EN.553.480: Numerical Linear Algebra

NOTES

James Guo

Fall 2024

The aim of the course is to deal with matrices with *much more* entries, with computation of eigenvalues and eigenvectors. In particular, the cost of computing the *determinant* is large, *i.e.*, the characteristic polynomial has high degree. Therefore, we want *computability* and *closeness* of the eigenvalues.

Numerically, we want to evaluate the *deviation* of the value and the actual computation.

Contents

Acknowledgments:

- The Lecture Notes records the course contents for `EN.553.480` Numerical Linear Algebra instructed by *Dr. Mario Micheli* at *Johns Hopkins University* in the Fall 2024 semester.
- The notes summaries the lecture contents, notes, and adapted contents from the following source:
  - *Numerical Linear Algebra* by Lloyd N. Trefethen and David Bau, III.
- The notes is a summary of the lectures, and it might contain minor typos or errors. Please point out any notable error(s).

Best regards, James Guo.
December 2024.

Remarks about this Note:

While compiling this collection of notes, I have attempted to abide to the following conventions:

- Vectors will be represented as bold cases, such as $\mathbf{x}$ and $\mathbf{y}$.
- Matrices with m rows and n columns of entries in field $\mathbb{F}$ (denoted $\mathbb{F}^{m\times n}$) will be represented with capitalized letters, such as A, where a_{ij} represents the the entry of row i and column j.
- The *field* $\mathbb{F}$ of the R-module, unless otherwise specified, will be assumed to be $\mathbb{F} = \mathbb{C}$, otherwise, it would most often be $\mathbb{F} = \mathbb{R}$.
- For matrix multiplication (which is an *action*) $A.\mathbf{x}$, it is a linear map $f : \mathbb{C}^n \to C^m$ as $\mathbf{y} = A.\mathbf{x}$, which can be alternatively represented as $f : \mathbf{x} \mapsto A\mathbf{x}$.
- For null space and range, the notation standards were used in modern algebra, *i.e.*, naming them as kernels (ker) and image (im).
- sup and inf were used in generic case over max and min unless there is a clear evidence that maximum or minimum is well-defined. Readers shall only drop those notation by their own discretion.

While recording each theorem/proposition/definition/remark/..., I have tried to make a name that briefly describes what it is about, just like the conventions in Dr. Sheldon Axler's *Linear Algebra Done Right*.

I Preliminaries

I.1 Vector Spaces and Subspaces

The Numerical Linear Algebra has its foundations on linear algebra.
Note: Unless otherwise specified, the proofs of the theorems in this section is omitted, as they are assumed backgrounds in a typical linear algebra course.

Definition I.1.1. Linearity.

A function f is linear if:

$$\begin{cases} f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) \text{ for all } \mathbf{x}, \mathbf{y} \in C^n, \\ f(\alpha \mathbf{x}) = \alpha f(\mathbf{x}) \text{ for all } \alpha \in \mathbb{C} \text{ and } \mathbf{x} \in \mathbb{C}^n. \end{cases}$$

⌟

Theorem I.1.2. Linear Function as a Matrix.

If $f : \mathbb{C}^n \to \mathbb{C}^m$ is linear transformation, then there exists $A \in \mathbb{C}^{m\times n}$ such that $f(x) = A.\mathbf{x}$ for all $\mathbf{x} \in \mathbb{C}^n$.

Remark I.1.3. Linear Combinations.

For $A \in \mathbb{C}^{n\times m}$, we may write:

$$A = [a_{i,j}] = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_m}, \end{bmatrix}$$

and for $\mathbf{x} \in \mathbb{C}^n$:

$$A.\mathbf{x} = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_m} \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1\mathbf{A_1} + x_2\mathbf{A_2} + \cdots + x_n\mathbf{A_n}.$$

⌟

Definition I.1.4. Matrix Multiplication.

Let $A \in \mathbb{C}^{l\times m}$ and $C \in \mathbb{C}^{m\times n}$, their matrix multiplication is defined as:

$$[b_{ij}] = B = AC = [a_{ik}][c_{kj}],$$

in which the entry b_{ij} in B is:

$$b_{ij} = \sum_{k=1}^{m} a_{ik}c_{kj}.$$

⌟

Remark I.1.5. Multiplication of Matrices as Columns.

For the above multiplication that $B = AC$, if we write B as vectors we have:

$$\begin{bmatrix} \mathbf{B_1} & \mathbf{B_2} & \cdots & \mathbf{B_n} \end{bmatrix} = A \begin{bmatrix} \mathbf{C_1} & \mathbf{C_2} & \cdots & \mathbf{C_n} \end{bmatrix} = \begin{bmatrix} A.\mathbf{C_1} & A.\mathbf{C_2} & \cdots & A.\mathbf{C_n} \end{bmatrix},$$

hence:

$$\mathbf{B_j} = A.\mathbf{C_j} = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_m} \end{bmatrix} . \begin{bmatrix} c_{1j} \\ c_{2j} \\ \vdots \\ c_{mj} \end{bmatrix},$$

hence each column of $B = AC$ is a linear combination of the columns of A.

⌟

Definition I.1.6. Inner and Outer Product.

Let $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$, their inner product (or dot product) is:

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^{\mathsf{T}}.\overline{\mathbf{v}} = u_1\overline{v_1} + u_2\overline{v_2} + \cdots + u_n\overline{v_n} \in \mathbb{C},$$

where as the outer product is:

$$\mathbf{u}.\mathbf{v}^{\mathsf{T}} = \begin{bmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ u_2v_1 & u_2v_2 & \cdots & u_2v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nv_1 & u_nv_2 & \cdots & u_nv_n \end{bmatrix} = \begin{bmatrix} v_1\mathbf{u} & v_2\mathbf{u} & \cdots & v_n\mathbf{u} \end{bmatrix}.$$

⌟

Definition I.1.7. Range and Null Space.

For $A \in \mathbb{C}^{m\times n}$, its range (or image) is:

$$\operatorname{im} A = \{\mathbf{y} : \mathbf{y} = A.\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{C}^n\} = \operatorname{span}\{\mathbf{A_1}, \mathbf{A_2}, \cdots, \mathbf{A_n}\} \subset \mathbb{C}^m.$$

Its null space (or kernel) is:

$$\ker A = \{\mathbf{x} : A.\mathbf{x} = \mathbf{0}\} \subset \mathbb{C}^n.$$

⌟

Theorem I.1.8. Rank-Nullity Theorem (Fundamental Theorem of Linear Maps).

For $A \in \mathbb{C}^{m\times n}$, the dimension of the range (or rank) and the dimension of the null space (or nullity) follows:

$$\dim(\operatorname{im} A) + \dim(\ker A) = n.$$

Moreover, for the transpose of A, we have:

$$\dim(\operatorname{im} A^{\mathsf{T}}) = \dim(\in A).$$

Proposition I.1.9. Properties on the Rank of A.

Assume that $A \in \mathbb{C}^{m\times n}$, the following holds:

(i) $\dim(\operatorname{im} A) \leqslant n$.

(ii) $\dim(\operatorname{im} A) \leqslant m$.

The result (ii) is a direct result of (i) and Rank-Nullity for transpose of A. Moreover, this implies that $\dim(\operatorname{im} A) \leqslant \min\{m, n\}$.

Definition I.1.10. Full Rank.

For $A \in \mathbb{C}^{m\times n}$, A is full rank if $\dim(\operatorname{im} A) = \min\{m, n\}$. ⌟

An example of a full rank matrix is $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$.

Definition I.1.11. Nonsingular Square Matrix.

A m-by-m square matrix A is nonsingular if $A.\mathbf{x} = \mathbf{0}$ has unique solution $\mathbf{x} = \mathbf{0}$. ⌟

Proposition I.1.12. Equivalent Facts for Square Matrices.

The following are equivalent:

(i) A is nonsingular,

(ii) A is invertible, *i.e.*, there exists A^{-1} such that $A^{-1}A = AA^{-1} = \mathrm{Id}$,

(iii) For all $\mathbf{b} \in \mathbb{C}^m$, $A.\mathbf{x} = \mathbf{b}$ has unique solution,

(iv) Columns of A are linearly independent,

(v) $\dim(\operatorname{im} A) = m$,

(vi) $\operatorname{im} A = \mathbb{C}^m$,

(vii) $\dim(\ker A) = 0$,

(viii) $\ker A = \{\mathbf{0}\}$,

(ix) $\det A \neq 0$,

(x) A is an isomorphism.

In particular, since A being nonsingular is the equivalent of being injective for square matrices, it is an isomorphism, hence all (ii) to (x) are equivalent to (i).

Remark I.1.13. Coordinates in $\mathbb{C}^m$.

There are many basis in $\mathbb{C}^m$:

(i) The Canonical Basis:

$$\alpha = \{\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_n}\},$$

where:

$$\mathbf{e_1} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{e_2} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \mathbf{e_1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Hence, for any $\mathbf{b} \in \mathbb{C}^m$, we have:

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = b_1\mathbf{e_1} + b_2\mathbf{e_2} + \cdots + b_m\mathbf{e_m}.$$

(ii) Another Basis (Arbitrary):

$$\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\}.$$

Here, for any $\mathbf{b} \in \mathbb{C}^m$, we are looking forward to having:

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = c_1\mathbf{v_1} + c_2\mathbf{v_2} + \cdots + c_m\mathbf{v_m} = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_m} \end{bmatrix}.\mathbf{c},$$

and since the matrix is invertible, it leads to that:

$$\mathbf{c} = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_m} \end{bmatrix}^{-1}.\mathbf{b},$$

in which by solving for $\mathbf{c}$ here, we have completed the decomposition.

⌟

Remark I.1.14. Remarks on Complex Numbers.

For $z \in \mathbb{C} = \mathbb{R}(\mathrm{i})$ in which $\mathrm{i}^2 = -1$, we may represent $z = a + \mathrm{i}b$, where $a, b \in \mathbb{R}$. Thus, we can visualize the real and imaginary parts, respectively.

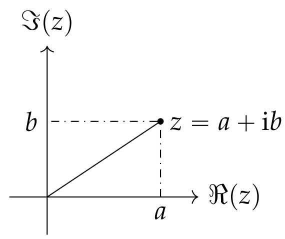


Figure I.1. Real and Complex Plane.

Here, we have the magnitude as:

$$|z| = \sqrt{a^2 + b^2},$$

and we have the complex conjugate as:

$$\bar{z} = a - \mathrm{i}b.$$

⌟

Example I.1.15. Matrix Operations.

Let $A \in \mathbb{C}^{m \times n}$, that is:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & a_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

its transpose is:

$$A^\intercal = \begin{bmatrix} a_{11} & a_{21} & \cdots \\ a_{12} & a_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

its complex conjugate is:

$$\overline{A} = \begin{bmatrix} \overline{a_{11}} & \overline{a_{12}} & \cdots \\ \overline{a_{21}} & \overline{a_{22}} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

and the Hermitian conjugate is:

$$A^* = (\overline{A})^\intercal.$$

⌟

In $\mathbb{R}^n$, the dot product for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, we have:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^\intercal.\mathbf{y} = \sum_{k=1}^{m} x_k y_k.$$

In particular, we can define the norm as:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^\intercal.\mathbf{x}} = \sqrt{\sum_{k=1}^{m} x_k^2}.$$

In $\mathbb{C}^n$, *i.e.*, when $\mathbb{F} = \mathbb{C}$, and the inner product for $\mathbf{x}, \mathbf{y} \in \mathbb{C}^m$, we have:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^*.\mathbf{y} = \sum_{k=1}^{m} \overline{x_k} y_k.$$

Again, we have the norm as:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{k=1}^{m} \overline{x_k} x_k} = \sqrt{\sum_{k=1}^{m} |x_k|^2},$$

since we have $\overline{x_k} x_k = |x_k|^2$, know as the *modulus*.

In particular $\mathbf{x} = \mathbf{0}$ if and only if $\|\mathbf{x}\| = 0$.

Proposition I.1.16. Distributivity of Matrix Operator.

