begin{bmatrix} 0 & F^T \\ F & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & F^T \\ F & 0 \end{bmatrix} = \begin{bmatrix} F^T F & 0 \\ 0 & FF^T \end{bmatrix}.$$

This matrix has the eigenvalues as FF^T and $F^T F$, which have the same nonzero eigenvalues. Thus $\lambda_{\max}(H^T H) = \lambda_{\max}(F^T F)$. i.e., $\|H\|_2 = \|F\|_2$. □

Question 5.3. *Consider Figure 5.1. Consider the corresponding contour plot for an arbitrary* 3*-by-*3 *matrix A with eigenvalues $\alpha_3 \leq \alpha_2 \leq \alpha_1$. Let C_1 and C_2 be the two great cirles along which $\rho(\boldsymbol{u}, A) = \alpha_2$. At what angle do they intersect?*

Remark. Since A is arbitary, we may assume A is real. Otherwise, $\boldsymbol{x}^* A\boldsymbol{x}$ may not be real number. What's more, assume A is diagnoalizable and the eigenvalues of A are not identical. If not, I don't think it will behave the same as what the textbook describes.

Solution. Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ are three unit orthogonal eigenvectors of A with corresponding eigenvalues as $\lambda_1, \lambda_2, \lambda_3$ repectively. Thus, all vectors $\boldsymbol{y} \in \mathbb{R}^3$ can be decomposed as $\boldsymbol{y} = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + b_3\boldsymbol{x}_3$. If a unit vector $\boldsymbol{y}$ satisfies $\rho(\boldsymbol{y}, A) = \alpha_2$, we have

$$\rho(\boldsymbol{y}, A) = \lambda_1 b_1^2 + \lambda_2 b_2^2 + \lambda_3 b_3^2 = \lambda_2.$$

Rearranging the above equation and using $b_1^2 + b_2^2 + b_3^2 = 1$ yields

$$\sqrt{\lambda_1 - \lambda_2}b_1 = \pm\sqrt{\lambda_2 - \lambda_3}b_3.$$

As we have assume not all eigenvalues are equal, we can suppose $\lambda_2 - \lambda_3 \neq 0$ and denote $\sqrt{\lambda_1 - \lambda_2}/\sqrt{\lambda_2 - \lambda_3} = c$ without loss of generality. Thus, we express the two great 'circles' by parametric equations as

$$\begin{aligned} b_1 &= \sin(\theta)/\sqrt{1+c^2}, \\ b_2 &= \cos(\theta), \\ b_3 &= \pm c\sin(\theta)/\sqrt{1+c^2}. \end{aligned}$$

Then, their insection points are the points with $\theta = 0$ and $\theta = \pi$. For $\theta = 0$, their tangent vectors are

$$\boldsymbol{t}_1 = 1/\sqrt{1+c^2}\boldsymbol{x}_1 + c/\sqrt{1+c^2}\boldsymbol{x}_3,$$
$$\boldsymbol{t}_2 = 1/\sqrt{1+c^2}\boldsymbol{x}_1 - c/\sqrt{1+c^2}\boldsymbol{x}_3.$$

Consequently, the corresponding intersection angle is $\arccos(\boldsymbol{t}_1^T\boldsymbol{t}_2/(\|\boldsymbol{t}_1\|_2\cdot\|\boldsymbol{t}_2\|_2)) = \arccos(\frac{1-c^2}{1+c^2})$.For $\theta = \pi$, by the same computation, the intersection angle is the same as $\theta = 0$. □

Question 5.4. *Use the Courant-Fischer minimax theorem (Theorem 5.2) to prove the* ***Cauchy interlace theorem****:*

1. *Suppose that $A = \begin{pmatrix} H & \boldsymbol{b} \\ \boldsymbol{b}^T & u \end{pmatrix}$ is an n-by-n symmetric matrix and H is $(n-1)$-by-$(n-1)$. Let $\alpha_n \le \ldots \le \alpha_1$ be the eigenvalues of A and $\theta_{n-1} \le \ldots \le \theta_1$ be the eigenvalues of H. Show that these two sets of eigenvalues* ***interlace****:*

$$\alpha_n \le \theta_{n-1} \le \ldots \le \theta_i \le \alpha_i \le \theta_{i-1} \le \alpha_{i-1} \le \ldots \le \theta_1 \le \alpha_1.$$

2. *Let $A = \begin{pmatrix} H & B \\ B^T & U \end{pmatrix}$ be n-by-n and H be m-by-m, with eigenvalues $\theta_m \le \ldots \le \theta_1$. Show that the eigenvalues of A and H interlace in the sense that $\alpha_{j+(n-m)} \le \theta_j \le \alpha_j$ (or equivalently $\alpha_j \le \theta_{j-(n-m)} \le \alpha_{j-(n-m)}$).*

Proof.

1. Denote the n-by-$(n-1)$ matrix P as $P = \begin{bmatrix} I_{n-1} \\ \boldsymbol{0} \end{bmatrix}$. Thus, $H = P^TAP$ and for any vector $\boldsymbol{x} \in \mathbb{R}^{n-1}$, $\boldsymbol{x}^T\boldsymbol{x} = (P\boldsymbol{x})^T(P\boldsymbol{x})$ holds because $P^TP = I_{n-1}$. Noting that for linear independent vectors $\boldsymbol{x}_1,\ldots.\boldsymbol{x}_{n-1} \in \mathbb{R}^{n-1}$, $P\boldsymbol{x}_1,\ldots.P\boldsymbol{x}_{n-1} \in \mathbb{R}^n$ are also linear independent. Therefore, if S is i-dimensional space in $\mathbb{R}^{n-1}$, PS is i-dimensional space in $\mathbb{R}^n$. Since the minimum of the subset is not less than the minimum of whole set, we have

$$\begin{aligned} \theta_i &= \min_{S^{n-i}} \max_{\boldsymbol{0}\neq\boldsymbol{x}\in S^{n-i}} \frac{\boldsymbol{x}^TH\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \\ &= \min_{S^{n-i}} \max_{\boldsymbol{0}\neq\boldsymbol{x}\in S^{n-i}} \frac{(P\boldsymbol{x})^TA(P\boldsymbol{x})}{(P\boldsymbol{x})^T(P\boldsymbol{x})} \\ &\ge \min_{S^{n-i}} \max_{\boldsymbol{0}\neq\boldsymbol{y}\in S^{n-i}} \frac{\boldsymbol{y}^TA\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \alpha_{i+1}. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \theta_i &= \max_{S^i} \min_{\boldsymbol{0}\neq\boldsymbol{x}\in S^i} \frac{\boldsymbol{x}^TH\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \\ &= \max_{S^i} \min_{\boldsymbol{0}\neq\boldsymbol{x}\in S^i} \frac{(P\boldsymbol{x})^TA(P\boldsymbol{x})}{(P\boldsymbol{x})^T(P\boldsymbol{x})} \\ &\le \max_{S^i} \min_{\boldsymbol{0}\neq\boldsymbol{y}\in S^i} \frac{\boldsymbol{y}^TA\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \alpha_i. \end{aligned}$$

Combine the results, and we have $\alpha_{i+1} \le \theta_i \le \alpha_i$, which is equivalent to what we should prove.

2. It is similar to what we have proved. Denote the n-by-m matrix P as $P = \begin{bmatrix} I_m \\ \mathbf{0} \end{bmatrix}$, and we have $H = P^T AP$. Using Courant-Fischer minimax theorem, we have

$$\begin{aligned} \theta_i &= \min_{S^{m+1-i}} \max_{\mathbf{0}\neq \boldsymbol{x}\in S^{m+1-i}} \frac{\boldsymbol{x}^T H\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \\ &= \min_{S^{m+1-i}} \max_{\mathbf{0}\neq \boldsymbol{x}\in S^{m+1-i}} \frac{(P\boldsymbol{x})^T A(P\boldsymbol{x})}{(P\boldsymbol{x})^T(P\boldsymbol{x})} \\ &\geq \min_{S^{m+1-i}} \max_{\mathbf{0}\neq \boldsymbol{y}\in S^{m+1-i}} \frac{\boldsymbol{y}^T A\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \alpha_{i+n-m}. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \theta_i &= \max_{S^i} \min_{\mathbf{0}\neq \boldsymbol{x}\in S^i} \frac{\boldsymbol{x}^T H\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \\ &= \max_{S^i} \min_{\mathbf{0}\neq \boldsymbol{x}\in S^i} \frac{(P\boldsymbol{x})^T A(P\boldsymbol{x})}{(P\boldsymbol{x})^T(P\boldsymbol{x})} \\ &\leq \max_{S^i} \min_{\mathbf{0}\neq \boldsymbol{y}\in S^i} \frac{\boldsymbol{y}^T A\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \alpha_i. \end{aligned}$$

Combine the results, and we have $\alpha_{i+n-m} \leq \theta_i \leq \alpha_i$, which is equivalent to what we should prove.

□

Question 5.5. *Let $A = A^T$ with eigenvalues $\alpha_1 \geq \ldots \geq \alpha_n$. Let $H = H^T$ with eigenvalues $\theta_1 \geq \ldots \geq \theta_n$. Let $A + H$ have eigenvalues $\lambda_1 \geq \ldots \lambda_n$. Use Courant-Fischer minimax theorem (Theorem) to show that $\alpha_j + \theta_n \leq \lambda_j \leq \alpha_j + \theta_1$. If H is positive definite, conclude that $\lambda_j > \alpha_j$. In other words, adding a symmetric positive definite matrix H to another symmetric matrix A can only increase its eigenvalues.*

Proof. According to Courant-Fischer minimax theorem, we have

$$\begin{aligned} \lambda_i &= \min_{S^{n+1-i}} \max_{\mathbf{0}\neq \boldsymbol{x}\in S^{n+1-i}} \frac{\boldsymbol{x}^T (A+H)\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \\ &\leq \min_{S^{n+1-i}} (\max_{\mathbf{0}\neq \boldsymbol{x}\in S^{n+1-i}} \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} + \max_{\mathbf{0}\neq \boldsymbol{x}\in S^{n+1-i}} \frac{\boldsymbol{x}^T H\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}) \\ &\leq \min_{S^{n+1-i}} (\max_{\mathbf{0}\neq \boldsymbol{x}\in S^{n+1-i}} \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} + \theta_1) = \alpha_i + \theta_1. \end{aligned}$$

And,

$$\begin{aligned} \lambda_i &= \max_{S^i} \min_{\mathbf{0}\neq \boldsymbol{x}\in S^i} \frac{\boldsymbol{x}^T (A+H)\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \\ &\geq \max_{S^i}(\min_{\mathbf{0}\neq \boldsymbol{x}\in S^i} \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} + \min_{\mathbf{0}\neq \boldsymbol{x}\in S^i} \frac{\boldsymbol{x}^T H\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}) \\ &\geq \max_{S^i}(\min_{\mathbf{0}\neq \boldsymbol{x}\in S^i} \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} + \theta_n) = \alpha_i + \theta_n. \end{aligned}$$

Combine the results, and we have $a_j + \theta_n \leq \lambda_j \leq \alpha_j + \theta_1$. As for positive definite H, we know that all its eigenvalues are positive. Thus,

$$\alpha_j < \alpha_j + \theta_n \leq \lambda_j.$$

□

Question 5.6. *Let $A = [A_1, A_2]$ be n-by-n, where A_1 is n-by-m and A_2 is n-by-$(n-m)$. Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ be the singular values of A and $\tau_1 \geq \tau_2 \ldots \geq \tau_m$ be the singular values of A_1. Use the Cauchy interlace theorem from Question 5.4 and part 4 of Theorem 3.3 to prove that $\sigma_j \geq \tau_j \geq \sigma_{j+n-m}$.*

Proof. Denote the matrix H as

$$H = \begin{bmatrix} 0 & A_1 & A_2 \\ A_1^T & 0 & 0 \\ A_2^T & 0 & 0 \end{bmatrix}.$$

According to part 4 of Theorem 3.3, the absolute values of eigenvalues of H are the singular values of A^T, which are equal to those of A. Thus, we can express the eigenvalues of H as $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq -\sigma_n \geq \ldots \geq -\sigma_1$. Consider the $(n+m)$-by-$(n+m)$ principal submatrix

$$H_1 = \begin{bmatrix} 0 & A_1 \\ A_1^T & 0 \end{bmatrix}.$$

As in Question 3.14, we have already know its eigenvalues are $\tau_1 \geq \ldots \tau_m \geq 0 = \ldots = 0 \geq -\tau_m \geq \ldots \geq -\tau_1$. Thus, by Cauchy interlace theorem, we have

$$\sigma_{i+(n-m)} \leq \tau_i \leq \sigma_i,$$

which is what we need to prove. □

Question 5.7. *Let $\boldsymbol{q}$ be a unit vector and $\boldsymbol{d}$ be any vector orthogonal to $\boldsymbol{q}$. Show that*

$$\|(\boldsymbol{q}+\boldsymbol{d})\boldsymbol{q}^T - I\|_2 = \|\boldsymbol{q}+\boldsymbol{d}\|_2.$$

Remark. My proof suppose the dimension of the vector space is greater than 2. If the dimension is only 2, the proof is similar.

Proof. Denote $\boldsymbol{z}$ as a unit vector which is orthogonal to both $\boldsymbol{q}$ and $\boldsymbol{d}$. Therefore, any vector $\boldsymbol{x}$ can be express as $\boldsymbol{x} = a\boldsymbol{q} + b\boldsymbol{d} + c\boldsymbol{z}$, and we have

$$\begin{aligned} \|(\boldsymbol{q}+\boldsymbol{d})\boldsymbol{q}^T - I\|_2^2 &= \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\left((\boldsymbol{q}+\boldsymbol{d})\boldsymbol{q}^T - I\right)\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} \\ &= \max_{a^2+b^2+c^2\neq} \frac{(a^2-2ab+b^2)\|\boldsymbol{d}\|_2^2 + c^2}{a^2+b^2\|\boldsymbol{d}\|_2^2+c^2} \\ &= \max_{a^2+b^2+c^2\neq 0} \left(1 + \frac{(\|\boldsymbol{d}\|_2^2-1)a^2 - 2ab\|\boldsymbol{d}\|_2^2}{a^2+b^2\|\boldsymbol{d}\|_2^2+c^2}\right) \\ &\leq \max_{a^2+b^2\neq 0} \left(1 + \frac{(\|\boldsymbol{d}\|_2^2-1)a^2 - 2ab\|\boldsymbol{d}\|_2^2}{a^2+b^2\|\boldsymbol{d}\|_2^2}\right). \end{aligned}$$

To determine the maximum of the above formula, we need to consider whether $a = 0$. If $a = 0$, the above formula yields 1; If not, set $s = b/a, c = \|\boldsymbol{d}\|_2^2$, and define

$$f(s) = \frac{c-1-2cs}{1+cs^2}, \text{and } f'(s) = 2c^2 \frac{s^2-(1-\frac{1}{c})s-\frac{1}{c}}{(1+cs^2)^2}.$$

Thus, $f(s)$ has two local maximum at $s = -\frac{1}{c}, s = +\infty$. Since $f(-\frac{1}{c}) = \frac{c^2+c}{c+1}, f(+\infty) = 0$, we have

$$\|(\boldsymbol{q}+\boldsymbol{d})\boldsymbol{q}^T - I\|_2^2 = c+1 = \|\boldsymbol{q}+\boldsymbol{d}\|_2^2.$$

□

Question 5.8. *Formulate and prove a theorem for singular vectors analogous to Theorem 5.4.*

Remark. My theorem has been shown below, and I assume that the target matrix A is m-by-n, where $m \geq n$. However, I am not sure whether it is the one that the question requests.

Theorem 5.4'. Let $A = U\Sigma V^T$, which is the reduced SVD of an m-by-n matrix A. Let $A + E = \hat{A} = \hat{U}\hat{\Sigma}\hat{V}^T$ be the perturbed SVD. Write $U = [\boldsymbol{u}_1 \ldots, \boldsymbol{u}_n], V = [\boldsymbol{v}_1 \ldots, \boldsymbol{v}_n], \hat{U} = [\hat{\boldsymbol{u}}_1 \ldots, \hat{\boldsymbol{u}}_n]$, and $\hat{V} = [\hat{\boldsymbol{v}}_1 \ldots, \hat{\boldsymbol{v}}_n]$. Let θ, φ denote the acute angles between $\boldsymbol{u}_i, \hat{\boldsymbol{u}}_i$ and $\boldsymbol{v}_i, \hat{\boldsymbol{v}}_i$ repectively. Define gap(i, A) as $\min_{j\neq i} |\sigma_i - \sigma_j|$ if $i \neq n$; and gap(n, A) as $\min_{j\neq n}(\sigma_n, |\sigma_n - \sigma_j|)$ when $i = n$. Then we have

$$\max(\sin 2\theta, \sin 2\varphi) \leq \frac{2\sqrt{2}\|E\|_2}{\text{gap}(i, A)}.$$

Proof. Denote $\boldsymbol{\omega}_i = \frac{1}{\sqrt{2}}\begin{pmatrix}\boldsymbol{v}_i\\ \boldsymbol{u}_i\end{pmatrix}$, $\hat{\boldsymbol{\omega}}_i = \frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\boldsymbol{v}}_i\\ \hat{\boldsymbol{u}}_i\end{pmatrix}$, and $\hat{\theta}$ the acute angle between $\boldsymbol{\omega}_i$ and $\hat{\boldsymbol{\omega}}_i$. Then, consider the matrix

$$H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}, \quad \hat{H} = \begin{bmatrix} 0 & A^T + E^T \\ A+E & 0 \end{bmatrix},$$

and $\delta H = \hat{H} - H$. By Theorem 3.3 and Question 5.2, we know that $\|\delta H\|_2 = \|E\|_2$ and $\sigma_1 \geq \ldots \sigma_n \geq 0 \geq \ldots \geq 0 \geq -\sigma_n \geq \ldots \geq -\sigma_1$ are the eigenvalues of H, $\hat{\sigma}_1 \geq \ldots \hat{\sigma}_n \geq 0 \geq \ldots \geq 0 \geq -\hat{\sigma}_n \geq \ldots \geq -\hat{\sigma}_1$ are the eigenvalues of $\hat{H}$. Thus, According to and Theorem 5.4, we know that

$$\sin 2\hat{\theta} \leq \frac{2\|E\|_2}{\text{gap}(i, A)}.$$

As all angles are acute and

$$\cos\hat{\theta} = \boldsymbol{\omega}_i^T \hat{\boldsymbol{\omega}}_i = \frac{1}{2}(\boldsymbol{u}^T\hat{\boldsymbol{u}} + \boldsymbol{v}^T\hat{\boldsymbol{v}}) = \frac{1}{2}(\theta + \varphi),$$

we just prove for the angle, which is larger than $\hat{\theta}$. For the smaller angle, consider $\|\boldsymbol{u}_i + \hat{\boldsymbol{u}}_i\|_2^2$ instead of $\|\boldsymbol{u} - \hat{\boldsymbol{u}}_i\|_2$. Without loss of generality, we may assume that θ is the larger one. Since

$$2\sin^2\frac{\theta}{2} = 1 - \cos\theta_i = \frac{1}{2}\|\boldsymbol{u}_i - \hat{\boldsymbol{u}}_i\|_2^2 \leq \|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}_i\|_2^2 = 4\sin^2\frac{\hat{\theta}}{2},$$

and $\theta \geq \hat{\theta}$, we have

$$\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} \leq 2\sqrt{2}\sin\frac{\hat{\theta}}{2}\cos\frac{\hat{\theta}}{2} = \sqrt{2}\sin\hat{\theta}.$$

Then,

$$\sin 2\theta = 2\sin\theta\cos\theta \leq 2\sqrt{2}\sin\hat{\theta}\cos\hat{\theta} = \sqrt{2}\sin 2\hat{\theta} \leq \frac{2\|E\|_2}{\mathrm{gap}(i,A)}.$$

□

Question 5.9. *Prove bound (5.6) from Theorem 5.5.*

Proof. Since A is symmetric, it has a set of orthonormal vectors $[\boldsymbol{q}_1,\dots,\boldsymbol{q}_n]$, which span the whole space. Thus, suppose the unit vector $\boldsymbol{x} = \sum_{j=1}^n b_j\boldsymbol{q}_j$. Since

$$A\boldsymbol{x} - \rho(\boldsymbol{x},A)\boldsymbol{x} = \boldsymbol{r},$$
$$b_i A\boldsymbol{q}_i = b_i\alpha_i\boldsymbol{q}_i,$$

subtract the second from the first, and we have

$$(A-\rho(\boldsymbol{x},A)I)\sum_{j\neq i} b_j\boldsymbol{q}_j = \boldsymbol{r} + (\rho(\boldsymbol{x},A)-\alpha_i)b_i\boldsymbol{q}_i.$$

As we can see that the left side does not contain $\boldsymbol{q}_i$, the right side does not contain $\boldsymbol{q}_i$, either. Thus, the part of $\boldsymbol{q}_i$ of $\boldsymbol{r}$ cancelled by $(\rho(\boldsymbol{x},A)-\alpha_i)b_i\boldsymbol{q}_i$, and $\|\boldsymbol{r}+(\rho(\boldsymbol{x},A)-\alpha_i)b_i\boldsymbol{q}_i\|_2 \leq \|\boldsymbol{r}\|_2$. Denote the right side as $\sum_{j\neq i} c_j\boldsymbol{q}_j$. We have

$$(A-\rho(\boldsymbol{x},A)I)\sum_{j\neq i} b_j\boldsymbol{q}_j = (\alpha_i-\rho(\boldsymbol{x},A))\sum_{j\neq i} b_j\boldsymbol{q}_j = \sum_{j\neq i} c_j\boldsymbol{q}_j.$$

Thus, $b_j = c_j/(\alpha_i-\rho(\boldsymbol{x},A))$ as we have assume the gap$'$ is not zero. Then

$$\sin\theta = \|\sum_{j\neq i} b_j\boldsymbol{q}_j\|_2 = (\sum_{j\neq i}\left(c_j/(\alpha_i-\rho(\boldsymbol{x},A))\right)^2)^{\frac{1}{2}} \leq \frac{1}{\mathrm{gap}'}(\sum_{j\neq i} c_j^2)^{\frac{1}{2}} \leq \frac{\|\boldsymbol{r}\|_2}{\mathrm{gap}'}.$$

□

Question 5.10. *A harder question. Skipped.*

Question 5.11. *Suppose $\theta = \theta_1+\theta_2$, where all three angles lie between* 0 *and* $\pi/2$. *Prove that* $\frac{1}{2}\sin\theta \leq \frac{1}{2}\sin 2\theta_1 + \frac{1}{2}\sin 2\theta_2$.