For any $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{n \times p}$, we have:

$$(AB)^\intercal = B^\intercal A^\intercal \text{ and } (AB)^* = B^* A^*.$$

Definition I.1.17. Orthogonality.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^m$, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\mathbf{x}^*.\mathbf{y} = 0$.
For sets $S = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\} \subset \mathbb{C}^m$ is orthogonal set if the vectors are pairwise orthogonal, *i.e.*, $\mathbf{v_i}^*.\mathbf{v_j} = 0$ for all $i \neq j$.

⌟

Theorem I.1.18. Orthogonality $\implies$ Linear Independence.

If $S = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\} \subset \mathbb{C}^m$ is a orthogonal set of *nonzero* vectors, then S is linearly independent.

Proof. Here, we let λ_k's be set such that:

$$\sum_{k=1}^{m} \lambda_k \mathbf{v_k} = \mathbf{0}.$$

For all $1 \leqslant i \leqslant m$ the inner product with $\mathbf{v_i}$, giving us that:

$$\sum_{k=1}^{m} \lambda_k \mathbf{v_k}^* \mathbf{v_i} = \mathbf{0}^* \mathbf{v_1} = 0.$$

Thus, by orthogonality and the nonzero vectors, we have:

$$\lambda_i = 0 \text{ for all } 1 \leqslant i \leqslant m,$$

which enforces S to be linearly independent. □

Corollary I.1.19. Surjective + Orthogonal $\Longrightarrow$ Basis.

Suppose that $S = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\} \subset V$ is a basis in which $V = \text{span}\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\}$, then S is a orthogonal basis of V.

Definition I.1.20. Orthogonal Basis.

An orthogonal basis can be $\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_r}\} \subset \mathbb{C}^m$ with property that $\mathbf{v_i}^*.\mathbf{v_j} = 0$ for all $i \neq j$.
In particular, we have $V = \text{span}(\beta)$ as a subspace of $\mathbb{C}^m$. ⌟

Theorem I.1.21. Orthogonal Projection.

Given an orthogonal basis:

$$\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_r}\} \subset \mathbb{C}^m$$

of $V = \text{span}(\beta)$ and a vector $\mathbf{b} \in V$, we have the coordinates of $\mathbf{b}$ with respect to β, *i.e.*, the unique scalars $c_1, \cdots, c_r$ such that $\mathbf{b} = \sum_{k=1}^{r} c_k \mathbf{v_k}$, are:

$$c_k = \frac{\mathbf{v_k}^*\mathbf{b}}{\|\mathbf{v_k}\|^2} \text{ for all } 1 \leqslant k \leqslant r.$$

Proof. For all $1 \leqslant i \leqslant m$, we take the inner product of the linear combinations with $\mathbf{v_i}^*$, we have:

$$\mathbf{v_i}^*\mathbf{b} = \sum_{k=1}^{r} c_k \mathbf{v_i}^*.\mathbf{v_k} = c_i \|\mathbf{v_i}\|^2,$$

as desired. □

Definition I.1.22. Orthonormal Basis.

A basis $\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_r}\} \subset \mathbb{C}^m$ is *orthonormal* if:

(i) β is orthogonal, and

(ii) each vector in β is unit, *i.e.*, $\|\mathbf{v_i}\| = 1$.

⌟

In the orthonormal basis, the orthogonal projections will, in turn, be:

$$\mathbf{b} = \sum_{k=1}^{r} (\mathbf{v_k}^*.\mathbf{b})\mathbf{v_i}.$$

Definition I.1.23. Orthogonal Matrix with $\mathbb{F} = \mathbb{R}$.

A matrix $A \in \mathbb{R}^{m\times m}$ is called orthogonal if $A^\intercal = A^{-1}$, *i.e.*, $A^\intercal A = \text{Id}_m$ and $AA^\intercal = \text{Id}_m$. ⌟

This can be similarly defined in matrices with complex entries.

Definition I.1.24. Unitary Matrix with $\mathbb{F} = \mathbb{C}$.

A matrix $Q \in \mathbb{C}^{m\times m}$ is called *unitary* if $Q^* = Q^{-1}$, *i.e.*, $Q^*Q = \mathrm{Id}_m$ and $QQ^* = \mathrm{Id}_m$. ⌟

Theorem I.1.25. Equivalences with Unitary.

For a matrix $Q \in \mathbb{C}^{m\times m}$, the following conditions are the equivalent:

(i) Q is unitary, *i.e.*, $Q^*Q = \mathrm{Id}_m$,

(ii) The columns of $Q = \begin{bmatrix}\mathbf{Q_1} & \mathbf{Q_2} & \cdots & \mathbf{Q_m}\end{bmatrix}$ are orthonormal, *i.e.*:
$$\|\mathbf{Q_i}\| = 1 \text{ and } \mathbf{Q_i}^*\mathbf{Q_j} = 0 \text{ for all } i \neq j,$$

(iii) For all $\mathbf{x} \in \mathbb{C}^m$, we have $\|Q.\mathbf{x}\| = \|\mathbf{x}\|$, *i.e.*, the action $f : \mathbf{x} \mapsto Q.\mathbf{x}$ is an *isometry*.

Proof. (i) $\Longrightarrow$ (ii): Suppose that $Q^*Q = \mathrm{Id}_m$, then we have:
$$\begin{bmatrix}\mathbf{Q_1}^*\\ \mathbf{Q_2}^*\\ \vdots\\ \mathbf{Q_m}^*\end{bmatrix}\cdot\begin{bmatrix}\mathbf{Q_1} & \mathbf{Q_2} & \cdots & \mathbf{Q_m}\end{bmatrix} = \begin{bmatrix}\mathbf{Q_1}^*\mathbf{Q_1} & \mathbf{Q_1}^*\mathbf{Q_2} & \cdots & \mathbf{Q_1}^*\mathbf{Q_m}\\ \mathbf{Q_2}^*\mathbf{Q_1} & \mathbf{Q_2}^*\mathbf{Q_2} & \cdots & \mathbf{Q_2}^*\mathbf{Q_m}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{Q_m}^*\mathbf{Q_1} & \mathbf{Q_m}^*\mathbf{Q_2} & \cdots & \mathbf{Q_m}^*\mathbf{Q_m}\end{bmatrix} = \begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix},$$
which follows along with (ii).
(i) $\Longrightarrow$ (iii): Still, suppose that $Q^*Q = \mathrm{Id}_m$, then for all $\mathbf{x} \in \mathbb{C}^m$, we have:
$$\|Q.\mathbf{x}\|^2 = (Q.\mathbf{x})^*(Q.\mathbf{x}) = \mathbf{x}^*Q^*Q\mathbf{x} = \mathbf{x}^*(Q^*Q)\mathbf{x} = \mathbf{x}^*\,\mathrm{Id}_m\,\mathbf{x} = \mathbf{x} * \mathbf{x} = \|\mathbf{x}\|^2,$$
as desired. □

Suppose that we want to solve that $Q.\mathbf{x} = \mathbf{b}$, where Q is unitary. If we denote:
$$Q = \begin{bmatrix}\mathbf{Q_1} & \mathbf{Q_2} & \cdots & \mathbf{Q_m}\end{bmatrix},$$
we have that:
$$x_1\mathbf{Q_1} + x_2\mathbf{Q_2} + \cdots + x_m\mathbf{Q_m} = \mathbf{b},$$
so we have:
$$\mathbf{x} = Q^*Q.\mathbf{x} = Q^*.\mathbf{b}.$$

I.2 Normed Vector Space

Definition I.2.1. Euclidean Length.

Let $\mathbf{x} \in \mathbb{C}^m$, the Euclidean length of $\mathbf{x}$ is:
$$\|\mathbf{x}\| = \sqrt{\mathbf{x} * .\mathbf{x}} = \sqrt{\sum_{k=1}^{m} |x_k|^2}$$
⌟

Definition I.2.2. Norm.

A norm is a function $\|\bullet\| : \mathbb{C}^n \to \mathbb{R}$ with the following properties for all $\mathbf{x}, \mathbf{y} \in \mathbb{C}^m$ and $\alpha \in \mathbb{C}$:

(i) Positivity: $\|\mathbf{x}\| \geqslant 0$,

(ii) Definiteness: $\mathbf{x} = 0 \iff \|\mathbf{x}\| = 0$,

(iii) Triangle Inequality: $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$, and

(iv) Homogeneity: $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$.

Example I.2.3. Examples of Norms.

The 1-norm is $\|\mathbf{x}\|_1 \sum_{k=1}^{m} |x_k|$.
The p-norm (where $1 \leqslant p < \infty$) is $\|\mathbf{x}\|_p = \left(\sum_{k=1}^{m} |x_k|^p\right)^{1/p}$.
The ∞-norm is $\|x\|_\infty = \max_{1\leqslant k\leqslant m} |x_i|$.

For a fixed norm, we have unit sphere:

$$S = \{\mathbf{x} : \|\mathbf{x}\| = 1\},$$

and the unit ball as:

$$B = \{\mathbf{x} : \|\mathbf{x}\| \leqslant 1\}.$$

Example I.2.4. 2-norm (Euclidean Norm) in $\mathbb{R}^2$.

The unite sphere of 2-norm would satisfy that $\|\mathbf{x}\|_2 = 1$, or equivalently, $\|\mathbf{x}\|^2 = 1$, hence equivalent to $x_1^2 + x_2^2 = 1$ since we are in the real field. Thus it is the unit circle.

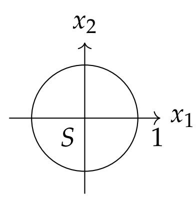


Figure I.2. Unit Sphere in 2-norm.

Example I.2.5. 1-norm in $\mathbb{R}^2$.

The unite sphere of 1-norm would satisfy that $\|\mathbf{x}\|_1 = 1$, or equivalently, $|x_1| + |x_2| = 1$ since we are in the real field. Thus it is the diamonds.

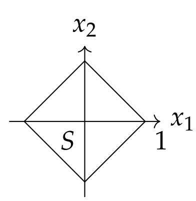


Figure I.3. Unit Sphere in 1-norm.

Example I.2.6. ∞-norm in $\mathbb{R}^2$.

The unite sphere of ∞-norm would satisfy that $\|\mathbf{x}\|_\infty = 1$, or equivalently, $\max\{|x_1|, |x_2|\} = 1$ since we are in the real field. Thus it is the diamonds.

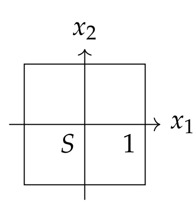


Figure I.4. Unit Sphere in ∞-norm.

Remark I.2.7. Note on Invalid p-norms.

When $0 < p < 1$, $\| \bullet \|_p$ is no longer a norm as it violates triangle inequality.

Example I.2.8. Calculation on Norms.

Let $\mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \in \mathbb{C}^2$, we have the norms as:

$$\begin{aligned} \|\mathbf{x}\|_1 &= |2| + |1| = 3, \\ \|\mathbf{x}\|_2 &= \sqrt{|2|^2 + |1|^2} = \sqrt{5}, \\ \|\mathbf{x}\|_\infty &= \max\{|2|, |1|\} = 2. \end{aligned}$$

Proposition I.2.9. Monotonicity of Norms.

For any $\mathbf{x} \in \mathbb{C}^n$, $\|\mathbf{x}\|_1 \geqslant \|\mathbf{x}\|_2 \geqslant \|\mathbf{x}\|_\infty$.

Remark I.2.10. Stretch of a Norm.

Let $A \in \mathbb{C}^{m\times n}$, consider the function $f : \mathbb{C}^n \to \mathbb{C}^m$ such that $\mathbf{x} \mapsto A.\mathbf{x}$. The stretch of $\mathbf{x}neq\mathbf{0}$ caused by multiplication by A is:

$$\frac{\|A.\mathbf{x}\|_*}{\|\mathbf{x}\|_*},$$

where the norms ($\| \bullet \|_*$) are identical.

Definition I.2.11. Matrix Norms.

Let $A \in \mathbb{C}^{m\times n}$, the matrix norm induced by a vector norm is:

$$\|A\|_{M^{m\times n}} := \sup_{\substack{\mathbf{x}\in\mathbb{C}^n \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|A.\mathbf{x}\|_*}{\|\mathbf{x}\|_*}.$$

Since we can restrict the search on the unit circle, which is *compact*, we can use max instead of the sup.