Proof. Consider the function

$$f(\theta_1) = \sin 2\theta - \sin 2\theta_1 - \sin 2(\theta-\theta_1).$$

Its derivative is

$$f'(\theta_1) = -2\cos 2\theta_1 + 2\cos 2(\theta-\theta_1).$$

As $2\theta-2\theta_1, 2\theta_1 \in [0,\pi]$, between which the function cos is monotonus, $f'(\theta_1)$ has only one root as $\theta_1 = \theta/2$. Thus, $f(\theta_1)$ decreases in $[0,\theta/2]$, and increases in $[\theta/2,\theta]$. It can only reach its maximum at the extremity $\theta_1 = 0$ and $\theta_1 = \theta$. i.e.,

$$f(\theta_1) \leq \max(f(0), f(\theta)) = 0.$$

Therefore, $\frac{1}{2}\sin\theta \leq \frac{1}{2}\sin 2\theta_1 + \frac{1}{2}\sin 2\theta_2$. □

Question 5.12. *Prove Corollary 5.2.*

Proof. Denote the matrix H as $H = \begin{bmatrix} 0 & G^T \\ G & 0 \end{bmatrix}$, and $\hat{H}$ as $\hat{H} = \begin{bmatrix} 0 & \hat{G}^T \\ \hat{G} & 0 \end{bmatrix}$. According to Theorem 3.3, $\sigma_1 \geq \ldots \sigma_n \geq 0 \geq \ldots \geq 0 \geq -\sigma_n \geq \ldots \geq -\sigma_1$ are the eigenvalues of H, $\hat{\sigma}_1 \geq \ldots \hat{\sigma}_n \geq 0 \geq \ldots \geq 0 \geq -\hat{\sigma}_n \geq \ldots \geq -\hat{\sigma}_1$ are the eigenvalues of $\hat{H}$. Since

$$\hat{H} = \begin{bmatrix} X^T & 0 \\ 0 & Y^T \end{bmatrix} H \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix},$$

if we can prove $\left\| \begin{bmatrix} X^T & 0 \\ 0 & Y^T \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} - I \right\|_2 = \epsilon$, by applying Theorem 5.6, we will prove the result. As the target two-norm equals $\left\| \begin{bmatrix} X^T X - I & 0 \\ 0 & Y^T Y - I \end{bmatrix} \right\|_2$ and

$$\left\| \begin{bmatrix} X^T X - I & 0 \\ 0 & Y^T Y - I \end{bmatrix} \right\|_2 = \max\left(\|X^T X - I\|_2, \|Y^T Y - I\|_2 \right),$$ [6]

the equation holds. Thus, we prove the result. □

Question 5.13. *Let A be a symmetric matrix. Consider running shifted QR iteration (Algorithm 4.5) with a Rayleigh quotient shift ($\sigma_i = a_{nn}$) at every iteration, yielding a sequence $\sigma_1, \sigma_2, \ldots$ of shifts. Also run Rayleigh quotient iteration (Algorithm 5.1), starting with $\boldsymbol{x}_0 = [0, 0, \ldots, 1]^T$, yielding a sequence of Rayleigh quotients $\rho_1, \rho_2, \ldots$. show that these sequences are identical: $\sigma_i = \rho_i$ for all i. This justifies the claim in section 5.3.2 that shifted QR iteration enjoys local cubic convergence.*

Proof. We prove it by induction. At the begining, for the shifted QR iteration, $\sigma_1 = a_{nn}$. For Rayleigh quotient iteration start with $\boldsymbol{x}_0 = [0, 0, \ldots, 1]$, $\rho_1 = \boldsymbol{x}_0^T A \boldsymbol{x}_0 = a_{nn}$. Thus, $\sigma_1 = \rho_1$. Then, suppose it holds until the kth step. For $(k+1)$th step, denote $P_k = Q_k Q_{k-1} \ldots Q_1$. Since A is symmetric, the orthogonal matrix Q_i are symmetric, too[7]. We kown that for the shifted QR iteration,

$$P_k(A_{k+1} - \sigma_k I) = (A - \sigma_k) P_k.$$

What's more,

$$P_k R_k = P_{k-1} Q_k R_k = P_{k-1}(A_k - \sigma_k I) = (A - \sigma_k) P_{k-1}.$$

Transpose on both sides, premultiply by P_{k-1} and postmultiply by P_k, yielding

$$P_{k-1} R_k^T = (A - \sigma_k) P_k.$$

Equate the last column, and we have

$$P_{k-1} \boldsymbol{e}_n = (A - \sigma_k) P_k \boldsymbol{e}_n / r_k, \text{ or equivalently } P_k \boldsymbol{e}_n = (A - \sigma_k)^{-1} P_{k-1} \boldsymbol{e}_n / r_k.$$

where r_k represents $R_k(n,n)$. As $P_0 \boldsymbol{e}_n = \boldsymbol{e}_n$ and $\|P_i \boldsymbol{e}_n\|_2 = 1$, it is a inverse iteration sequence with the same start and same shifts as Rayleigh quotient iteration, by assumption. Thus, $P_k \boldsymbol{e}_n = \boldsymbol{x}_k$, which suggest $\sigma_k = A(n,n) = \boldsymbol{e}_n^T A_k \boldsymbol{e}_n = \boldsymbol{e}_n^T P_k A P_k \boldsymbol{e}_n = \boldsymbol{x}_k^T A \boldsymbol{x}_k = \rho_k$. □

[6] It is because that such a block diagonal matrix has the eigenvalues as the union of those of its diagonal blocks.

[7] It is because $(QR)^T = R^T Q^T = QR$, yielding $Q = R^T Q^T R$ the right side of which is symmetric.

Question 5.14. *Prove Lemma 5.1.*

Proof. The prove will be trivial if $\boldsymbol{x}$ or $\boldsymbol{y}$ are zero vector. Thus, we may assume that both of them are nonzero. Then, there exists a row of $\boldsymbol{x}\boldsymbol{y}^T$ that can linearly represent the other rows. Without loss of generality, suppose it is the first row. Otherwise, we may change it to the first row by exchange rows and columns. Denote the ith entry of $\boldsymbol{x}$ and $\boldsymbol{y}$ as x_1 and y_1 repectively. Then, by assumption, $x_1 \neq 0$. We have

$$\begin{aligned}
\det(I + \boldsymbol{x}\boldsymbol{y}^T) &= \det \begin{bmatrix} 1 + x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & 1 + x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \dots & 1 + x_n y_n \end{bmatrix} \\
&= \det \begin{bmatrix} 1 + x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ -x_2/x_1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -x_n/x_1 & 0 & \dots & 1 \end{bmatrix} \\
&= \det \begin{bmatrix} 1 + \sum_{i=1}^n x_i y_i & x_1 y_2 & \dots & x_1 y_n \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \\
&= 1 + \sum_{i=1}^n x_i y_i = 1 + \boldsymbol{x}^T \boldsymbol{y}.
\end{aligned}$$

□

Remark. There is a simpler way to prove the question. First, observe that $\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{x} = (\boldsymbol{y}^T\boldsymbol{x})\boldsymbol{x}$, which means $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{x}\boldsymbol{y}^T$ corresponding to $\boldsymbol{y}^T\boldsymbol{x}$. As the rank of $\boldsymbol{x}\boldsymbol{y}^T$ is one, $\boldsymbol{x}\boldsymbol{y}^T$ has only one nonzero eigenvalue. Then by Question 4.5, the eigenvalues of $I + \boldsymbol{x}\boldsymbol{y}^T$ are 1's and $1 + \boldsymbol{y}^T\boldsymbol{x}$. The result follows.

Question 5.15. *Prove that if $t(n) = 2t(n/2) + cn^3 + O(n^2)$, then $t(n) \approx c\frac{4}{3}n^3$. This justifies the complexity analysis of the divide-and-conquer algorithm (Algorithm 5.2).*

Proof. Since we have

$$\begin{aligned}
t(n) &= 2t(n/2) + cn^3 + O(n^2), \\
2t(n/2) &= 4t(n/4) + \frac{1}{4}cn^3 + O(n^2), \\
&\cdots\cdots\cdots \\
2^{\log_2 n - 1} t(2) &= 2^{\log_2 n} t(1) + \frac{1}{2^{2\log_2 n - 2}} cn^3 + O(n^2),
\end{aligned}$$

by adding them together, we get

$$t(n) = 2^{\log_2 n} t(1) + c\frac{4}{3}(1 - n^{-2})n^3 + O(n^2) \approx \frac{4}{3}cn^3.$$

□

Question 5.16. *Let $A = D + \rho\boldsymbol{u}\boldsymbol{u}^T$, where $D = diag(d_1,\dots,d_n)$ and $\boldsymbol{u} = [u_1,\dots,u_n]^T$. Show that if $d_i = d_{i+1}$ or $u_i = 0$, then d_i is an eigenvalues of A. If $u_i = 0$, show that the eigenvector corresponding to d_i is $\boldsymbol{e}_i$, the ith column of the identity matrix. Derive a similarly simple expression when $d_i = d_{i+1}$. This shows how to handle deflation in the divide-and-conquer algorithm, Algorithm 5.2.*

Proof. When $u_i = 0$, then the ith column of $\boldsymbol{u}\boldsymbol{u}^T$ is zero vector. Thus,

$$(D + \rho\boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{e}_i = d_i\boldsymbol{e}_i.$$

As for $d_i = d_{i+1}$, without loss of generality, we can assume that $u_i \neq 0$ and $d_{i-1} \neq d_i$, $d_{i+1} \neq d_{i+2}$; otherwise, such case will either be the previous one or similar to the following. We claim that $\boldsymbol{e}_{i+1} - u_{i+1}/u_i\boldsymbol{e}_i$ is the eigenvector corresponding to d_i, because

$$\begin{aligned}
(A - d_iI)(\boldsymbol{e}_{i+1} - u_{i+1}/u_i\boldsymbol{e}_i) &= \begin{pmatrix} u_1^2 + d_1 - d_i & \dots & u_1u_i & u_1u_{i+1} & \dots & u_1u_n \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_1u_i & \dots & u_i^2 & u_iu_{i+1} & \dots & u_iu_n \\ u_1u_{i+1} & \dots & u_iu_{i+1} & u_{i+1}^2 & \dots & u_{i+1}u_n \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_1u_n & \dots & u_iu_n & u_{i+1}u_n & \dots & u_n^2 + d_n - d_i \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ -u_{i+1}/u_i \\ 1 \\ \vdots \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 1 & \dots & u_1 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & u_i & 0 & \dots & 0 \\ 0 & \dots & u_{i+1} & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & u_n & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} d_1 - d_i & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_1 & \dots & u_i & u_{i+1} & \dots & u_n \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & d_n - d_i \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ -u_{i+1}/u_i \\ 1 \\ \vdots \\ 0 \end{pmatrix} \\
&= \boldsymbol{0}.
\end{aligned}$$

□

Remark. My proof is based on solving $(A - \lambda I)\boldsymbol{x} = 0$. There is another way based on deflation. First, as we can use a Givens rotation to cancel an entry of a vector, if $d_i = d_{i+1}$, denote G as the Givens rotation which satisfies

$$G\begin{pmatrix} u_1 \\ \vdots \\ u_i \\ u_{i+1} \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} u_1 \\ \vdots \\ u_i \\ 0 \\ \vdots \\ u_n \end{pmatrix}.$$

Then,

$$GAG^T = GDG^T + \rho G\boldsymbol{u}\boldsymbol{u}^TG^T = D + \rho\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T,$$

where $\hat{u}_i = 0$.

Question 5.17. *Let ψ and ψ' be given scalars. Show how to compute scalars c and $\hat{c}$ in the function definition $h(\lambda) = \hat{c} + \frac{c}{d-\lambda}$ so that at $\lambda = \xi$, $h(\xi) = \psi$ and $h'(\xi) = \psi'$. This result is needed to derive the secular equation solver in section 5.3.3.*

Solution. As

$$h'(\lambda) = \frac{c}{(d-\lambda)^2},$$

the conditions $h(\xi) = \psi$ and $h'(\xi) = \psi'$ become

$$\psi = \hat{c} + \frac{c}{d-\xi}, \quad \psi' = \frac{c}{(d-\xi)^2},$$

whose solution is

$$c = \psi'(d-\xi)^2, \quad \hat{c} = \psi - \psi'(d-\xi).$$

□

Question 5.18. *Use the SVD to show that if A is an m-by-n real matrix with $m \geq n$, then there exists an m-by-n matrix Q with orthonormal columns ($Q^TQ = I$) and an n-by-n positive semidefinite matrix P such that $A = QP$. This decomposition is called the* ***polar decomposition*** *of A, because it is analogous to the polar form of a complex number $z = e^{i\arg(z)} \cdot |z|$. Show that if A is nonsingular, then the polar decomposition is unique.*

Proof. Suppose $A = U\Sigma V^T$, which is the reduced SVD of A. To show how to do the polar decomposition, consider

$$A = U\Sigma V^T = UV^TV\Sigma^{1/2}\Sigma^{1/2}V^T = (UV^T)(\Sigma^{1/2}V^T)^T(\Sigma^{1/2}V^T).$$

The UV^T is m-by-n orthonormal and $(\Sigma^{1/2}V^T)^T(\Sigma^{1/2}V^T)$ is a positive semidefinite matrix, which satisfy the condition of polar decomposition.
For the uniqueness, suppose there are two polar decomposition of A as $A = QP = \hat{Q}\hat{P}$. Consider $A^TA = P^2 = \hat{P}^2$. As P and $\hat{P}$ are positive definite matrices when A has full rank, P and $\hat{P}$ have the same eigenvalues. Suppose $P = W^T\Lambda W$, $\hat{P} = S^T\Lambda S$. We get

$$SW^T\Lambda WS^T = \Lambda.$$

It follows that

$$\hat{P} = S^T\Lambda S = S^TSW^T\Lambda WS^TS = W^T\Lambda W = P,$$

and $AP^{-1} = Q = \hat{Q} = A\hat{P}^{-1}$. □

Remark. In fact, the uniqueness follows from the uniqueness of $\sqrt{A^TA}$. Standard textbooks about linear algebra often define $\sqrt{A^TA} = W^T\Lambda W$ without checking whether it is well defined as the orthogonal matrix W may not be unique. Thus, I present the proof here.

Question 5.19. *Prove Lemma 5.5*

Proof.

1. Denote $\boldsymbol{x}$ as any vector in $\mathbb{R}^{2n}$. Then,

$$P^T\boldsymbol{x} = [\boldsymbol{e}_1^T, \dots, \boldsymbol{e}_{2n}^T]^T\boldsymbol{x} = [x_1, x_{n+1}, x_2, \dots . x_n, x_{2n}]^T,$$
$$\boldsymbol{x}^T P = \boldsymbol{x}^T[\boldsymbol{e}_1, \dots, \boldsymbol{e}_{2n}] = [x_1, x_{n+1}, x_2, \dots, x_n, x_{2n}].$$

Thus, denote $a_{i,j}$ as the ith row, jth column entry of A. We have

$$P^T AP = \begin{pmatrix} a_{1,1} & a_{2,1} & \dots & a_{2n,1} \\ a_{1,n+1} & a_{2,n+1} & \dots & a_{2n,n+1} \\ a_{1,2} & a_{2,2} & \dots & a_{2n,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & \dots & a_{2n,n} \\ a_{1,2n} & a_{2,2n} & \dots & a_{2n,2n} \end{pmatrix} P = \begin{pmatrix} a_{1,1} & a_{n+1,1} & \dots & a_{n,1} & a_{2n,1} \\ a_{1,n+1} & a_{n+1,n+1} & \dots & a_{n,n+1} & a_{2n,n+1} \\ a_{1,2} & a_{n+1,2} & \dots & a_{n,2} & a_{2n,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{1,n} & a_{n+1,n} & \dots & a_{n,n} & a_{2n,n} \\ a_{1,2n} & a_{n+1,2n} & \dots & a_{n,2n} & a_{2n,2n} \end{pmatrix}.$$

i.e.,

$$P^T AP = \begin{pmatrix} 0 & a_1 & \dots & 0 & 0 \\ a_1 & 0 & \dots & 0 & 0 \\ 0 & b_1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & a_n \\ 0 & 0 & \dots & a_n & 0 \end{pmatrix}.$$

2. We can verify by simple computations:

$$\begin{pmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & \ddots & \ddots & \\ & & & a_{n-1} & b_{n-1} \\ & & & & a_n \end{pmatrix} \begin{pmatrix} a_1 & & & & \\ b_1 & a_2 & & & \\ & \ddots & \ddots & & \\ & & b_{n-2} & a_{n-1} & \\ & & & b_{n-1} & a_n \end{pmatrix} = \begin{pmatrix} a_1^2 + b_1^2 & a_2b_1 & & \\ a_2b_1 & a_2^2 + b_2^2 & a_3b_2 & \\ & \ddots & \ddots & \\ & & a_nb_{n-1} & a_n^2 \end{pmatrix}.$$

As for the remaining result, suppose $B = U\Sigma V^T$, which is the SVD of B. Then, $BB^T = U\Sigma^2 U^T$. Therefore, the singular value are the square roots of the eigenvalues of T_{BB^T}, and the left singular vectors of B are the eigenvectors of T_{BB^T}.

3. We can verify by simple computations:

$$\begin{pmatrix} a_1 & & & & \\ b_1 & a_2 & & & \\ & \ddots & \ddots & & \\ & & b_{n-2} & a_{n-1} & \\ & & & b_{n-1} & a_n \end{pmatrix} \begin{pmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & \ddots & \ddots & \\ & & & a_{n-1} & b_{n-1} \\ & & & & a_n \end{pmatrix} = \begin{pmatrix} a_1^2 & a_1b_1 & & \\ a_1b_1 & a_2^2 + b_1^2 & a_2b_2 & \\ & \ddots & \ddots & a_{n-1}b_{n-1} \\ & & a_{n-1}b_{n-1} & a_{n-1}^2 + b_{n-1}^2 \end{pmatrix}.$$

As for the remaining result, suppose $B = U\Sigma V^T$, which is the SVD of B. Then, $B^T B = V\Sigma^2 V^T$. Therefore, the singular value are the square roots of the eigenvalues of T_{B^TB}, and the right singular vectors of B are the eigenvectors of T_{B^TB}.

□

Question 5.20. *Prove Lemma 5.7*

Proof. We can verify it by simple computations:

$$D_1BD_2 = \begin{pmatrix} \chi_1 & & & & \\ & \frac{\chi_2\chi_1}{\zeta_1} & & & \\ & & \frac{\chi_3\chi_2\chi_1}{\zeta_2\zeta_1} & & \\ & & & \ddots & \\ & & & & \frac{\chi_n\cdots\chi_1}{\zeta_{n-1}\cdots\zeta_1} \end{pmatrix} \begin{pmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & a_3 & b_3 & \\ & & & \ddots & \ddots \\ & & & & a_n \end{pmatrix} \begin{pmatrix} 1 & & & & \\ & \frac{\zeta_1}{\chi_1} & & & \\ & & \frac{\zeta_2\zeta_1}{\chi_2\chi_1} & & \\ & & & \ddots & \\ & & & & \frac{\zeta_{n-1}\ldots\zeta_1}{\chi_{n-1}\cdots\chi_1} \end{pmatrix}$$

$$= \begin{pmatrix} \chi_1 a_1 & \chi_1 b_1 & & & \\ & \frac{\chi_2\chi_1}{\zeta_1}a_2 & \frac{\chi_2\chi_1}{\zeta_1}b_2 & & \\ & & \frac{\chi_3\chi_2\chi_1}{\zeta_2\zeta_1}a_3 & \frac{\chi_3\chi_2\chi_1}{\zeta_2\zeta_1}b_3 & \\ & & & \ddots & \ddots \\ & & & & \frac{\chi_n\cdots\chi_1}{\zeta_{n-1}\ldots\zeta_1}a_n \end{pmatrix} \begin{pmatrix} 1 & & & & \\ & \frac{\zeta_1}{\chi_1} & & & \\ & & \frac{\zeta_2\zeta_1}{\chi_2\chi_1} & & \\ & & & \ddots & \\ & & & & \frac{\zeta_{n-1}\ldots\zeta_1}{\chi_{n-1}\cdots\chi_1} \end{pmatrix}$$

$$= \begin{pmatrix} \chi_1 a_1 & \zeta_1 b_1 & & & \\ & \chi_2 a_2 & \zeta_2 b_2 & & \\ & & \chi_3 a_3 & \zeta_3 b_3 & \\ & & & \ddots & \ddots \\ & & & & \chi_n a_n \end{pmatrix}.$$

□

Question 5.21. *Prove Theorem 5.13. Also, reduce the exponent* $4n-2$ *in Theorem 5.13 to* $2n-1$.

Proof. First, we can evaluate $\|D_1D_1 - I\|_2$ as follows:

$$\begin{aligned} \|D_1D_1 - I\|_2 &= \left\| \text{diag}\{\chi_1^2 - 1, \frac{\chi_2^2\chi_1^2}{\zeta_1^2} - 1, \ldots, \frac{\chi_n^2\cdots\chi_1^2}{\zeta_{n-1}^2\cdots\zeta_1^2} - 1\} \right\|_2 \\ &= \max_{1\le i\le n} \left| \frac{\prod_{j=1}^{i}\chi_j^2 - \prod_{j=1}^{i-1}\zeta_j^2}{\prod_{j=1}^{i-1}\zeta_j^2} \right| \\ &\le \max_{1\le i\le n} \tau^{2i-2} \left| \prod_{j=1}^{i}\chi_j^2 - \prod_{j=1}^{i-1}\zeta_j^2 \right| \\ &\le \max_{1\le i\le n} \tau^{2i-2}(\tau^{2i} - \tau^{-2i+2}) \\ &\le \max_{1\le i\le n} \tau^{4i-2} - 1 = \tau^{4n-2} - 1. \end{aligned}$$

For $\|D_2D_2 - I\|_2$, using the same techinique, we can prove $\|D_2D_2 - I\|_2 = \tau^{4n-4} - 1$. Thus, $\tau^{4n-2} - 1 \equiv \max(\|D_1D_1 - I\|_2, \|D_2D_2 - I\|_2)$, and by Corollary 5.2, we have

$$|\hat{\sigma}_i - \sigma_i| \le (\tau^{4n-2} - 1)\sigma_i = ((\epsilon + 1)^{4n-2} - 1)\sigma_i = (4n-2)\epsilon\sigma_i + O(\epsilon^2).$$