Proof. Let $\mathbf{x} \in \mathbb{C}^n$ be nonzero and arbitrary, we want to show that the stretch of $\mathbf{x}$ and $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$, that is $\mathbf{x} = \|\mathbf{x}\|\mathbf{u}$, hence we have:

$$\frac{\|A.xx\|}{\|\mathbf{x}\|} = \frac{\|A.(\|\mathbf{x}\|\mathbf{u})\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{x}A.\mathbf{u}\|}{\|\mathbf{x}\|} = \|\|A.\mathbf{u}\| = \frac{\|A.\mathbf{u}\|}{\|\mathbf{u}\|},$$

which implies that they are equivalent. $\square$

Theorem I.2.12. Equivalence of Matrix Norm.

Let $A \in \mathbb{C}^{m\times n}$, the matrix norm can be computed as:

$$\|A\|_{M^{m\times n}} = \max_{\substack{\mathbf{x}\in\mathbb{C}^n \\ \|\mathbf{x}\|_*=1}} \|A.\mathbf{x}\|_*.$$

Example I.2.13. Computing Matrix Norm.

Let $A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}$, we may compute its matrix norm induced by the vector norms for norm $\| \bullet \|_1$, as:

$$\|A\|_1 = \max_{\substack{\mathbf{x} \in \mathbb{C}^n \\ \|\mathbf{x}\|=1}} \|A.\mathbf{u}\|_1,$$

where for all $\mathbf{u} = \begin{bmatrix} u_1 & u_2 \end{bmatrix}^\intercal$, we have $A.\mathbf{u} = u_1\mathbf{A_1} + u_2\mathbf{A_2}$. Since $\|u\|_1 = 1$, we have $|u_1| + |u_2| = 1$. Consider $A.\mathbf{e_1} = (1, 0)$ and $A.\mathbf{e_2} = (2, 2)$, we may demonstrate the transformation as:

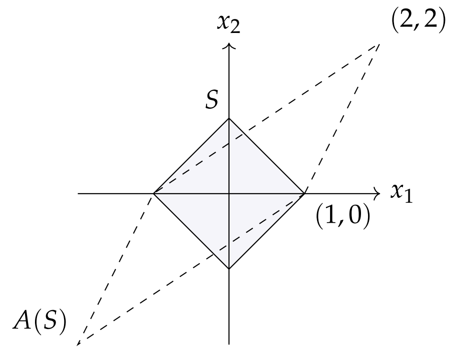


Figure I.5. Transformation of Unit Sphere in 1*-norm with Matrix A.*

Notice that the maximum of the 1-norm after transformation A is, in fact, $\|A.(0, 1)\|_1 = \|(2, 2)\|_1 = |2| + |2| = 4$. Hence, we have $\|A\| = 4$.

Note that the same computation can be done with 2-norms or other norms, however, the computation will be more lengthy and complicated. For example, $\|A\|_2 \approx 2.9308$ and $\|A\|_\infty = 3$.

Theorem I.2.14. 1**-norm of Matrix is Maximum of Norm of Vectors.**

For any $A \in \mathbb{C}^{m \times n}$, where $A = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_n} \end{bmatrix}$. The 1-norm of A is given by:

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \|\mathbf{A_j}\|_1 = \max_{1 \leqslant j \leqslant n} \left(\sum_{k=1}^{m} |a_{kj}| \right).$$

Observe that this conclusion aligns with the above example, where $\|\mathbf{A_1}\| = 1$ and $\|\mathbf{A_2}\| = 4$, so $\|A\|_1 = 4$.

Proof. Suppose $A \in \mathbb{C}^{m \times n}$, $A = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_n}, \end{bmatrix}$, and $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}^\intercal$ such that $\|\mathbf{u}\|_1 = 1$, hence we can represent:

$$A.\mathbf{u} = u_1\mathbf{A_1} + u_2\mathbf{A_2} + \cdots + u_n\mathbf{A_n}.$$

When we take the norm for both sides, we have:

$$\|A.\mathbf{u}\| = \|u_1\mathbf{A_1} + u_2\mathbf{A_2} + \cdots + u_n\mathbf{A_n}\|.$$

Then, by triangle inequality and properties of norm, we can have:

$$\begin{aligned} \|A.\mathbf{u}\| &\leqslant \|u_1\mathbf{A_1}\| + \|u_2\mathbf{A_2}\| + \cdots + \|u_n\mathbf{A_n}\| \\ &= |u_1|\|\mathbf{A_1}\| + |u_2|\|\mathbf{A_2}\| + \cdots + |u_n|\|\mathbf{A_n}\| \\ &\leqslant \max_{1 \leqslant j \leqslant n} \|\mathbf{A_j}\|. \end{aligned}$$

Hence, for all $\mathbf{u}$, we have:

$$\|A.\mathbf{u}\| \leqslant \max_{1 \leqslant j \leqslant n} \|\mathbf{A_j}\|_1.$$

Note that the above equality holds when $\mathbf{u} = \mathbf{e_{j*}}$ where j^* is the largest where $\|A.\mathbf{A_{j*}}\|$ is the largest of $\|\mathbf{A_1}\|, \cdots, \|\mathbf{A_n}\|$. Hence, the maximum over all the unit vectors of $A.\mathbf{u}$ is equal to the maximum of the magnitude of the columns, hence the equality holds trivially. □

Theorem I.2.15. ∞-norm of Matrix.

For any $A \in \mathbb{C}^{m\times n}$, where $A = \begin{bmatrix}\mathbf{R_1} & \mathbf{R_2} & \cdots & \mathbf{R_m}\end{bmatrix}^\mathsf{T}$. The ∞-norm of A is given by:

$$\|A\|_\infty = \max_{1\leqslant k\leqslant m} \|\mathbf{R_k}^*\|_1.$$

Example I.2.16. Calculation of Matrix Norm.

Let $A = \begin{bmatrix}1 & 2\\ 0 & 2\end{bmatrix}$, note that:

$$\begin{aligned}
\|\mathbf{A_1}\| &= |1| + |0| = 1,\\
\|\mathbf{A_2}\| &= |0| + |2| = 2,\\
\|\mathbf{R_1}^*\| &= |1| + |2| = 3,\\
\|\mathbf{R_2}^*\| &= |0| + |2| = 2.
\end{aligned}$$

Hence, $\|A\|_1 = 4$ and $\|A\|_\infty = 3$. ⌟

Theorem I.2.17. Cauchy-Schwarz Inequality.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, we have:

$$|\mathbf{x}^*\mathbf{y}| \leqslant \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2,$$

and the equality holds if and only if they are scalar multiples of each other, *i.e.*, parallel.

Recall that in $\mathbb{R}^2$ or $\mathbb{R}^3$, and any $\mathbf{x}, \mathbf{y}$ in the space, we have:

$$\mathbf{x}\cdot\mathbf{y} = \mathbf{x}^\mathsf{T}\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta.$$

Hence, in such case:

$$-1 \leqslant \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \leqslant 1.$$

Theorem I.2.18. Hölder's Inequality.

For any p and q such that:

$$\frac{1}{p} + \frac{1}{q} = 1,$$

which is called *harmonic conjugates*, then:

$$|\mathbf{x}^*\mathbf{y}| \leqslant \|\mathbf{x}\|_p\|\mathbf{y}\|_q.$$

Proposition I.2.19. Matrix Norm for Outer Product.

For any $\mathbf{x} \in \mathbb{C}^m$, $y \in \mathbb{C}^n$, and the matrix $A = \mathbf{x}\mathbf{y}^* \in \mathbb{C}^{m \times n}$. We have:

$$\|A\|_2 = \|\mathbf{x}\|\|y\|.$$

Proof. Note that by the equivalent definition of norms:

$$\|A\| = \max_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{C}^n}} \|A.\mathbf{u}\|.$$

Note that by associativity:

$$\|A.\mathbf{u}\| = \|(\mathbf{x}\mathbf{y}^*)\mathbf{u}\| = \|\mathbf{x}(\mathbf{y}^*\mathbf{u})\| = \|(\mathbf{y}^*\mathbf{u})\mathbf{x}\| = |\mathbf{y}^*\mathbf{u}|\|\mathbf{x}\|,$$

and by Cauchy-Schwarz, we have:

$$\|A.\mathbf{u}\| = |\mathbf{y}^*\mathbf{u}|\|\mathbf{x}\| \leqslant \|\mathbf{y}\|\|\mathbf{u}\|\|\mathbf{x}\|,$$

with equality when $\mathbf{u}$ is parallel to $\mathbf{y}$, thus:

$$\|A\| = \max_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{C}^n}} \|A.\mathbf{u}\| = \|\mathbf{x}\|\|\mathbf{y}\|\|\mathbf{u}\| = \|\mathbf{x}\|\|\mathbf{y}\|.$$

□

Theorem I.2.20. Inequality for Matrix and Vector Norm.

For any vector norm and induced matrix norm, and for all $A \in \mathbb{C}^{m \times n}$ and $\mathbf{x} \in \mathbb{C}^n$, we have:

$$\|A.\mathbf{x}\| \leqslant \|A\|_M\|\mathbf{x}\|.$$

Proof. By definition:

$$\|A\| = \max_{\substack{\mathbf{x}\in\mathbb{C}^m \\ \|\mathbf{x}\|\neq 0}} \frac{\|A.\mathbf{x}\|}{\|\mathbf{x}\|},$$

hence for eneric $\mathbf{x} \in \mathbb{C}^m$, we have:

$$\|A\| \geqslant \frac{\|A.\mathbf{x}\|}{\|\mathbf{x}\|},$$

hence implying that $\|A.\mathbf{x}\| \leqslant \|A\|\|\mathbf{x}\|$. □

Corollary I.2.21. Inequality for Matrix Multiplication.

For any matrix norm induced by vector norm, and for all $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{n \times p}$, then:

$$\|AB\| \leqslant \|A\|\|B\|$$

Proof. Note that:

$$\|AB\| = \max_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{C}^p}} \|AB.\mathbf{u}\|.$$

For the inequality:

$$\|AB.\mathbf{u}\| \leqslant \|A\|\|B.\mathbf{u}\| \leqslant \|A\|\|B\|\|\mathbf{u}\|$$

Taking when the equality holds (by Cauchy-Schwarz), the equality holds. □

Remark I.2.22. Recalling Axioms of Normed Vector Space.

For all $A, B \in \mathbb{C}^{m\times n}$, for all $\alpha \in \mathbb{C}$, the following must hold:

(i) Positivity and definiteness: $\|A\| \geqslant 0$ and $\|A\| = 0$ if and only if $A = 0$,

(ii) Triangular inequality: $\|A + B\| \leqslant \|A\| + \|B\|$, and

(iii) Homogeneity: $\|\alpha A\| = |\alpha|\|A\|$.

⌟

Definition I.2.23. Forbenius norm.

Let $A \in \mathbb{C}^{m\times n}$, its Forbenius norm is defined to be:

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{m}|a_{i,j}^2|\right)^{1/2}.$$

⌟

Theorem I.2.24. Equivalent Definition of Forbenius Norm.

Let $A \in \mathbb{C}^{m\times n}$, its Forbenius norm is:

$$\|A\|_F = \sqrt{\mathrm{Tr}(A^*A)}.$$

Proof. It is easy to find that the diagonals of A^*A are $\mathbf{A_i}^*\mathbf{A_i}$, and since the trace is the sum of the squares, then:

$$\mathrm{Tr}(A^*A) = \sum_{i=1}^{n}\|\mathbf{A_i}\|^2 = \sum_{i=1}^{n}|a_{i,1}|^2 + \cdots + \sum_{i=1}^{n}|a_{i,m}|^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}|a_{i,j}|^2,$$

and the sums can be switched since it is finite sum. □

Example I.2.25. Isometry.

The following matrices are isometric:

(i) Rotation matrices,

(ii) Reflection matrices, and

(iii) Permutation of coordinates.

⌟

Proposition I.2.26. Invariant of Unitary Matrix.