To improve the bound, it is necessary to improve $\max(\|D_1D_1 - I\|_2, \|D_2D_2 - I\|_2)$. Consider, $\hat{D}_1 = (\frac{1}{\sqrt{\chi_n}}\prod_{j=1}^{[n/2]}\zeta_j/\chi_j)D_1, \hat{D}_2 = (\sqrt{\chi_n}\prod_{j=1}^{[n/2]}\chi_j/\zeta_j)D_2$[8]. Then,

$$\hat{D}_1 = \operatorname{diag}\left(\frac{\chi_1}{\sqrt{\chi_n}}\prod_{j=1}^{[n/2]}\zeta_j/\chi_j, \frac{\chi_2}{\sqrt{\chi_n}}\prod_{j=2}^{[n/2]}\zeta_j/\chi_j, \dots, \sqrt{\chi_n}\prod_{j=[n/2]+1}^{n-1}\chi_j/\zeta_j\right),$$

$$\hat{D}_2 = \operatorname{diag}\left(\sqrt{\chi_n}\prod_{j=1}^{[n/2]}\chi_j/\zeta_j, \sqrt{\chi_n}\prod_{j=2}^{[n/2]}\chi_j/\zeta_j, \dots, \sqrt{\chi_n}\prod_{j=[n/2]+1}^{n-1}\zeta_j/\chi_j\right)$$

It is easy to verify that $\hat{B} = \hat{D}_1 B \hat{D}_2 = D_1 B D_2$, and

$$\|\hat{D}_1\hat{D}_1 - I\|_2 \le \max\left(\max_{1\le i\le[n/2]}\left|\frac{\chi_i^2}{\chi_n}\prod_{j=i}^{[n/2]}\zeta_j^2/\chi_j^2 - 1\right|, \max_{[n/2]+1\le i\le n}\left|\frac{\chi_i^2}{\chi_n}\prod_{j=[n/2]+1}^{i-1}\chi_j^2/\zeta_j^2 - 1\right|\right) \le \tau^{2n-1} - 1,$$

$$\|\hat{D}_2\hat{D}_2 - I\|_2 \le \max\left(\max_{1\le i\le[n/2]}\left|\chi_n\prod_{j=i}^{[n/2]}\chi_j^2/\zeta_j^2 - 1\right|, \max_{[n/2]+1\le i\le n}\left|\chi_n\prod_{j=[n/2]+1}^{i-1}\zeta_j^2/\chi_j^2 - 1\right|\right) \le \tau^{2n-1} - 1.$$

Thus, we improve that bound from $4n-2$ to $2n-1$. □

Question 5.22. *Prove that Algorithm 5.13 computes the SVD of G, assuming that G^TG converges to a diagonal matrix.*

Remark. I am not sure whether those σ_i are ordered by their corresponding magnitude, and it seems that G must have full rank. Otherwise, $U = [G(:,1)/\sigma_1, G(:,2)/\sigma_2, \dots, G(:,n)/\sigma_n]$ cannot be orthonormal as $G(:,i)$ are linear dependent or some $\sigma_i = 0$.

Proof. Denote $\hat{G}$ as the resulted matrix such that $\hat{G}^T\hat{G} = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2) = \Lambda$. As $\hat{G}^T\hat{G}$, it follows

$$\sigma_j^2 = \sum_{i=1}^n \hat{G}(i,j)^2 = \|\hat{G}(:,j)\|_2^2.$$

Then, we claim that $U = [\hat{G}(:,1)/\sigma_1, \hat{G}(:,2)/\sigma_2, \dots, \hat{G}(:,n)/\sigma_n]$ is orthonormal, since

$$U^TU = \begin{bmatrix}\hat{G}(:,1)^T/\sigma_1 \\ \hat{G}(:,2)^T/\sigma_2 \\ \vdots \\ \hat{G}(:,n)^T/\sigma_n\end{bmatrix}\left[\hat{G}(:,1)/\sigma_1, \hat{G}(:,2)/\sigma_2, \dots, \hat{G}(:,n)/\sigma_n\right] = I,$$

$$\begin{aligned} UU^T &= \left[\hat{G}(:,1)/\sigma_1, \hat{G}(:,2)/\sigma_2, \dots, \hat{G}(:,n)/\sigma_n\right]\begin{bmatrix}\hat{G}(:,1)^T/\sigma_1 \\ \hat{G}(:,2)^T/\sigma_2 \\ \vdots \\ \hat{G}(:,n)^T/\sigma_n\end{bmatrix} \\ &= \sum_{i=1}^n \hat{G}(:,i)\hat{G}(:,i)^T/\sigma_i^2 = \hat{G}\Lambda^{-1}\hat{G}^T = \hat{G}(\Lambda^{-1}\hat{G}^T\hat{G})\hat{G}^{-1} = I. \end{aligned}$$

[8] $[a]$ denotes the largest integer that smaller than a. Another feasible const is $\sqrt{\chi_1}\prod_{j=1}^{(n/2)}\zeta_j\chi_j$, where (a) means the smallest ineger that larger than a.

Finally, as

$$U\Sigma V^T = [\hat{G}(:,1)/\sigma_1, \hat{G}(:,2)/\sigma_2, \ldots, \hat{G}(:,n)/\sigma_n]\sqrt{\Lambda}J^T = \hat{G}J^T = \hat{G}J^{-1} = G,$$

it follows that Algorithm 5.13 computes the SVD of G. □

Question 5.23. *A harder question. Skipped.*

Question 5.24. *A harder question. Skipped.*

Question 5.25. *A harder question. Skipped.*

Question 5.26. *Suppose that $\boldsymbol{x}$ is an n-vector. Define the matrix C by $c_{ij} = |x_i| + |x_j| - |x_i - x_j|$. Show that $C(\boldsymbol{x})$ is positive semidefinite.*

Proof. As

$$c_{ij} = |x_i| + |x_j| - |x_i - x_j| = \begin{cases} 2\min(|x_i|, |x_j|), & x_i x_j \geq 0, \\ 0, & x_i x_j < 0, \end{cases}$$

we may assume that $x_1 \geq x_2 \geq \ldots \geq x_k \geq 0 \geq x_{k+1} \ldots \geq x_n$. Otherwise, we can use permutation matrix to transform other situation into this. Then, the matrix C has the form

$$C = 2\begin{bmatrix} x_1 & x_2 & \ldots & x_k & & & & \\ x_2 & x_2 & \ldots & x_k & & & & \\ \vdots & \vdots & \ddots & \vdots & & & & \\ x_k & x_k & \ldots & x_k & & & & \\ & & & & -x_{k+1} & -x_{k+2} & \ldots & -x_n \\ & & & & -x_{k+2} & -x_{k+2} & \ldots & -x_n \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & -x_n & -x_n & \ldots & -x_n \end{bmatrix} = \begin{bmatrix} C_1 & \\ & C_2 \end{bmatrix}.$$

Thus, it suffices to prove for the $\boldsymbol{x}$ which satisfies $|\boldsymbol{x}| = \boldsymbol{x}$, and consequently C takes the form

$$C = \begin{bmatrix} x_1 & x_2 & \ldots & x_k \\ x_2 & x_2 & \ldots & x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_k & x_k & \ldots & x_k \end{bmatrix}.$$

For simlicity, call the matrices which have this form as A-matrices. We prove that A-matrices are semipositive by induction. First, as $x_i \geq 0$ by assumption, the base case holds. Then, suppose that all $(k-1)$-by-$(k-1)$ A-matrices are semipositive. For any k-by-k A-matrix C, its determinant can be computed as

$$C = x_k \begin{bmatrix} x_1 & x_2 & \ldots & x_k \\ x_2 & x_2 & \ldots & x_k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix} = x_k \begin{bmatrix} & x_2 - x_1 & x_3 - x_1 & \ldots & x_k - x_1 \\ & & x_3 - x_2 & \ldots & x_k - x_2 \\ & & & \ddots & \vdots \\ & & & & x_k - x_{k-1} \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix}$$

$$= (-1)^{2k} x_k \prod (x_i - x_{i+1}) \geq 0.$$

Thus, all the principal submatrices of C has nonnegative determinant, which implies C is semipositive. □

Question 5.27. *Let*

$$A = \begin{pmatrix} I & B \\ B^* & I \end{pmatrix}$$

with $\|B\|_2 < 1$. *Show that*

$$\|A\|_2 \|A^{-1}\|_2 = \frac{1 + \|B\|_2}{1 - \|B\|_2}.$$

Proof. To prove the equation, we need to compute the $\|A\|_2$ and $\|A^{-1}\|_2$ exactly. Denote B as n-by-m matrix, and partion any vector $\boldsymbol{x} \in \mathbb{R}^{n+m}$ as $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}$ where $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{z} \in \mathbb{R}^m$. Then we can evaluate the largest and smallest eigenvalues of A as follows:

$$\lambda_{\max} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* A \boldsymbol{x} = \max_{\|\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2=1} \begin{pmatrix} \boldsymbol{y}^* & \boldsymbol{z}^* \end{pmatrix} \begin{pmatrix} I & B \\ B^* & I \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = 1 + 2 \max_{\|\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2=1} \operatorname{Re}(\boldsymbol{y}^* B \boldsymbol{z}),$$

$$\lambda_{\min} = 1 + 2 \min_{\|\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2=1} \operatorname{Re}(\boldsymbol{y}^* B \boldsymbol{z}).$$

As $2|\operatorname{Re}(\boldsymbol{y}^* B \boldsymbol{z})| = \|\boldsymbol{y}^* B \boldsymbol{z} + \boldsymbol{z}^* B^* \boldsymbol{y}\|_2 \le 2\|\boldsymbol{y}\|_2 \|B\|_2 \|\boldsymbol{z}\|_2 \le (\|\boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2)\|B\|_2 = \|B\|_2$, the above formulas has an upper bound as $1 + \frac{1}{2}\|B\|_2$, and a lower bound as $1 - \frac{1}{2}\|B\|_2$. To see both of them are attainable, set $\hat{\boldsymbol{z}}$ as the vector which satisfies $\|B\hat{\boldsymbol{z}}\|_2 = \|B\|_2$, and $\hat{\boldsymbol{y}}$ which is a unit vector and parallel to $\hat{\boldsymbol{z}}^* B^*$. Thus,

$$\frac{\boldsymbol{y}^* B \boldsymbol{z}}{\|\boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2} = \frac{\pm 1}{2\|B\|_2} \hat{\boldsymbol{z}}^* B^* B \boldsymbol{z} = \frac{\pm 1}{2\|B\|_2} \|B\boldsymbol{z}\|_2^2 = \frac{\pm 1}{2} \|B\|_2.$$

What's more, since we know $\|B\|_2 < 1$, the matrix A are positive definite. Thus, the largest eigenvalue of A^{-1} is the inverse of smallest eigenvalue of A. Therefore,

$$\|A\|_2 \|A^{-1}\|_2 = \frac{1 + \|B\|_2}{1 - \|B\|_2}.$$

□

Question 5.28. *A square matrix A is said to be* ***skew Hermitian*** *if* $A^* = -A$. *Prove that*

1. *the eigenvalues of a skew Hermitian are purely imaginary.*
2. $I - A$ *is nonsingular.*
3. $C = (I - A)^{-1}(I + A)$ *is unitary, C is called the* ***Cayley transform*** *of A.*

Proof.

1. Denote λ as an eigenvalue of matrix A, and $\boldsymbol{x}$ as the corresponding unit eigenvector. Thus, $A\boldsymbol{x} = \lambda \boldsymbol{x}$ and $-\boldsymbol{x}^* A = \bar{\lambda} \boldsymbol{x}^*$. Consequently,

$$\boldsymbol{x}^* A \boldsymbol{x} = \lambda \boldsymbol{x}^* \boldsymbol{x} = \lambda,$$
$$-\boldsymbol{x}^* A \boldsymbol{x} = \bar{\lambda} \boldsymbol{x}^* \boldsymbol{x} = \bar{\lambda}.$$

Add them together, and we have $\lambda + \bar{\lambda} = 0$, which means λ is pure imaginary.