If $Q \in \mathbb{C}^{m\times n}$ is unitary, and $P \in \mathbb{C}^{n\times m}$, then for all $A \in \mathbb{C}^{m\times n}$, we have:

(i) Left unit: $\|QA\|_2 = \|A\|_2$ and $\|QA\|_F = \|A\|_F$.

(ii) Right unit: $\|AP\|_2 = \|A\|_2$ and $\|AP\|_F = \|A\|_F$.

Proof. (i) By definition:

$$\|QA\|_2 = \max_{\substack{\mathbf{u}\in\mathbb{C}^n \\ \|\mathbf{u}\|_2=1}}\|QA.\mathbf{u}\|_2 = \max_{\substack{\mathbf{u}\in\mathbb{C}^n \\ \|\mathbf{u}\|_2=1}}\|Q.(A.\mathbf{u})\|_2 = \max_{\substack{\mathbf{u}\in\mathbb{C}^n \\ \|\mathbf{u}\|_2=1}}\|A.\mathbf{u}\|_2 = \|A\|_2.$$

For Forbenius norm, we have:

$$\|QA\|_F = \sqrt{\operatorname{Tr}\left((QA)^*(QA)\right)} = \sqrt{\operatorname{Tr}(A^*Q^*QA)} = \sqrt{\operatorname{Tr}(A^*A)} = \|A\|_F.$$

(ii) For the right unitary:

$$\|AP\|_2 = \max_{\substack{\mathbf{u}\in\mathbb{C}^n \\ \|\mathbf{u}\|_2=1}} \|AP.\mathbf{u}\|_2 = \max_{\substack{\mathbf{u}\in\mathbb{C}^n \\ \|\mathbf{u}\|_2=1}} \|A.\mathbf{v}\|_2,$$

where $\mathbf{v} \in \mathbb{C}^m$ and $\|\mathbf{v}\|_2 = \|P\mathbf{u}\|_2 = 1$. Note that the map is valid since P is isometry, hence it is invertible, thus $S \to S$ is invertible, hence, we have the above to be $\|A\|_2$.
For Forbenius norm, we have:

$$\|AP\|_F = \sqrt{\operatorname{Tr}\left((AP)^*(AP)\right)} = \sqrt{\operatorname{Tr}(P^*A^*AP)} = \sqrt{\operatorname{Tr}(PP^*A^*A)} = \sqrt{\operatorname{Tr}(A^*A)} = \|A\|_F. \qquad \square$$

I.3 Eigenspace

Definition I.3.1. Eigenvalue.

Let $A \in \mathbb{C}^{m\times m}$ be a square matrix, $\mathbf{x}$ is an eigenvector of A with associate eigenvalue $\lambda \in \mathbb{C}$ such that:

$$\mathbf{x} \neq \mathbf{0} \text{ and } A.\mathbf{x} = \lambda\mathbf{x}.$$

⌟

Example I.3.2. Finding Eigenvalue and Eigenvector.

Let $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, we may notice that for $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, we have:

$$A.\mathbf{x} = 2\mathbf{x}.$$

Else, for the rotation matrix $A = \begin{bmatrix} \cos\pi/4 & -\sin\pi/4 \\ \sin\pi/4 & \cos\pi/4 \end{bmatrix}$, and note that there is no real eigenvalues, the complex eigenvalues are:

$$\lambda = e^{\pm i\pi/4}.$$

⌟

Theorem I.3.3. Singular $\iff 0$ is Eigenvalue.

Let A be a m-by-m matrix, A is singular if and only if 0 is an eigenvalue.

Proof. A is singular $\iff$ There exists $\mathbf{x} \neq \mathbf{0}$ such that $A.\mathbf{x} = \mathbf{0} \iff A.\mathbf{x} = 0\mathbf{x} \iff \lambda = 0$ is eigenvalue. $\square$

Definition I.3.4. Characteristic Polynomial.

The characteristic polynomial of a matrix $A \in \mathbb{C}^{m\times m}$ is:

$$\det(A - \lambda\,\mathrm{Id}) = 0.$$

⌟

Hence, we may define the polynomial that:

$$\det(A) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)},$$

where σ is the permutation, S_m is the n-th cyclic group, and sgn is the sign function of the permutation, *i.e.*, even or odd.
The degree of the characteristic polynomial is the dimension of the square matrix, and by the fundamental theorem of algebra, since $\mathbb{C}$ is algebraically closed, we are guaranteed with full set of complex eigenvalues.

Proposition I.3.5. Roots of Characteristic Polynomial are Eigenvalues.

The roots of the characteristic polynomial are the eigenvalues with respective multiplicity. If an eigenvalue has multiplicity 1, then it is a simple eigenvalue.

Remark I.3.6. Diagonalized Matrices.

When A is diagonal, the eigenvalues are the entries on the diagonal. ⌟

Theorem I.3.7. Determinant and Eigenvalues.

$\det A$ is the product of the eigenvalues counted with multiplicity.

Theorem I.3.8. Complex Conjugates in Real Field.

Suppose $A \in \mathbb{R}_{m,m}$ has only real entries, then the coefficients in p_A are also real. If $\lambda = a + ib$ is an eigenvalues, then:

(i) $\overline{\lambda} = a - ib$ is also an eigenvalue, and

(ii) $\overline{\lambda}$ has the same multiplicity with λ.

Proof. Note that:

$$\overline{p_A(\lambda)} = \overline{\sum_{k=1}^{n} p_k \lambda^k} = 0,$$
$$\sum_{k=1}^{n} p_k \overline{\lambda^k} = 0.$$

Hence $\overline{\lambda}$ is also a root. □

If A is real and symmetric, then we have:

$$A^\intercal = A \text{ and } A^* = A,$$

then we have only real eigenvalues.

Definition I.3.9. Eigenspace.

The eigenspace associated with eigenvalue λ_i is:

$$\ker(A - \lambda_i \operatorname{Id}) = \{x : (A - \lambda_i \operatorname{Id}).x = 0\}.$$

Every vector in this subspace is a eigenvector of A with eigenvalue λ_i. ⌟

Definition I.3.10. Geometric Multiplicity.

The geometric multiplicity of an eigenvalue λ_i is defined as:

$$\dim(\text{eigenspace of } \lambda_i) = \dim\left(\ker(A - \lambda_i \operatorname{Id})\right) = \#\text{ of Linearly Independent Eigenvectors of } \lambda_i.$$ ⌟

Proposition I.3.11. Monotonicity of Multiplicity.

For each matrix $A \in \mathbb{C}^{n\times n}$ with an eigenvalue λ, we have:

$$1 \leqslant \text{Geometric Multiplicity}(\lambda) \leqslant \text{Algebraic Multiplicity}(\lambda).$$

The second equality holds when A is diagonal.

Example I.3.12. Example of Strict Inequality.

Let $A = \begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$, note that 5 is an eigenvalue, we may observe that its algebraic multiplicity is 3.

Consider $A - 5\,\mathrm{Id} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} := B$, and note that:

$$\ker B = \{(x_1, 0, x_3) : x_1, x_3 \in \mathbb{C}\}.$$

Hence, the geometric multiplicity, we have 2, which is strictly less than the algebraic multiplicity. ⌟

Definition I.3.13. Defective Eigenvalue and Non-Defective Matrix.

If Geometric Multiplicity(λ) < Algebric Multiplicity(λ), then λ_i is a *defective* eigenvalue.
$A \in \mathbb{C}^{m\times m}$ is *non-defective* if it has m independent eigenvectors, which is equivalently Geometric Multiplicity(λ) < Algebric Multiplicity(λ) for all eigenvalues of A. ⌟

Proposition I.3.14. Distinct Eigenvalues $\Longrightarrow$ LI Eigenvectors.

Suppose $A \in \mathbb{C}^{m\times m}$ is a matrix with eigenvectors $\{x_1, \cdots, x_k\}$ whose corresponding eigenvalues as $\{\lambda_1, \cdots, \lambda_k\}$. If all eigenvalues are distinct, then the eigenvectors are linearly independent.
If $k = m$, then all eigenvectors of A are linearly independent.

The above case does not account for invertible matrix, as zero can be an eigenvalue.

Definition I.3.15. Similar Matrices.

Suppose $A, B \in \mathbb{C}^{m\times m}$, they are similar if $A = SBS^{-1}$ for some invertible matrix S. ⌟

Proposition I.3.16. Consequences of Similar Matrices.

Suppose $A, B \in \mathbb{C}^{m\times m}$ are similar, then:

(i) $\det A = \det B$,

(ii) $p_A(\lambda) = p_B(\lambda)$, and

(iii) The set of all eigenvalues of A and B are identical.

Definition I.3.17. Diagonalizable.

$A \in \mathbb{C}^{m\times m}$ is diagonal if it is similar to a diagonal matrix $D \in \mathbb{C}^{m\times m}$, *i.e.*, there exists invertible matrix $S \in \mathbb{C}^{m\times m}$ such that $A = SDS^{-1}$. ⌟

Theorem I.3.18. Diagonalizable $\iff$ Non-defective.

Suppose $A \in \mathbb{C}^{m\times m}$. A is diagonalizable if and only if A has m linearly independent eigenvectors, *i.e.*, A is non-defective.

Proof. Note that $A = SDS^{-1}$ is equivalent to $AS = SD$, hence equivalent to:

$$\begin{bmatrix} A\mathbf{S_1} & A\mathbf{S_2} & \cdots & A\mathbf{S_m} \end{bmatrix} = \begin{bmatrix} d_{1,1}\mathbf{S_1} & d_{2,2}\mathbf{S_2} & \cdots & d_{m,m}\mathbf{S_m} \end{bmatrix}.$$

Hence, we equivalently have $A.\mathbf{S_i} = d_{i,i}\mathbf{S_i}$, so $\mathbf{S_i}$ is a set of eigenvector of A with eigenvalues $d_{i,i}$. Since S is invertible, all columns of S are linearly independent. □

Corollary I.3.19. Distinct Eigenvalues $\implies$ Diagonalizable.

If $A \in \mathbb{C}^{m\times m}$ has m distinct eigenvalues, then A is diagonalizable.

This corollary is an immediate consequence of Diagonalizable $\iff$ Non-defective.

The Complex Spectral Theorem can be generalized to the real Spectral Theorem.

Proposition I.3.20. Kernel of A and A^*A are Same.

For any $A \in \mathbb{C}^{m\times n}$, we have:

$$\ker A = \ker(A^*A).$$

Proof. We first show that $\ker A \subset \ker(A^*A)$. Suppose $\mathbf{x} \in \ker A$, then $A.\mathbf{x} = \mathbf{0}$, so $A^*A.\mathbf{x} = \mathbf{0}$, hence $\mathbf{x} \in \ker(A^*A)$.

For the other inclusion, we suppose $\mathbf{x} \in \ker(A^*A) = 0$, then $\mathbf{x}^*A^*A\mathbf{x} = \mathbf{x}^*.\mathbf{0}$. By collecting the terms, we have $(A.\mathbf{x})^*(A.\mathbf{x}) = 0$, which implies that $\|A.\mathbf{x}\|^2 = 0$, and by the axiom of normed vector space, we have $A.\mathbf{x} = \mathbf{0}$, which implies that $\mathbf{x} \in \ker A$. □

I.4 Diagonalization and Singular Value Decomposition

Remark I.4.1. Matrices as Stretches.

Let $A \in \mathbb{C}^{n\times m}$ be a matrix with complex entries. The map $T : \mathbb{C}^n \to \mathbb{C}^m$ which for any $\mathbf{x} \in \mathbb{C}^n$, having $\mathbf{x} \mapsto A.\mathbf{x}$, deforms a unit circle $S = \{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$ into an ellipse. ⌟

Definition I.4.2. Singular Values of Matrix.

Let $A \in \mathbb{C}^{n\times m}$ be a matrix with complex entries. We let $\mathbf{v_1}$ be the vector such that:

$$\|A\mathbf{v_1}\| = \max_{\substack{\mathbf{u}\in\mathbb{C}^n \\ \|\mathbf{u}\|_2=1}} \|A\mathbf{u}\|_2,$$

and we let $\mathbf{u_1}$ be the unit vector of v_1, and the singular value σ_1 satisfies that:

$$\sigma_1 = \|A.\mathbf{v_1}\|.$$

There also exists $\mathbf{v_2} \in \mathbb{C}^2$ such that $\mathbf{v_2} \perp \mathbf{v_1}$ and $A.\mathbf{v_2} \perp A.\mathbf{v_1}$, and we define the other singular value σ_2:

$$\sigma_2 = \|A.\mathbf{v_2}\|.$$

⌟

Proposition I.4.3. Properties of Full Rank 2-by-2 Matrices.

Let A be a 2-by-2 matrix with full rank, there exists $\mathbf{v_1}, \mathbf{v_2} \in \mathbb{R}^2$ with the following properties:

(i) $\beta = \{\mathbf{v_1}, \mathbf{v_2}\}$ is an orthonormal basis of $\mathbb{R}^2$.

(ii) If we define $A.\mathbf{v_1} = \sigma_1 \mathbf{u_1}$ and $A.\mathbf{v_2} = \sigma_2 \mathbf{u_2}$, we have $\sigma_1 \geqslant \sigma_2 > 0$ and $A.\mathbf{v_1} \perp A.x\mathbf{v_2}$.

In particular, we can rewrite it as:

$$\begin{bmatrix} A.\mathbf{v_1} & A.\mathbf{v_2} \end{bmatrix} = \begin{bmatrix} \sigma_1 \mathbf{u_1} & \sigma_2 \mathbf{u_2} \end{bmatrix},$$

which results in:

$$A \underbrace{\begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} \end{bmatrix}}_{V} = \underbrace{\begin{bmatrix} \mathbf{u_1} & \mathbf{u_2} \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\Sigma}.$$

Since U and V are unitary, we have:

$$A = U \Sigma V^*.$$

In general, for matrix $A \in \mathbb{C}^{m \times n}$ with $\dim(\operatorname{im} A) = r$, we can write:

$$\begin{array}{ccccccc} A & = & U & \circ & \Sigma & \circ & V. \\ \Cap & & \Cap & & \Cap & & \Cap \\ \mathbb{C}^{m \times n} & & \mathbb{C}^{m \times m} & & \mathbb{C}^{m \times n} & & \mathbb{C}^{n \times n} \end{array}$$

Proposition I.4.4. Properties of $\bullet^*\bullet$ for Matrices.

For any matrix $A \in \mathbb{C}^{m \times n}$, with $\dim(\operatorname{im} A) = r$, we have:

(i) A^*A is n-by-n matrix and Hermitian (self-adjoint), *i.e.*, $(A^*A)^* = A^*A$,

(ii) $\dim\big(\dim(A^*A)\big) = r$,

(iii) A^*A has r *nonzero* eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$, while $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_n = 0$.

(iv) A^*A and AA^* have the same *nonzero* eigenvalues.

Proof. (i) Trivial.

(ii) Concerning the rank-nullity theorem, we have:

$$\dim\big(\operatorname{im}(A^*A)\big) = n - \dim\big(\ker(A^*A)\big) = n - \dim(\ker A) = \dim(\operatorname{im} A).$$

(iii) Consider the eigenspace for $\lambda = 0$, we have:

$$\mathrm{E}_0 = \ker(A - 0 \cdot \mathrm{Id}) = \ker A.$$

Then, we have the dimension of E_0 as $n-r$, so A has r *nonzero* eigenvalues. Suppose $\lambda \neq 0$ is an eigenvalue of A^*A, so there exists $\mathbf{x} \neq \mathbf{0}$ such that $A^*A.\mathbf{x} = \lambda\mathbf{x}$, then we have $\mathbf{x}^*A^*A.\mathbf{x} = \lambda\mathbf{x}^*.\mathbf{x}$, so we have $\|A.\mathbf{x}\|^2 = \lambda\|\mathbf{x}\|^2$, so:

$$\lambda = \frac{\|A.\mathbf{x}\|^2}{\|\mathbf{x}\|^2} > 0.$$

(iv) Suppose that $\lambda \neq 0$ is an eigenvalue of A^*A, then there exists $\mathbf{x} \neq \mathbf{0}$ such that $A^*A.\mathbf{x} = \lambda\mathbf{x}$, so:

$$AA^*A.\mathbf{x} = \lambda A.\mathbf{x},$$

hence the eigenvalue of AA^* λ and the eigenvector is $A.\mathbf{x}$ (since $\mathbf{x}$ is eigenvector and $\lambda \neq 0$, $A.\mathbf{x} \neq \mathbf{0}$, otherwise it is a contradiction). $\square$

Proposition I.4.5. Single Value Decomposition.

Suppose $A \in \mathbb{C}^{m\times n}$ such that $\dim(\operatorname{im} A) = r \leqslant \min\{m, n\}$ in which $A : \mathbb{C}^n \to \mathbb{C}^m$, then there exists:

(i) an orthonormal basis $\{\mathbf{v_1} \cdots, \mathbf{v_r}, \underbrace{\mathbf{v_{r+1}}, \cdots, \mathbf{v_n}}_{\text{basis for } \ker A}\}$ of $\mathbb{C}^n$,

(ii) another orthonormal basis $\{\underbrace{\mathbf{u_1} \cdots, \mathbf{u_r}}_{\text{basis for } \operatorname{im} A}, \mathbf{u_{r+1}}, \cdots, \mathbf{u_m}\}$ of $\mathbb{C}^m$, and

(iii) r singular values, with $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$,

such that:

(i) $U^*U = UU^* = \mathrm{Id}$,

(ii) $V^*V = VV^* = \mathrm{Id}$, and

(iii) $\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \sigma_r & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$

Remark I.4.6. Single Value for Full Rank.

Suppose A is square and *full rank*, it is $r \times r$ and $\ker A = \{\mathbf{0}\}$, $\operatorname{im} A = \mathbb{C}^r$. Moreover, the have the bases $\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_r}\}$ and $\gamma = \{\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_r}\}$, while the Σ matrix is:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}$$

is square, diagonal, and full rank. ⌟

Theorem I.4.7. Spectral Theorem.

Suppose $B \in \mathbb{C}^{n\times n}$ is Hermitian (*i.e.*, $B^* = B$), then B has n real eigenvalues and it is orthogonally

diagonalizable, *i.e.*, B has n orthonormal eigenvectors $\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\}$, such that, if we define:
$$S := \begin{bmatrix} \mathbf{v_1} & \cdots & \mathbf{v_n}, \end{bmatrix}$$
we have:
$$S^*AS = D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$
Hence $A = SDS^*$.

Remark I.4.8. Remark on Real Spectral Theorem.

For the case above, we want to construct unitary $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{m\times m}$, and $\Sigma \in \mathbb{C}^{m\times n}$ that is diagonalizable such that $A = U\Sigma V^*$. The sketch of the proof would be solving for U, Σ, V, where we first compute:
$$A^*A = (U\Sigma V^*)^*(U\Sigma V^*) = V\Sigma^* U^* U \Sigma V^*,$$
hence implying that:
$$A^*A = V(\Sigma^*\Sigma)V^*,$$
which implies that:
$$V = \begin{bmatrix} \mathbf{v_1} & \cdots & \mathbf{v_n}, \end{bmatrix}$$
which gives:
$$\Sigma^*\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \sigma_r^2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$
Hence, the singular values are $\sigma_i = \sqrt{\lambda_i}$ with $1 \leqslant i \leqslant r$. ⌟

For the process, we diagonalize A^*A with respect to orthonormal basis $mathbf{v_i}\}$ and take $\mathbf{v_i}$'s to form columns of V and compute $\sigma_i\sqrt{\lambda_i}$ to form Σ.
Then, we choose $\mathbf{u_i}$'s via:
$$\mathbf{u_i} = \frac{A.\mathbf{v_i}}{\sigma_i} \text{ for } i = 1, 2, \cdots, r.$$
For the remaining $u_{r+1}, \cdots, u_m$ are chosen to be orthonormal to $u_1, \cdots, u_m$, using Graham Schmidt process.

Example I.4.9. Finding SVD.

Let $A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ -4 & 4 \end{bmatrix}$, we note that $m = 3$, $n = 2$, and $r = 1$ since there is only one linearly independent column.

(i) First, we find A^*A, that is:
$$A^*A = \begin{bmatrix} 18 & -18 \\ -18 & 18 \end{bmatrix},$$

and we note that eigenvalues as:

$$A^*A.\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \text{ and } A^*A.\begin{bmatrix}1\\-1\end{bmatrix} = 36\begin{bmatrix}1\\-1\end{bmatrix}$$

.

(ii) Now, since we have the eigenvalues as 36 and 0, whose eigenvectors are $\mathbf{v_1} = 1/\sqrt{2}\begin{bmatrix}1\\-1\end{bmatrix}$ and $\mathbf{v_2} = 1/\sqrt{2}\begin{bmatrix}1\\1\end{bmatrix}$, then the first singular value is $\sigma_1 = 6$, then we have:

$$\mathbf{u_1} = \frac{A\mathbf{v_1}}{\sigma_1} = \frac{1}{3\sqrt{2}}\begin{bmatrix}1\\-1\\-4\end{bmatrix}$$

(iii) Then, we look for $\mathbf{u_2} \perp \mathbf{u_3}$, in which we want:

$$\frac{1}{3\sqrt{2}}\begin{bmatrix}y_1 & y_2 & y_3\end{bmatrix}\begin{bmatrix}1\\-1\\-4\end{bmatrix} = 0,$$

which returns to:

$$y_1 = y_2 + 4y_3$$

As of right now, we have:

$$y = \begin{bmatrix}y_2 + 4y_3\\ y_2\\ y_3\end{bmatrix} = y_2\begin{bmatrix}1\\1\\0\end{bmatrix} + y_3\begin{bmatrix}4\\0\\1\end{bmatrix}.$$

Note that we want them to be orthogonal, that is:

$$\begin{bmatrix}1\\1\\0\end{bmatrix} \text{ and } \begin{bmatrix}2\\-2\\1\end{bmatrix}.$$

Technically, we should use Graham Schmidt to find an orthonormal basis, but this can be observed easily.

⌟

Proposition I.4.10. Block Matrix Multiplication.

Suppose $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$, we can break:

$$m = \sum_{i=1}^{\alpha} m_i,$$

$$n = \sum_{i=1}^{\beta} n_i,$$

$$p = \sum_{i=1}^{\gamma} p_i.$$

In particular, we can write:

$$AB = \begin{bmatrix} A_{1,1} & \cdots & A_{1,\beta} \\ \vdots & \ddots & \vdots \\ A_{\alpha,1} & \cdots & A_{\alpha,\beta} \end{bmatrix} \begin{bmatrix} B_{1,1} & \cdots & B_{1,\gamma} \\ \vdots & \ddots & \vdots \\ B_{\beta,1} & \cdots & B_{\beta,\gamma} \end{bmatrix},$$

where computation is distributive as if they are scalar entries.

There is an alternative way to compute $\gamma = \{\mathbf{u_1}, \cdots, \mathbf{u_r}, \mathbf{u_{r+1}}, \cdots, \mathbf{u_m}\}$, hence:

$$A = U\Sigma V^* \iff AV = U\Sigma V^* V = U\Sigma.$$

The right hand side can be simplified into:

$$U\Sigma = \begin{bmatrix} \mathbf{u_1} & \mathbf{u_2} & \cdots & \mathbf{u_m} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \sigma_r & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \sigma_1 \mathbf{u_1} & \sigma_2 \mathbf{u_2} & \cdots & \sigma_m \mathbf{u_m} \end{bmatrix}.$$

For the left hand side, we have:

$$AV = \begin{bmatrix} A\mathbf{v_1} & A\mathbf{v_2} & \cdots & A\mathbf{v_r} & A\mathbf{v_{r+1}} & \cdots & A\mathbf{v_n} \end{bmatrix}.$$

Hence, for the first r columns, we have:

$$\mathbf{u_i} = \frac{A.\mathbf{v_i}}{\sigma_i} \text{ for } i = 1, 2, \cdots, r,$$

where as for the last vectors, they are the kernel of the map, and must be set to be an orthonormal set that is also orthogonal to prior entries.