2. According to Question 4.5, the eigenvalues of $I - A$ are $1-\lambda_i$, where λ_i are the eigenvalues of A. Since all the eigenvalues of A are pure imaginary,

$$\prod |1-\lambda_i| = \prod \sqrt{1+\lambda_i^2},$$

which suggests that $I - A$ is invertible.

3. Since $C^* = (I-A)(I+A)^{-1}$, we have

$$\begin{aligned}
C^*C &= (I-A)(I+A)^{-1}(I-A)^{-1}(I+A)\\
&= (I-A)(I-A+A-AA)^{-1}(1+A)\\
&= (I-A)(I-A)^{-1}(I+A)^{-1}(1+A)\\
&= I,\\
CC^* &= (I-A)^{-1}(I+A)(I-A)(I+A)^{-1}\\
&= (I-A)^{-1}(I-A)(I+A)(I+A)^{-1}\\
&= I.
\end{aligned}$$

Thus, C is unitary.

□

6 Solutions for Chapter VI: Iterative Methods for Linear Systems

Question 6.1. *Prove Lemma 6.1.*

Proof. It suffices to verify $T_N \boldsymbol{z}_j = \lambda_j \boldsymbol{z}_j$ as follows:

$$\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix} \begin{pmatrix} \sin(\frac{j\pi}{N+1}) \\ \sin(\frac{2j\pi}{N+1}) \\ \vdots \\ \sin(\frac{jN\pi}{N+1}) \end{pmatrix} = \begin{pmatrix} 2\sin(\frac{j\pi}{N+1}) - \sin(\frac{2j\pi}{N+1}) \\ \vdots \\ 2\sin(\frac{kj\pi}{N+1}) - \sin(\frac{(k+1)j\pi}{N+1}) - \sin(\frac{(k-1)j\pi}{N+1}) \\ \vdots \\ 2\sin(\frac{Nj\pi}{N+1}) - \sin(\frac{(N-1)j\pi}{N+1}) \end{pmatrix}$$

As $\sin(A) + \sin(A) = 2\sin(\frac{A+B}{2})\cos(\frac{A-B}{2})$, it follows

$$2\sin(\frac{kj\pi}{N+1}) - \sin(\frac{(k+1)j\pi}{N+1}) - \sin(\frac{(k-1)j\pi}{N+1}) = 2\sin(\frac{kj\pi}{N+1}) - 2\sin(\frac{kj\pi}{N+1})\cos(\frac{j\pi}{N+1}).$$

As for the first and last entries, it also holds because $\sin(\frac{0\cdot j\pi}{N+1}) = \sin(\frac{(N+1)\cdot j\pi}{N+1}) = 0$. □

Question 6.2. *Prove the following formulas for triangular factorizations of T_N.*

1. *The Cholesky factorization $T_N = B_N^T B_N$ has a upper bidiagonal Cholesky factor B_N with*

$$B_N(i,i) = \sqrt{\frac{i+1}{i}}, B_N(i,i+1) = \sqrt{\frac{i}{i+1}}.$$

2. *The result of Gaussian elimination with partial pivoting on T_N is $T_N = L_N U_N$, where the triangular factors are bidiagonal:*

$$\begin{aligned} L_N(i,i) &= 1,\ L_N(i+1,i) = -\frac{i}{i+1}, \\ U_N(i,i) &= \frac{i+1}{i},\ U_N(i,i+1) = -1. \end{aligned}$$

3. *$T_N = D_N D_N^T$, where D_N is the N-by-$(N+1)$ upper bidiagonal matrix with 1 on the main diagonal and -1 on the superdiagonal.*

Proof.

1. It suffices to verify that $T_N = B_N^T B_N$. According to Question 5.19, the diagonal entries of $B_N^T B_N$ are $\frac{i+1}{i} + \frac{i-1}{i} = 2$, and the offdiagonal entries are $\sqrt{\frac{i+1}{i}\frac{i}{i+1}} = 1$. Thus, $T_N = B_N^T B_N$.
2. As T_N is band matrix, $T(i,i)$ would be updated only at $(i-1)$th step. For the first step, $L_N(1,1) = 1$, $L_N(2,1) = -\frac{1}{2}$ and $U_N(1,1) = 2$, $U_N(1,2) = -1$. Suppose it holds for $(k-1)$th step and does not require swapping rows. Then, for kth step, we have

$$T_N = L_{k-1}\begin{pmatrix} \ddots & \ddots & & & \\ \ddots & \frac{k}{k-1} & -1 & & \\ & -1 & 2 & -1 & \\ & & & \ddots & \ddots \end{pmatrix} = L_{k-1}\begin{pmatrix} \ddots & & & \\ & 1 & & \\ & \frac{k-1}{k} & 1 & \\ & & & \ddots \end{pmatrix}\begin{pmatrix} \ddots & \ddots & & & \\ \ddots & \frac{k}{k-1} & -1 & & \\ & & \frac{k+1}{k} & -1 & \\ & & & \ddots & \ddots \end{pmatrix}.$$

   Therefore, it also holds for kth step and does not require swapping rows.
3. It suffices to verify that $T_N = D_N^T D_N$ as follows:

$$D_N D_N^T = \begin{pmatrix} 1 & -1 & & & \\ & \ddots & \ddots & & \\ & & 1 & -1 & \\ & & & 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & & & \\ -1 & \ddots & & \\ & \ddots & 1 & \\ & & -1 & 1 \\ & & & -1 \end{pmatrix} = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & 2 & -1 \\ & & -1 & 2 \end{pmatrix} = T_N$$

□

Question 6.3. *Confirm equation (6.13).*

Proof. It suffices to verify that $(-\frac{\partial^2}{\partial x^2}-\frac{\partial^2}{\partial y^2})\sin(i\pi x)\sin(j\pi y) = (i^2\pi^2+j^2\pi^2)\sin(i\pi x)\sin(j\pi y)$ as follows:

$$\begin{aligned}(-\frac{\partial^2}{\partial x^2}-\frac{\partial^2}{\partial y^2})\sin(i\pi x)\sin(j\pi y) &= -\frac{\partial^2}{\partial x^2}\sin(i\pi x)\sin(j\pi y)-\frac{\partial^2}{\partial y^2}\sin(i\pi x)\sin(j\pi y)\\ &= i^2\pi^2\sin(i\pi x)\sin(j\pi y)+j^2\pi^2\sin(i\pi x)\sin(j\pi y).\end{aligned}$$

□

Question 6.4. *1. Prove Lemma 6.2.*

2. Prove Lemma 6.3.

3. Prove that the Sylvester equation $AX - XB = C$ is equivalent to $(I_n \otimes A - B^T \otimes I_m) vec(X) = vec(C)$.

4. Prove that $vec(AXB) = (B^T \otimes A) \cdot vec(X)$.

Proof.

1. As the textbook has already proved part III of Lemma 6.2, we only need to prove part I and part II. For part I, partition X by columns as $X = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]$. Then

$$\text{vec}(AX) = \begin{pmatrix} A\boldsymbol{x}_1 \\ \vdots \\ A\boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} A & & \\ & \ddots & \\ & & A \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = (I_n \otimes A)\cdot \text{vec}(X).$$

The proof for part II is similar. It suffices to verify that $XB(i,j) = \sum_{k=1}^n B(k,j)X(i,k)$, which holds by the definition of matrix product.

2. Since the (i,j) block of $(A\otimes B)(C\otimes D)$ is $\sum_{k=1}^n A(i,k)C(k,j)BD = (AC)\otimes(BD)$, part I follows. For part II, use the result of part I, yielding

$$(A\otimes B)(A^{-1}\otimes B^{-1}) = I_m \otimes I_n = I_{m+n}.$$

Thus, $(A\otimes B)^{-1} = (A^{-1}\otimes B^{-1})$. As for part III, we can verify it as follows

$$(A\otimes B)^T = \begin{pmatrix} A(1,1)B^T & A(2,1)B^T & \ldots & A(n,1)B^T \\ A(1,2)B^T & A(2,2)B^T & \ldots & A(n,2)B^T \\ \vdots & \vdots & \ddots & \vdots \\ A(1,n)B^T & A(2,n)B^T & \ldots & A(n,n)B^T \end{pmatrix} = A^T \otimes B^T.$$

3. Take operator "vec" on both side of the Sylvester equation, yielding

$$\text{vec}(AX - XB) = \text{vec}(AX) - \text{vec}(XB) = (I\otimes A)\text{vec}(X) - (B^T \otimes I)\text{vec}(X) = \text{vec}(C).$$

4. It can be verified as

$$\text{vec}(AXB) = (I\otimes A)\text{vec}(XB) = (I\otimes A)\cdot(B^T\otimes I)\text{vec}(X) = (B^T\otimes A)\text{vec}(X).$$

□

Question 6.5. *Suppose that $A^{n\times n}$ is diagonalizable, so A has n independent eigenvectors: $A\boldsymbol{x}_i = \alpha_i \boldsymbol{x}_i$, or $AX = X\Lambda_A$, where $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ and $\Lambda_A = diag(\alpha_i)$. Similarly, suppose that $B^{m\times m}$ is diagonalizable, so B has m independent eigenvectors: $B\boldsymbol{y}_i = \beta_i \boldsymbol{y}_i$, or $BY = Y\Lambda_B$, where $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m]$ and $\Lambda_B = diag(\beta_j)$. Prove the following results.*

1. *The mn eigenvalues of $I_m \otimes A + B \otimes I_n$ are $\lambda_{ij} = \alpha_i + \beta_j$, i.e., all possible sums of pairs of eigenvalues of A and B. The corresponding eigenvectors are $\boldsymbol{z}_{ij}$, where $\boldsymbol{z}_{ij} = \boldsymbol{x}_i \otimes \boldsymbol{y}_j$, whose $(km+l)$th entry is $\boldsymbol{x}_i(k)\boldsymbol{y}_j(l)$. Written another way,*

$$(I_m \otimes A + B \otimes I_n)(Y \otimes X) = (Y \otimes X)(I_m \otimes \Lambda_A + \Lambda_B \otimes I_n).$$

2. *The Sylvester equation $AX + XB^T = C$ is nonsingular if and only if the sum $\alpha_i + \beta_j = 0$ for all eigenvalues α_i of A and β_j of B. The same is true for the slightly different Sylvester equation $AX + XB = C$.*

3. *The mn eigenvalues of $A \otimes B$ are $\lambda_{ij} = \alpha_i \beta_j$, i.e., all possible products of pairs of eigenvalues of A and B. The corresponding eigenvectors are $\boldsymbol{z}_{ij}$, where $\boldsymbol{z}_{ij} = \boldsymbol{x}_i \otimes \boldsymbol{y}_j$, whose $(km+l)$th entry is $\boldsymbol{x}_i(k)\boldsymbol{y}_j(l)$. Written another way,*

$$(B \otimes A)(Y \otimes X) = (Y \otimes X)(\Lambda_B \otimes \Lambda_A).$$

Remark. It seems that part II of the question is wrong. The sum $\alpha_i + \beta_j$ should not equal zero. It has been proved in Question 4.6. Also, I do not know why the author swapped A and B in the matrix form, athough they are equivalent.
Proof.

1. It suffices to verify the equation as follows:

$$\begin{aligned}(I_m \otimes A + B \otimes I_n)(Y \otimes X) &= Y \otimes AX + BY \otimes X = Y \otimes X\Lambda_A + Y\Lambda_B \otimes X \\ &= (Y \otimes X)(I_m + \Lambda_A) + (Y \otimes X)(\Lambda_B + I_n) \\ &= (Y \otimes X)(I_m \otimes \Lambda_A + \Lambda_B \otimes I_n).\end{aligned}$$

2. According to Question 6.4, the first Sylvester equation equals to solve

$$(I_m \otimes A + B \otimes I_n)\text{vec}(X) = \text{vec}(C),$$

which is a linear equation. To solve it, it is necessary and sufficient $\det(I_m \otimes A + B \otimes I_n) \neq 0$. As what part I has proved, the eigenvalues of the above matrix is $\alpha_i + \beta_j$. Thus, for all i, j, $\alpha_i + \beta_j \neq 0$. For the second Sylvester equation, it suffices to prove that the eigenvalues of $I_m \otimes A + B^T \otimes I_n$ are also $\alpha_i + \beta_j$, which is true if we subsititue B as B^T in part I.

3. It suffices to verify the equation as follows:

$$(B \otimes A)(Y \otimes X) = (BY \otimes AX) = (Y\Lambda_B) \otimes (X\Lambda_A) = (Y \otimes X)(\Lambda_B \otimes \Lambda_A).$$

Question 6.6. *Programming question. Skipped.*

Question 6.7. *Prove Lemma 6.7.*

Proof.

1. Since the mth Chebyshev polynomial is defined by $T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x)$ with initial value $T_0(x) = 1, T_1(x) = x$, we can prove it by induction. As $T_0(1) = 1, T_1(1) = 1$ and assume it holds for $(m-1)$th Chebyshev polynomial, it follows
$$T_m(1) = 2T_{m-1}(1) - T_{m-2}(1) = 1.$$

2. As $T_1(x) = x = 2^{1-1}x, T_0(x) = 1$[9], suppose it holds for $(m-1)$th Chebyshev polynomial. It follows
$$T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x) = 2^{m-1}x^m + O(x^{m-1}).$$

3. The recursive formula has unique solution when initial value is given, though it may take many forms. Thus, it suffices to verify such T_m satisfies the recursive formula by induction, since it holds for base case. When $|x| \le 1$, it follows[10]
$$\begin{aligned} T_m = \cos(m \cdot \arccos x) &= \cos\left((m-1)\arccos x + \arccos x\right) \\ &= x\cos(\left(m-1)\arccos x\right) - \sin\left((m-2)\arccos x + \arccos x\right)\sin(\arccos x) \\ &= xT_{m-1} - x\sin\left((m-2)\arccos x\right)\sin(\arccos x) - T_{m-2}\sin(\arccos x)^2 \\ &= xT_{m-1} - T_{m-2} + T_{m-2}\cos(\arccos x)^2 + x/2\Big(T_{m-1} - T_{m-3}\Big) \\ &= \frac{3}{2}xT_{m-1} - T_{m-2} + \frac{x}{2}(2xT_{m-2} - T_{m-3}) \\ &= 2xT_{m-1} - T_{m-2}. \end{aligned}$$
When $|x| \ge 1$, it follows[11]
$$\begin{aligned} T_m = \cosh(m\operatorname{arccosh} x) &= xT_{m-1} + \sinh\left((m-1)\operatorname{arccosh} x\right)\sinh\left(\operatorname{arccosh} x\right) \\ &= xT_{m-1} + x\sinh\left((m-2)\operatorname{arccosh} x\right)\sinh\left(\operatorname{arccosh} x\right) + T_{m-2}\sinh\left(\operatorname{arccosh} x\right)^2 \\ &= xT_{m-1} - T_{m-2} + x^2T_{m-2} + x/2T_{m-1} - x/2T_{m-3} \\ &= \frac{3}{2}T_{m-1} - T_{m-2} + x/2(2xT_{m-2} - T_{m-3}) \\ &= 2T_{m-1} - T_{m-2}. \end{aligned}$$

4. $|T_m(x)| \le 1$ because $|T_m(x)| = |\cos\left(m\arccos x\right)|$ when $|x| \le 1$.

5. As $\cosh x > 0$, the zeros of $T_m(x)$ must lie in segment $[-1, 1]$. When $|x| \le 1$, $T(x) = \cos\left(m\arccos x\right)$, which has zero at $x_i = \cos\left((2i-1)\pi/(2m)\right)$. Since $T_m(x)$ is polynomial with degree m, all its zeros are $x_i = \cos\left((2i-1)\pi/(2m)\right)$.

[9] Although T_0 does not satisfy the formula, yet it does not matter.

[10] $\cos(\arccos x) = x$ holds because $|x| \le 1$.

[11] $\cosh(\operatorname{arccosh} x) = x$ holds because $|x| \ge 1$.

6. Since $\operatorname{arccosh} x = \ln(x \pm \sqrt{x^2-1})$ when $|x| > 1$, it follows

$$T_m = \frac{1}{2}\big(\exp(m\ln(x \pm \sqrt{x^2-1})) + \exp(-m\ln(x \pm \sqrt{x^2-1}))\big)$$
$$= \frac{1}{2}\big((x \pm \sqrt{x^2-1})^m + (x \pm \sqrt{x^2-1})^{-m}\big).$$

As $x + \sqrt{x^2-1} = 1/(x - \sqrt{x^2-1})$, it can be united as one formula as

$$\frac{1}{2}\big((x + \sqrt{x^2-1})^m + (x + \sqrt{x^2-1})^{-m}\big).$$

7. Substitute x as $1+\epsilon$ into the above formula, yielding[12]

$$\frac{1}{2}\big((1+\epsilon+\sqrt{\epsilon^2+2\epsilon})^m + (1+\epsilon+\sqrt{\epsilon^2+2\epsilon})^{-m}\big) \geq \frac{1}{2}(1+\sqrt{2\epsilon})^m \geq .5(1+m\sqrt{2\epsilon}).$$

Question 6.8. *Programming question. Skipped.*

Question 6.9. *Confirm that evaluating the formula in (6.47) by performing the matrix-vector multiplications from right to left is mathematically the same as Algorithm 6.13.*

Proof. We prove it step by step. First, according to Question 6.4.,

$$(Z^T \otimes Z^T) \cdot \text{vec}(h^2 F) = h^2 \text{vec}(ZFZ^T) = h^2 \text{vec}(Z^T F Z).^{13}$$

The first step is correct. Since

$$I \otimes \Lambda + \Lambda \otimes I = \text{diag}(\lambda_1 I + \Lambda, \dots, \lambda_n \Lambda),$$

it follows

$$(I \otimes \Lambda + \Lambda \otimes I)^{-1} = \text{diag}(\frac{1}{2\lambda_1}, \frac{1}{\lambda_1+\lambda_2}, \dots, \frac{1}{\lambda_2+\lambda_1}, \dots, \frac{1}{2\lambda_n}).$$

Consequently,

$$(I \otimes \Lambda + \Lambda \otimes I)^{-1} \text{vec}(h^2 F')(jn+i) = \frac{h^2 f'_{i,j+1}}{\lambda_{j+1}+\lambda_i}.$$

The second step is correct. Finally, use part IV of Question 6.4., again, yielding

$$\text{vec}(V) = (Z \otimes Z)\text{vec}(V') = \text{vec}(ZV'Z) = \text{vec}(ZV'Z^T).$$

□

Question 6.10. *The question is too long, so I omit it. For details, please refer to the textbook.*

Proof.

[12] The last one holds by Bernoulli inequality.

[13] $Z = Z^T$.

1. First, we prove that when two matrices commute they have at least one common eigenvector. Suppose λ is an eigenvalue of matrix A with corresponding eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_s$. Since $AB\boldsymbol{x}_i = BA\boldsymbol{x}_i = \lambda B\boldsymbol{x}_i$, each $B\boldsymbol{x}_i$ is in the eigenspace with value λ of A. Consequently, $B\boldsymbol{x}_i = \sum_{j=1}^{s} a_{ji}\boldsymbol{x}_j$. Or, in matrix form as

$$\left(B\boldsymbol{x}_1, \dots, B\boldsymbol{x}_s\right) = \left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_s\right) \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{s1} & a_{s2} & \dots & a_{ss} \end{pmatrix}.$$

Denote the right side matrix as S. It has at least one eigenvector $\boldsymbol{y}$ corresponding to μ. Thus,

$$B\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_s\right]\boldsymbol{y} = \left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_s\right] \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{s1} & a_{s2} & \dots & a_{ss} \end{bmatrix} \boldsymbol{y} = \mu\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_s\right]\boldsymbol{y}.$$

As $\boldsymbol{x}_i$ are linear independent, $\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_s\right]\boldsymbol{y} \neq \boldsymbol{0}$. Therefore, $\left[\boldsymbol{x}_1, \dots, \boldsymbol{x}_s\right]\boldsymbol{y}$ is a common eigenvector. Then, we prove the question by induction. The base case holds because it involves only scalar operations. Suppose it holds for $(n-1)$-by-$(n-1)$ matrices. Then, for n-by-n matrices, since A and B have one common unit eigenvector $\boldsymbol{x}$, expand it into an orthonormal matrix Q. We have

$$Q^T A Q = \begin{pmatrix} \lambda & \\ & A_{n-1} \end{pmatrix}, Q^T B Q = \begin{pmatrix} \mu & \\ & B_{n-1} \end{pmatrix}.$$

Since $Q^T A Q Q^T B Q = Q^T A B Q = Q^T B A Q - Q^T B Q Q^T A Q$, it follows that A_{n-1} and B_{n-1} commute. Thus, by induction, there exists $\tilde{Q}$ such that $\tilde{Q}^T A_{n-1}\tilde{Q} = \Lambda_{A_{n-1}}$, $\tilde{Q}^T B_{n-1}\tilde{Q} = \Lambda_{B_{n-1}}$. Then

$$Q^T \begin{pmatrix} 1 & \\ & \tilde{Q}^T \end{pmatrix} A \begin{pmatrix} 1 & \\ & \tilde{Q} \end{pmatrix} Q = \begin{pmatrix} \lambda & \\ & \Lambda_{B_{n-1}} \end{pmatrix}, Q^T \begin{pmatrix} 1 & \\ & \tilde{Q}^T \end{pmatrix} B \begin{pmatrix} 1 & \\ & \tilde{Q} \end{pmatrix} Q = \begin{pmatrix} \lambda & \\ & \Lambda_{B_{n-1}} \end{pmatrix}.$$

2. First, consider the most trivial case, when

$$\tilde{T} = \begin{pmatrix} & -1 & & \\ -1 & & \ddots & \\ & \ddots & & -1 \\ & & -1 & \end{pmatrix}.$$

Since $\tilde{T} = T_N - 2I$, by Question 4.5, the eigenvalues of $\tilde{T}$ are $2\cos(\frac{j\pi}{N+1})$ with corresponding eigenvector $\boldsymbol{z}_j = [\sin(\frac{j\pi}{N+1}), \dots, \sin(\frac{jN\pi}{N+1})]$. Then, as $\hat{T} = \alpha I - \theta\tilde{T}$, it follows that the eigenvalues of $\hat{T}$ are $\alpha - 2\theta\cos(\frac{j\pi}{N+1})$ with corresponding eigenvectors $\boldsymbol{z}_j = [\sin(\frac{j\pi}{N+1}), \dots, \sin(\frac{jN\pi}{N+1})]$.