Theorem I.4.11. Properties about Matrix Norms.

For any $m \times n$ matrix A, we have:

(i) $\|A\|_2 = \sigma_1$, *i.e.*, the largest singular value (as σ_i's are positive and ordered from large to small), and

(ii) $\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}$.

Recall that $\| \bullet \|_2$ and $\| \bullet \|_F$ are invariant by multiplication by unitary matrices.

Proof. (i) Note that:

$$\begin{aligned} \|A\|_2 &= \|U\Sigma V^*\|_2 = \|U\Sigma\|_2 = \|\Sigma\|_2 \\ &= \max_{\substack{\mathbf{u} \in \mathbb{C}^n \\ \|\mathbf{u}\|_2 = 1}} \|\Sigma.\mathbf{u}\|_2 = \max_{\substack{\mathbf{u} \in \mathbb{C}^n \\ \|\mathbf{u}\|_2 = 1}} \sqrt{\|u_1\|^2\sigma_1^2 + |u_2|\sigma_2^2 + \cdots + |u_r|^2\sigma_r^2} = \sqrt{|1|^2\sigma_1^2} = \sigma_1. \end{aligned}$$

(ii) For Forbenius norm, we have:

$$\begin{aligned} \|A\|_F &= \|U\Sigma V^*\|_F = \|U\Sigma\|_F = \|\Sigma\|_F \\ &= \sqrt{|\sigma_1|^2 + |\sigma_2|^2 + \cdots + |\sigma_r|^2} = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}. \end{aligned}$$

□

Proposition I.4.12. Hermitian $\Longrightarrow$ Orthogonally Diagonalizable.

Suppose A is Hermitian, then A is orthogonally diagonalizable.
Moreover, the singular values of A are $\sigma_i = |\lambda_i|$, where λ_i's are the *ordered* (by absolute value) of eigenvalues of A.

Proof. Let $A = SDS^*$, in particular:

$$A = \underbrace{\begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_n} \end{bmatrix}}_{\text{orthonormal set of eigenvectors of } A} \underbrace{\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}}_{\text{Not necessarily SVD, as eigenvalues can be non-positive}} \begin{bmatrix} \mathbf{v_1}^* \\ \mathbf{v_2}^* \\ \vdots \\ \mathbf{v_n}^* \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} \operatorname{sgn}(\lambda_1)\mathbf{v_1} & \operatorname{sgn}(\lambda_2)\mathbf{v_2} & \cdots & \operatorname{sgn}(\lambda_n)\mathbf{v_n} \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} |\lambda_1| & 0 & \cdots & 0 \\ 0 & |\lambda_2| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & |\lambda_n| \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \mathbf{v_1}^* \\ \mathbf{v_2}^* \\ \vdots \\ \mathbf{v_n}^* \end{bmatrix}}_{V},$$

where the sign function is defined to be:

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x \geqslant 0, \\ -1 & \text{if } x < 0. \end{cases}$$

□

Now, as we consider the SVD for $A \in \mathbb{C}^{m \times n}$, where $\dim(\operatorname{im} A) = r$, we have:

$$A = \begin{bmatrix} \overbrace{\mathbf{u_1}\, \mathbf{u_2} \cdots \mathbf{u_r}}^{r \text{ columns}} & \overbrace{\mathbf{u_{r+1}} \cdots \mathbf{u_m}}^{m-r \text{ columns}} \end{bmatrix} \begin{bmatrix} \overbrace{\begin{matrix}\sigma_1 & \cdots & 0\end{matrix}}^{\tilde{\Sigma}} & 0 & \cdots & 0 \\ \vdots \quad \ddots \quad \vdots & \vdots & \ddots & \vdots \\ \begin{matrix}0 & \cdots & \sigma_r\end{matrix} & 0 & \cdots & 0 \\ \begin{matrix}0 & \cdots & 0\end{matrix} & 0 & \cdots & 0 \\ \vdots \quad \ddots \quad \vdots & \vdots & \ddots & \vdots \\ \begin{matrix}0 & \cdots & 0\end{matrix} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \left.\begin{matrix} \mathbf{v_1}^* \\ \vdots \\ \mathbf{v_r}^* \end{matrix}\right\} r \text{ rows} \\ \left.\begin{matrix} \mathbf{v_{r+1}}^* \\ \vdots \\ \mathbf{v_n}^* \end{matrix}\right\} n-r \text{ rows} \end{bmatrix}$$

Computationally, we find that:

(i) $\mathbf{v_1}, \cdots, \mathbf{v_r}, \mathbf{v_{r+1}}, \cdots, \mathbf{v_n}$ are orthonormal eigenvectors of A^*A with respective eigenvalues, and

(ii) $\mathbf{u_1}, \cdots, \mathbf{u_r}, \mathbf{u_{r+1}}, \cdots, \mathbf{u_m}$ are orthonormal eigenvectors of AA^* with respective eigenvalues.

Recall the block computation, we have:

$$A = \begin{bmatrix} \tilde{U} & U_2 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{V}^* \\ V_2^* \end{bmatrix} = \begin{bmatrix} \tilde{U}\tilde{\Sigma} + U_2 0 & \tilde{U}0 + U_2 0 \end{bmatrix} \begin{bmatrix} \tilde{V}^* \\ V_2^* \end{bmatrix} = \tilde{U}\tilde{\Sigma}\tilde{V}^*.$$

Definition I.4.13. Reduced Singular Value Decomposition.

With the construction above, with $A \in \mathbb{C}^{m \times n}$ being Hermitian, we have reduced Singular Value Decomposition, that is:

$$\begin{array}{ccccccc} A & = & \tilde{U} & \circ & \tilde{\Sigma} & \circ & \tilde{V}. \\ \pitchfork & & \pitchfork & & \pitchfork & & \pitchfork \\ \mathbb{C}^{m\times n} & & \mathbb{C}^{m\times r} & & \mathbb{C}^{r\times r} & & \mathbb{C}^{r\times n} \end{array}$$

Note that with reduced SVD, we have:

$$A = \begin{bmatrix} \sigma_1 \mathbf{u_1} & \sigma_2 \mathbf{u_2} & \cdots & \sigma_r \mathbf{u_r} \end{bmatrix} = \sigma_1 \mathbf{u_1} \mathbf{v_1^*} + \sigma_2 \mathbf{u_2} \mathbf{v_2^*} + \cdots + \sigma_r \mathbf{u_r} \mathbf{v_r^*},$$

which is called the *rank-one* decomposition of A. Hence A is the sum of rank-one matrices with weights $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$.

Remark I.4.14. Approximating a Matrix.

By choosing a $p \leqslant r$, we have the rank-p approximation of A such thatL

$$\widehat{A_p} = \sigma_1 \mathbf{u_1} \mathbf{v_1} + \cdots + \sigma_p \mathbf{u_p} \mathbf{v_p^*},$$

where we have $\dim(\operatorname{im} \widehat{A_p}) = p$.

Example I.4.15. Using SVD to Compress Image(s).

Suppose we use a 400-by-400 matrix to represent a gray-scale image, *i.e.*, $A \in [0, 255]^{400\times 400}$, that is:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,400} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,400} \\ \vdots & \vdots & \ddots & \vdots \\ A_{400,1} & A_{400,2} & \cdots & A_{400,400} \end{bmatrix},$$

where each entry represents the gray-scale in that pixel, where 0 represents black and 255 represents white, and the grays are within $(0, 255)$, while becoming lighter as the number increases.
Consider the reduced SVD, we may write A into:

$$A = \sigma_1 \mathbf{u_1} \mathbf{v_1^*} + \sigma_2 \mathbf{u_2} \mathbf{v_2^*} + \cdots + \sigma_r \mathbf{u_{400}} \mathbf{v_{400}^*},$$

in which we only consider the first p terms, we are able to reduce the rank in order to store less data for the image. In this way, for rank p, we only need to store $(1 + 400 + 400) \times p$ values rather than everything.

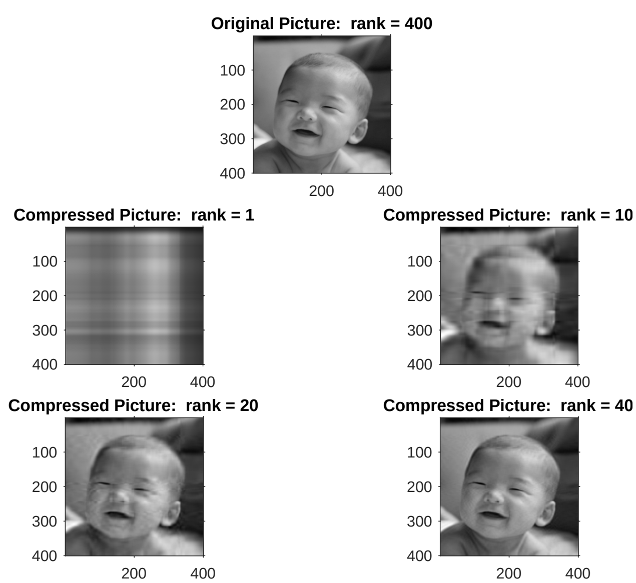


Figure I.6. Compressing an image of full rank 400 into lower ranks.

With the example image above, we note that when rank is 10, the image is roughly recognizable, where as at rank 40, the image is clear enough.

I.5 Projection

Definition I.5.1. Idempotent.

For matrix $P \in \mathbb{C}^{m\times m}$ is *idempotent* ot projector if $P^2 = P$.

Hence, for all $\mathbf{x} \in \mathbb{C}^m$, we have $P.\mathbf{x} = P^2.\mathbf{x}$, *i.e.*, $P.\mathbf{x} = P(P.\mathbf{x})$.

Definition I.5.2. Orthogonal Projector on Vector.

For a fixed nonzero $\mathbf{v} \in \mathbb{C}^m$, we define that orthogonal projector onto $\mathbf{v}$ as:

$$P_{\mathbf{v}} = \frac{1}{\|\mathbf{v}\|^2}\mathbf{v}\mathbf{v}^* = \frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}}.$$

Geometrically, we may represent the projector as:

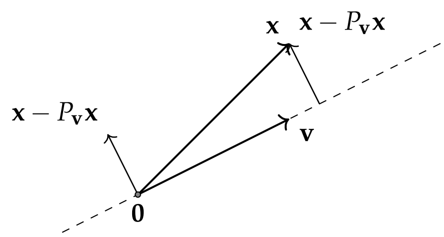


Figure I.7. Geometric feature of the Orthogonal Projector.

Theorem I.5.3. Geometric Properties with Orthogonal Projection.

For all $\mathbf{x} \in \mathbb{C}^m$, we have:

(i) $P_{\mathbf{v}}\mathbf{x} \in \operatorname{span}\{\mathbf{v}\}$, *i.e.*, $\mathbf{y} \parallel \mathbf{v}$, and

(ii) $\mathbf{x} - P_{\mathbf{v}}\mathbf{x} \perp \mathbf{v}$.

Proof. (i) For the projector, we have:

$$P_{\mathbf{v}}\mathbf{x} = \frac{1}{\|\mathbf{v}\|^2}\mathbf{v}\mathbf{v}^*\mathbf{x} = \frac{\mathbf{v}^*\mathbf{x}}{\|\mathbf{v}\|^2}\mathbf{v} = k\mathbf{v} = \operatorname{span}\{\mathbf{v}\}.$$

(ii) For the inner product:

$$\mathbf{v}^*\left(\mathbf{x} - P_{\mathbf{v}}\mathbf{x}\right) = \mathbf{v}^*\mathbf{x} - \mathbf{v}^*\left(\frac{1}{\|\mathbf{v}\|^2}\mathbf{v}\mathbf{v}^*\mathbf{x}\right) = \mathbf{v}^*\mathbf{x} - \frac{1}{\|\mathbf{v}\|^2}\mathbf{v}^*\mathbf{v}\mathbf{v}^*\mathbf{x} = 0.$$

□

Proposition I.5.4. Orthogonal Projector is Idempotent.