3. Since $T = I \otimes A + \tilde{T} \otimes H$, it follows

$$\begin{aligned}(I \otimes Q)T(I \otimes Q^T) &= (I \otimes Q)(I \otimes A)(I \otimes Q^T) + (I \otimes Q)(\tilde{T} \otimes H)(I \otimes Q^T)\\ &= I \otimes QAQ^T + \tilde{T} \otimes QHQ^T = I \otimes \Lambda_A + \tilde{T} \otimes \Lambda_H.\end{aligned}$$

   By Question 6.5, the eigenvalues of $\tilde{T} \otimes \Lambda_H$ are $\lambda_{ij} = 2\cos(\frac{j\pi}{N+1})\theta_i$ with corresponding eigenvector $\boldsymbol{x}_{ij} = \boldsymbol{z}_i \otimes \boldsymbol{e}_j$. Then,

$$(I \otimes \Lambda_A)\boldsymbol{x}_{ij} = (I \otimes \Lambda_A)(\boldsymbol{z}_i \otimes \boldsymbol{e}_j) = \boldsymbol{z}_i \otimes (\alpha_j)\boldsymbol{e}_j = \alpha_j(\boldsymbol{z}_i \otimes \boldsymbol{e}_j) = \alpha_j\boldsymbol{x}_{ij}.$$

   Denote $X = [\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{nn}] = Z \otimes I$. It follows

$$(I \otimes \Lambda_A + \tilde{T} \otimes \Lambda_H)X = X\text{diag}(\alpha_1 + \lambda_{11}, \ldots, \alpha_n + \lambda_{nn}).$$

   The eigenvectors are $(I \otimes Q)(Z \otimes I) = Z \otimes Q$.

4. Denote $STS^{-1} = \Lambda$ as the eigendecomposition of T. Then, to solve $T\boldsymbol{x} = \boldsymbol{b}$ yields $\boldsymbol{x} = S^{-1}\Lambda^{-1}S\boldsymbol{b}$. As we have the explicit form of Λ and S, we claim by this method we can solve $T\boldsymbol{x} = \boldsymbol{b}$ in $O(n^3)$ while dense LU costs $O(n^6)$ and band LU costs $O(n^4)$.
   As Z is symmetric orthonormal and finding Q and Q^{-1} costs $O(n^3)$ time, if we can bound the matrix vector product $(Z \otimes Q)\boldsymbol{b}$ and $(Z^{-1} \otimes Q^{-1})\boldsymbol{c}$, the claim holds. Since the two matrix vector product is similar, we use $(Z \otimes Q)\boldsymbol{b}$ as illustration. Partition $\boldsymbol{b}$ by every n entries, and form it into a n-by-n matrix B. Then, the corresponding matrix vector product becomes matrix matrix product as

$$(Z \otimes Q)\boldsymbol{b} = (Z \otimes Q)\text{vec}(B) = \text{vec}(QBZ^T),$$

   which costs $O(n^3)$ time.

5. If A and B are symmetric tridiagonal Toeplitz matrices, $Q = Z$ and the cost of matrix matrix product can be reduced to $O(n^2 \log n)$ by fast sine transformation (or FFT). Thus, the total cost reduces to $O(n^2 \log n)$.

□

Question 6.11. *Suppose that R is upper triangular and nonsingular and that C is upper hessenberg. confirm that RCR^{-1} is upper Hessenberg.*

Proof. As the inverse of an upper triangular matrix is upper triangular, it suffices to verify multiplication by upper triangular preserve Hessenberg. Since $C(i,j) = 0$ when $i+1 > j$ and $R(i,j) = 0$ when $i > j$, it follows

$$RC(i,j) = \sum_{k=1}^{n} R(i,k)C(k,j) = \sum_{k=1}^{i-1} R(i,k)C(k,j) + \sum_{k=i}^{n} R(i,k)C(k,j) = 0, \text{ when } i > j+1,$$

$$CR(i,j) = \sum_{k=1}^{n} C(i,k)R(k,j) = \sum_{k=1}^{i-2} C(i,k)R(k,j) + \sum_{k=i-1}^{n} C(i,k)R(k,j) = 0, \text{ when } i > j+1.$$

□

Question 6.12. *Confirm that the Krylov subspace $\mathcal{K}_k(A, \boldsymbol{y}_1)$ has dimension k if and only if the Arnoldi algorithm (Algorithm 6.9) or the Lanczos algorithm (Algorithm 6.10) can compute $\boldsymbol{q}_k$ without quitting first.*

Proof. The proofs for Arnoldi algorithm and Lanczos algorithm are similar. Thus, it suffices to prove the result for Arnoldi algorithm. If we can prove that at ith step of Arnoldi algorithm, $\boldsymbol{z}_i = A\boldsymbol{q}_i$ lies in the space $\mathrm{span}(A[\boldsymbol{y}_1, \dots. \boldsymbol{y}_i])$. The question is proved, because the criteria of quitting is whether $h_{j+1,j} = 0$ occurs for any j, which equals to $\mathrm{span}(A[\boldsymbol{y}_1, \dots, \boldsymbol{y}_j]) \subset \mathrm{span}[\boldsymbol{y}_1, \dots, \boldsymbol{y}_j]$, i.e., Krylov subspace $\mathcal{K}_k(A, \boldsymbol{y}_1)$ has dimension less that k.
We prove $\boldsymbol{z}_i = A\boldsymbol{q}_i \in \mathrm{span}(A[\boldsymbol{y}_1, \dots. \boldsymbol{y}_{i+1}])$ by induction. For the base case, it holds because $\boldsymbol{q}_j = \boldsymbol{b}/\|\boldsymbol{b}\|_2$. If it holds for the $(j-1)$th step, we can conclude $\boldsymbol{q}_i \in [\boldsymbol{y}_1, \dots. \boldsymbol{y}_{j-1}]$ for all $i \le j$, since $\boldsymbol{q}_i = \boldsymbol{z}_{i-1} - \sum_{k=1}^{i-1} \boldsymbol{q}_k$. Then for the jth step, $\boldsymbol{z} = A\boldsymbol{q}_j \in \mathrm{span}(A[\boldsymbol{y}_1, \dots. \boldsymbol{y}_j])$. □

Question 6.13. *Confirm that when $A^{n\times n}$ is symmetric positive definite and $Q^{n\times k}$ has full column rank, then $T = Q^T AQ$ is also symmetric positive definite.*

Proof. For any vector $\boldsymbol{x} \in \mathbb{R}^k$, it follows

$$\boldsymbol{x}^T Q^T AQ\boldsymbol{x} = (Q\boldsymbol{x})^T AQ\boldsymbol{x} \ge 0.$$

As A is s.p.d, only when $Q\boldsymbol{x} = \boldsymbol{0}$ the equality holds. Then, since Q has full column rank, the only solution of $Q\boldsymbol{x} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$. Thus, T is also s.p.d. □

Question 6.14. *Prove Theorem 6.9.*

Remark. The Theorem has a little error. That is, $\hat{A} = M^{-1/2}AM^{1/2}$. The correct one should be $\hat{A} = M^{-1/2}AM^{-1/2}$

Proof. It suffices to prove the relationship bewtween the variables and we can do it by induction. For the base case, we can verify as follows:

$$\begin{aligned}
\hat{\boldsymbol{z}} &= \hat{A}\cdot\hat{\boldsymbol{p}}_1 = M^{-1/2}AM^{-1}\boldsymbol{b} = M^{-1/2}A\boldsymbol{p}_1 = M^{-1/2}\boldsymbol{z},\\
\hat{\nu}_1 &= (\hat{\boldsymbol{r}}_0^T\hat{\boldsymbol{r}}_0)/(\hat{\boldsymbol{p}}_1^T\hat{\boldsymbol{z}}) = (\boldsymbol{b}^T M^{-1}\boldsymbol{b})/(\boldsymbol{b}^T M^{-1}\boldsymbol{z}) = \nu_1,\\
\hat{\boldsymbol{x}}_1 &= \hat{\boldsymbol{x}}_0 + \hat{\nu}_1\hat{\boldsymbol{p}}_1 = \nu_1 M^{1/2}\boldsymbol{p}_1 = M^{1/2}\boldsymbol{x}_1,\\
\hat{\boldsymbol{r}}_1 &= \hat{\boldsymbol{r}}_{\mathbf{0}} - \hat{\nu}_1\hat{\boldsymbol{z}} = M^{-1/2}\boldsymbol{r}_0 - \nu_1 M^{-1/2}\boldsymbol{z} = M^{-1/2}\boldsymbol{r}_1,\\
\hat{\mu}_2 &= (\hat{\boldsymbol{r}}_1^T\hat{\boldsymbol{r}}_1)/(\hat{\boldsymbol{r}}_0^T\hat{\boldsymbol{r}}_0) = (\boldsymbol{r}_1^T M^{-1}\boldsymbol{r}_1)/(\boldsymbol{r}_0^T M^{-1}\boldsymbol{r}_0) = \mu_2,\\
\hat{\boldsymbol{p}}_2 &= \hat{\boldsymbol{r}}_1 + \hat{\mu}_2\hat{\boldsymbol{p}}_1 = M^{-1/2}\boldsymbol{r}_1 + M^{-1/2}\mu_2\boldsymbol{p}_1 = M^{1/2}\boldsymbol{p}_2.
\end{aligned}$$

Suppose it holds for $(k-1)$th step. Then, for kth step, it follows

$$\begin{aligned}
\hat{\boldsymbol{z}} &= \hat{A}\cdot\hat{\boldsymbol{p}}_k = M^{-1/2}A\boldsymbol{p}_k = M^{-1/2}\boldsymbol{z},\\
\hat{\nu}_k &= (\hat{\boldsymbol{r}}_{k-1}^T\hat{\boldsymbol{r}}_{k-1})/(\hat{\boldsymbol{p}}_k^T\hat{\boldsymbol{z}}) = (\boldsymbol{r}_{k-1}{}^T M^{-1}\boldsymbol{r}_{k-1})/(\boldsymbol{p}_k^T\boldsymbol{z}) = \nu_k,\\
\hat{\boldsymbol{x}}_k &= \hat{\boldsymbol{x}}_{k-1} + \hat{\nu}_k\hat{\boldsymbol{p}}_k = M^{1/2}\boldsymbol{x}_{k-1} + \nu_k M^{1/2}\boldsymbol{p}_k = M^{1/2}\boldsymbol{x}_k,\\
\hat{\boldsymbol{r}}_k &= \hat{\boldsymbol{r}}_{\boldsymbol{k-1}} - \hat{\nu}_k\hat{\boldsymbol{z}} = M^{-1/2}\boldsymbol{r}_{k-1} - \nu_k M^{-1/2}\boldsymbol{z} = M^{-1/2}\boldsymbol{r}_k,\\
\hat{\mu}_{k+1} &= (\hat{\boldsymbol{r}}_k^T\hat{\boldsymbol{r}}_k)/(\hat{\boldsymbol{r}}_{k-1}^T\hat{\boldsymbol{r}}_{k-1}) = (\boldsymbol{r}_k^T M^{-1}\boldsymbol{r}_k)/(\boldsymbol{r}_{k-1}^T M^{-1}\boldsymbol{r}_{k-1}) = \mu_{k+1},\\
\hat{\boldsymbol{p}}_{k+1} &= \hat{\boldsymbol{r}}_k + \hat{\mu}_{k+1}\hat{\boldsymbol{p}}_k = M^{-1/2}\boldsymbol{r}_k + M^{-1/2}\mu_{k+1}\boldsymbol{p}_k = M^{1/2}\boldsymbol{p}_{k+1}.
\end{aligned}$$

□