Let $P_{\mathbf{v}}$ be the orthogonal projector onto $\mathbf{v}$, $P_{\mathbf{v}}^2 = P_{\mathbf{v}}$.

Proof. We may deduce that:

$$P_{\mathbf{v}}^2 = \left(\frac{1}{\|\mathbf{v}\|^2}\mathbf{v}\mathbf{v}^*\right)\left(\frac{1}{\|\mathbf{v}\|^2}\right)\mathbf{v}\mathbf{v}^* = \frac{1}{\|\mathbf{v}\|^4}\mathbf{v}\mathbf{v}^*\mathbf{v}\mathbf{v}^*$$
$$= \frac{\mathbf{x}^*\mathbf{v}}{\|\mathbf{v}\|^4}\mathbf{v}\mathbf{v}^* = \frac{1}{\|\mathbf{v}\|^2}\mathbf{v}\mathbf{v}^* = P_{\mathbf{v}}.$$

□

Recall that if $\beta = \{\mathbf{v_1}, \cdots, \mathbf{v_r}\}$ is an orthogonal basis of a subspace $V \subset \mathbb{C}^m$, then for all $\mathbf{x} \in V$, we can write:

$$\mathbf{x} = c_1\mathbf{v_1} + c_2\mathbf{v_2} + \cdots + c_r\mathbf{v_r},$$

where each $c_i = \dfrac{\mathbf{v_i}^*\mathbf{x}}{\|\mathbf{v_i}\|^2}$ for all $i = 1, 2, \cdots, r$.

Only when we have a orthogonal basis, we have the following sum:

$$\mathbf{x} = \frac{\mathbf{v_1}^*\mathbf{x}}{\|\mathbf{v_1}\|^2}\mathbf{v_1} + \frac{\mathbf{v_2}^*\mathbf{x}}{\|\mathbf{v_2}\|^2}\mathbf{v_2} + \cdots + \frac{\mathbf{v_r}^*\mathbf{x}}{\|\mathbf{v_r}\|^2}\mathbf{v_r} = P_{\mathbf{v_1}}\mathbf{x} + P_{\mathbf{v_2}}\mathbf{x} + \cdots + P_{\mathbf{v_r}}\mathbf{x}.$$

However, on non orthogonal basis, this might not be true.

Let $\{\mathbf{v_1}, \mathbf{v_2}\}$ be an orthogonal basis and let $\{\mathbf{w_1}, \mathbf{w_2}\}$ be a non-orthogonal basis in $\mathbb{R}^2$, we can present the following projections of $\mathbf{x} \in \mathbb{R}^2$.

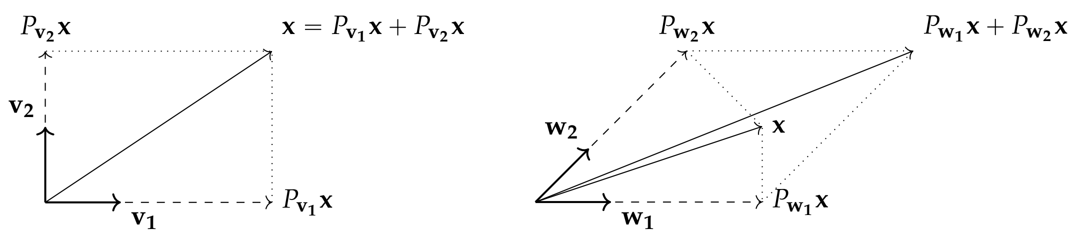


Figure I.8. Projection with orthogonal basis (left) and non-orthogonal basis (right).

Proposition I.5.5. Complementary Projector is Project.

If $P \in \mathbb{C}^{m\times m}$ is a projector, then $\mathrm{Id} - P$ is also a projector.

Proof. Note that:

$$(\mathrm{Id} - P)^2 = (\mathrm{Id} - P)(\mathrm{Id} - P) = \mathrm{Id}^2 - P - P + P^2 = \mathrm{Id}^2 - P - P + P = \mathrm{Id}^2 - P.$$

□

Remark I.5.6. Kernel and Image of Complementary Projector.

Consider a orthogonal projector on a line $\mathbf{v} \neq \mathbf{0}$, and for $\mathbf{v} \in \mathbb{C}^m$, with $P_{\mathbf{v}} = \mathbf{v}\mathbf{v}^*/\|\mathbf{v}\|^2$, so we have:

$$(\mathrm{Id} - P)\mathbf{x} = \mathbf{x} - P_{\mathbf{v}}\mathbf{x}.$$

Recall the kernel and image of the projector:

$$\mathrm{im}(P_{\mathbf{v}}) = \mathrm{span}\{\mathbf{v}\},$$
$$\ker(P_{\mathbf{v}}) = \{\mathbf{w} : \mathbf{v}^*\mathbf{w} = 0\} = (\mathrm{span}\{\mathbf{v}\})^{\perp}.$$

However, for the complementary projector, we have:

$$\mathrm{im}(\mathrm{Id} - P_{\mathbf{v}}) = \{\mathbf{w} : \mathbf{v}^*\mathbf{w} = 0\} = \ker(P_{\mathbf{v}}),$$
$$\ker(\mathrm{Id} - P_{\mathbf{v}}) = \mathrm{span}\{\mathbf{v}\} = \mathrm{im}(P_{\mathbf{v}}).$$

For any *Idempotent* matrices $P^2 = P$, we first find $\ker P$ such that:

$$\mathbf{w} = \mathbf{x} - P.\mathbf{x},$$

We find that:

$$P.\mathbf{w} = P.\mathbf{x} - P^2.\mathbf{x} = P.\mathbf{x} - P.\mathbf{x} = \mathbf{0}.$$

Hence, $\mathbf{w} \in \ker P$.
Also, we have $\mathbf{w} - (\operatorname{Id} - P).\mathbf{x} \in \operatorname{im}(\operatorname{Id} - P)$.

Theorem I.5.7. Idempotent $\Longrightarrow$ Image and Kernel Relation.

If $P = P^2$, then:

$$\operatorname{im}(\operatorname{Id} - P) = \ker P.$$

Moreover:

$$\operatorname{im} P = \ker(\operatorname{Id} - P).$$

Proof. (i) ($\operatorname{im}(\operatorname{Id} - P) \subseteq \ker P$:) We let $\mathbf{w} \in \operatorname{im}(\operatorname{Id} - P)$ be generic, then there exists $\mathbf{x}$ such that:

$$\mathbf{w} = (\operatorname{Id} - P).\mathbf{x} = \mathbf{x} - P.\mathbf{x},$$

hence:

$$P.\mathbf{w} = P.\mathbf{x} - P^2.\mathbf{x} = P.\mathbf{x} - P.\mathbf{x} = \mathbf{0}.$$

Therefore, $w \in \ker P$.
($\operatorname{im}(\operatorname{Id} - P) \supseteq \ker P$:) Let $\mathbf{w} \in \ker P$ be generic, *i.e.*, $P.\mathbf{w} = \mathbf{0}$, then we have:

$$(\operatorname{Id} - P).\mathbf{w} = \mathbf{w} - P.\mathbf{w},$$

which results in $\mathbf{w} = (\operatorname{Id} - P).\mathbf{w}$, so $w \in \operatorname{im}(\operatorname{Id} - P)$, as desired.

(ii) Let $Q = \operatorname{Id} - P$, by the previous part, we have:

$$\operatorname{im}(\operatorname{Id} - Q) = \ker Q,$$

hence, by $\operatorname{Id} - Q = \operatorname{Id} - \operatorname{Id} + P = P$, hence:

$$\operatorname{im} P = \ker(\operatorname{Id} - Q).$$

$\square$

Proposition I.5.8. Idempotent $\Longrightarrow$ Intersection of Image and Kernel is Trivial.

Let $P = P^2$ be idempotent:

$$\operatorname{im} P \cap \ker P = \{\mathbf{0}\}.$$

Proof. We let $\mathbf{w} \in \operatorname{im} P \cap \ker P$ be trivial, then:

$$\begin{cases} \mathbf{w} \in \operatorname{im} P = \ker(\operatorname{Id} - P), \\ \mathbf{w} \in \ker P. \end{cases}$$

This implies that:

$$\begin{cases} (\mathrm{Id} - P).\mathbf{w} = \mathbf{0}, \\ P.\mathbf{w} = \mathbf{0}, \end{cases}$$

which implies that $\mathbf{w} - P.\mathbf{w} = \mathbf{0}$, so $\mathbf{w} = \mathbf{0}$. □

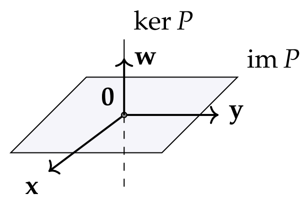


Figure I.9. Kernel and image of P as idempotent.

Note that suppose $\dim(\operatorname{im} P) = r$, by rank-nullity, we have $\dim(\ker P) = m - r$, hence all dimensions are captured by the image and the kernel, that can be represented by direct sum that:

$$\operatorname{im} P \oplus \ker P = \mathbb{C}^m.$$

Thus, any $\mathbf{x} \in \mathbb{C}^m$ can be uniquely decomposed into:

$$\mathbf{x} = \mathbf{y} + \mathbf{w},$$

where $\mathbf{y} \in \operatorname{im} P$ and $\mathbf{w} \in \ker P$.

To compute the $\mathbf{y}$ and $\mathbf{w}$ above, we have:

$$\mathbf{y} \in \operatorname{im} P = \ker(\mathrm{Id} - P),$$

hence resulting in:

$$(\mathrm{Id} - P).\mathbf{y} = \mathbf{0},$$

hence $\mathbf{y} = P.\mathbf{y}$, and by acting P on $\mathbf{x}$, we have:

$$P.\mathbf{x} = P.\mathbf{y} + P.\mathbf{w} = \mathbf{y},$$

hence resulting in:

$$\begin{cases} \mathbf{y} = P.\mathbf{x}, \\ \mathbf{w} = \mathbf{x} - \mathbf{y}. \end{cases}$$

Remark I.5.9. Eigenvalues and Diagonalizability.

Assume $P = P^2$ is idempotent:

(i) For all $\mathbf{x} \in \ker P$, we have $P.\mathbf{x} = \mathbf{0}$, hence $P.\mathbf{x} = 0.\mathbf{x}$.

(ii) For all $\mathbf{x} \in \operatorname{im} P = \ker(\mathrm{Id} - P)$, we have $(\mathrm{Id} - P).\mathbf{x} = \mathbf{0}$, so $\mathbf{x} = P.\mathbf{x}$ hence $P.\mathbf{x} = 1.\mathbf{x}$.

(iii) Hence, for $\lambda = 0$, the eigenvalue has geometric multiplicity of r, and the eigenvector has geometric multiplicity of $m - r$.

In conclusion, idempotent implies diagonalizability. ⌟

Proposition I.5.10. Gram-Schmidt Process.

Let $\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_p}\}$ be a basis of a subspace $V \subset \mathbb{C}^m$, the Gram-Schmidt Process allows us to construct an orthogonal basis $\{\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_p}\}$ such that $\text{span}\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_p}\} = \text{span}\{\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_p}\}$.

Proof. The construction of the Gram-Schmidt is as follows:

(i) Let $\mathbf{u_1} = \mathbf{v_1}$.

(ii) We construct u_2 that is orthogonal to v_1, let:

$$\mathbf{u_2} = \mathbf{v_2} - \text{component of } \mathbf{v_2} \text{ that is parallel to } \mathbf{u_1} = \mathbf{v_2} - \frac{\mathbf{u_1^* v_2}}{\|\mathbf{u_1}\|^2}\mathbf{u_1}.$$

Note that readers can verify that $\text{span}\{\mathbf{u_1}, \mathbf{u_2}\} = \text{span}\{\mathbf{v_1}, \mathbf{v_2}\}$.

(iii) For $\mathbf{u_3}$m we think of it as:

$$\mathbf{u_3} = \mathbf{v_3} - \frac{\mathbf{u_1^* v_3}}{\|\mathbf{u_1}\|^2}\mathbf{u_1} - \frac{\mathbf{u_2^* v_3}}{\|\mathbf{u_2}\|^2}\mathbf{u_2}.$$

This can be illustrated in $\mathbb{R}^3$ here, as follows:

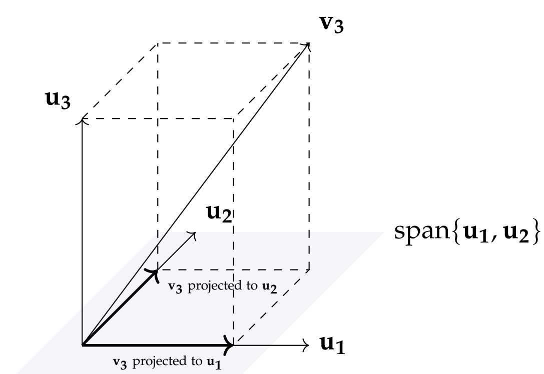


Figure I.10. Subtracting the orthogonal projections gives the orthogonal component.

(k) At step k, we have:

$$\mathbf{u_k} = \mathbf{v_k} - \sum_{j=1}^{k-1} \frac{\mathbf{u_j^* v_k}}{\|\mathbf{u_j}\|^2}\mathbf{u_j}.$$

Hence, we have completed the construction for finitely dimensional basis. □

Remark I.5.11. 3rd Step of Gram Schmidt.

Consider the built:

$$Q_2 = \begin{bmatrix} \mathbf{q_1} & \mathbf{q_2} \end{bmatrix}$$

Here, the projector on $\text{span}\{\mathbf{q_1}, \mathbf{q_2}\}$ has:

$$T_2 = Q_2 Q_2*'$$

which, by block matrix multiplication, leads to:

$$T_2\mathbf{A_3} = Q_2 Q_2^* \mathbf{A_3} = (\mathbf{q_1^*}\mathbf{A_3})\mathbf{q_1} + (\mathbf{q_2^*}\mathbf{A_3})\mathbf{q_2},$$

Hence, we have:

$$\mathbf{u_3} = \mathbf{A_3} - T_2\mathbf{A_3} = (\mathrm{Id} = T_2)\mathbf{A_3},$$

so we can define:

$$P_3 = \mathrm{Id} - T_2, \text{ so } \mathbf{u_3} = P_3\mathbf{A_3}, \text{ hence } \mathbf{q_3} = \frac{\|\mathbf{u_3}\|}{\mathbf{u_3}}.$$

⌟

Theorem I.5.12. For Idempotent, Hermitian $\Longleftrightarrow$ $\operatorname{im} P \perp \ker P$.

Suppose $P = P^2$ is idempotent:

$$P^* = P \iff \ker P \perp \operatorname{im} P.$$

Proof. ($\Longrightarrow$:) Note that for any linear map P, $\ker P \perp \operatorname{im}(P^*)$, hence $P = P^*$ implies $\ker P \perp \operatorname{im} P$.
($\Longleftarrow$:) We choose:

$$\beta = \{\mathbf{q_1}, \mathbf{q_2}, \cdots, \mathbf{q_r}\},$$
$$\gamma = \{\mathbf{q_{r+1}}, \mathbf{q_{r+1}}, \cdots, \mathbf{q_m}\},$$

as the orthonormal basis of $\operatorname{im} P$ and $\ker P$, respectively. From assumption that $\ker P \perp \operatorname{im} P$, then $\beta \perp \gamma$, so we have:

$$P.\mathbf{q_1} = q_1, P.\mathbf{q_2} = q_2, \cdots, P.\mathbf{q_r} = q_r, P.\mathbf{q_{r+1}} = \mathbf{0}, \cdots, P.\mathbf{q_m} = \mathbf{0}.$$

hence, our matrix can be represented as:

$$\begin{bmatrix} P.\mathbf{q_1} & P.\mathbf{q_2} & \cdots & P.\mathbf{q_r} & P.\mathbf{q_{r+1}} & \cdots & P.\mathbf{q_m} \end{bmatrix} = \begin{bmatrix} \mathbf{q_1} & \cdots & \mathbf{q_r} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

Hence, we have:

$$Q^*PQ = \begin{bmatrix} \mathbf{q_1^*} \\ \vdots \\ \mathbf{q_r^*} \\ \mathbf{q_{r+1}^*} \\ \vdots \\ \mathbf{q_m^*} \end{bmatrix} \begin{bmatrix} \mathbf{q_1} & \cdots & \mathbf{q_r} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix},$$

in which the top $r \times r$ rows are non-trivial. Thus:

$$Q^*PQ = D,$$

hence $P = QDQ^*$, so $P^* = (QDQ^*)^* = (Q^*)^*D^*Q^* = QDQ^* = P$. □

In fact, we may simplify the above prove by noticing orthogonality. Since the eigenvalues are non-negative, we have $P = QDQ^*$ being the SVD of P where $U = Q$, $\Sigma = D$ and $V = Q$.
If we are considering the parts of the *nonzero* entries of the matrices, we have $P = \tilde{Q}\tilde{\Sigma}\tilde{Q}^* = \tilde{Q}\,\mathrm{Id}\,\tilde{Q}^* = \tilde{Q}\tilde{Q}^*$.
Note that:

$$\tilde{Q} = \begin{bmatrix} \mathbf{q_1} & \cdots & \mathbf{q_2} \end{bmatrix}$$

is the orthonormal basis of the range, hence inducing a rank-one decomposition, that is:

$$P = \mathbf{q_1}\mathbf{q_1^*} + \mathbf{q_2}\mathbf{q_2^*} + \cdots + \mathbf{q_r}\mathbf{q_r^*}.$$

Recall that doe the projection, we have:

$$P_{\mathbf{q_i}} = \frac{\mathbf{q_i}\mathbf{q_i^*}}{\|\mathbf{q_i}\|^2} = \mathbf{q_i}\mathbf{q_i^*}.$$

Again, we would ask the question. *What if the basis we have for* im P *is not orthonormal?*

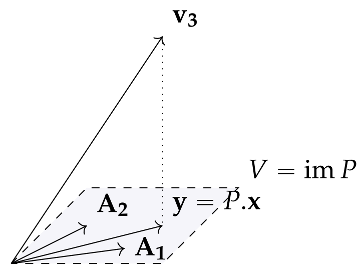


Figure I.11. Projections when the basis is not orthogonal.

In such case, we have:

$$V = \operatorname{span}\{\mathbf{A_1}, \mathbf{A_2}, \cdots, \mathbf{A_r}\},$$

which are linearly independent but not necessarily orthonormal. Hence, we want to find an $m \times n$ matrix P with the property that for all $\mathbf{x} \in \mathbb{C}^m$:

$$\mathbf{y} = P.\mathbf{x} \in V \dashv\backslash\lceil \mathbf{w} = \mathbf{x} - \mathbf{y} \perp V.$$

Proposition I.5.13. Matrix with full rank has Hermitian compose itself Invertible.

For A being a $m \times r$ matrix with $\dim(\operatorname{im} A) = r$, A^*A is invertible.

Definition I.5.14. Pseudo Inverse of a Matrix.

Suppose A is $m \times r$ and $\dim(\operatorname{im} A) = r$, we want to solve that:

$$A.\mathbf{x} = \mathbf{b} \text{ with } \mathbf{b} \in \operatorname{im} A.$$

Hence, for $A^*A\mathbf{x} = A^*\mathbf{b}$ since A^*A is square and invertible, then $\mathbf{x} = (A^*A)^{-1}A^*\mathbf{b}$, and we have:

$$A^+ = (A^*A)^{-1}A^*$$

as the pseudo-inverse of A. ⌟

Theorem I.5.15. Pseudo Inverse Conditions.

For a matrix $P \in \mathbb{C}^{m \times m}$ that satisfies for all $\mathbf{x} \in \mathbb{C}^m$ that:

(i) $\mathbf{y} = P.\mathbf{x} \in V$, and

(ii) $\mathbf{w} = \mathbf{x} - \mathbf{y} \perp V$,

is $P = A(A^*A)^{-1}A^*$ for A being the matrix composed of the linearly independent vectors.

Proof. By (i), for any $\mathbf{x} \in \mathbb{C}^m$, we want $\mathbf{y} = P.\mathbf{x} \in V$, *i.e.*:

$$\mathbf{y} = P.\mathbf{x} = c_1\mathbf{A_1} + c_2\mathbf{A_2} + \cdots + c_r\mathbf{A_r} = \underbrace{\begin{bmatrix} \mathbf{A_1} & \cdots & \mathbf{A_r} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_r \end{bmatrix}}_{\mathbf{c}} = A.\mathbf{c}.$$

By (ii), we have:

$$\begin{aligned} \mathbf{x} - P.\mathbf{x} \perp \mathbf{A_1} \implies & \qquad \mathbf{A_1}^*(\mathbf{x} - P.\mathbf{x}) = 0 \\ \vdots \qquad & \qquad \vdots \\ \mathbf{x} - P.\mathbf{x} \perp \mathbf{A_r} \implies & \qquad \mathbf{A_r}^*(\mathbf{x} - P.\mathbf{x}) = 0 \end{aligned}$$

Hence, we can get:

$$\begin{bmatrix} \mathbf{A}_1^* \\ \vdots \\ \mathbf{A}_r^* \end{bmatrix} .(\mathbf{x} - P.\mathbf{x}) = \mathbf{0}.$$

Hence, $A^*(\mathbf{x} - P.\mathbf{x}) = \mathbf{0}$, then $A^*(\mathbf{x} - A.\mathbf{c}) = \mathbf{0}$, xp $A^*.\mathbf{x} = A^*A\mathbf{c}$, which leads to $\mathbf{c} = (A^*A)^{-1}A^*.\mathbf{x}$, so:

$$P.\mathbf{x} = A.\mathbf{c} = A(A^*A)^{-1}A^*\mathbf{x} \text{ for all } \mathbf{x}.$$

□

Remark I.5.16. Special Cases with Projector.

If $\{\mathbf{A_1}, \cdots, \mathbf{A_r}\}$ are orthonormal, then $A^*A = r\,\mathrm{Id}$ and $P = AA^*$.
If $V = \mathrm{span}\{\mathbf{v}\}$, so $A = \begin{bmatrix} \mathbf{v} \end{bmatrix}$, then:

$$P = \mathbf{v}(\mathbf{v}^*\mathbf{v})^{-1}\mathbf{v}^* = \frac{\mathbf{v}\mathbf{v}^*}{\|\mathbf{v}\|^2},$$

which the orthogonal projection on a line. ⌟

I.6 QR Decomposition

Proposition I.6.1. Gram Schmidt for Normalization.

Let $\beta = \{\mathbf{A_1}, \mathbf{A_2}, \cdots, \mathbf{A_n}\}$ be a basis, the Gram-Schmidt process, gives $\gamma = \{\mathbf{q_1}, \mathbf{q_2}, \cdots, \mathbf{q_n}\}$, so we have:

$$\begin{aligned} \mathbf{u_1} &= \mathbf{A_1} \longrightarrow \mathbf{q_1} = \frac{\mathbf{u_1}}{\|\mathbf{u_1}\|} \\ \mathbf{u_2} &= \mathbf{A_2} - \frac{\mathbf{u}_1^*\mathbf{A_2}}{\|\mathbf{u_1}\|}\mathbf{u_1} \longrightarrow \mathbf{q_2} = \frac{\mathbf{u_2}}{\|\mathbf{u_2}\|} \\ \mathbf{u_3} &= \mathbf{A_3} - \frac{\mathbf{u}_1^*\mathbf{A_3}}{\|\mathbf{u_1}\|}\mathbf{u_1} - \frac{\mathbf{u}_2^*\mathbf{A_3}}{\|\mathbf{u_2}\|}\mathbf{u_2} \longrightarrow \mathbf{q_3} = \frac{\mathbf{u_3}}{\|\mathbf{u_3}\|}. \end{aligned}$$

For this part, our goal is to have given a matrix $A \in \mathbb{C}^{m\times n}$, we want to factor it as $A = QR$, where $Q \in \mathbb{C}^{m\times m}$ is unitary and $R \in \mathbb{C}^{m\times n}$ is upper triangular, *i.e.*:

$$\underbrace{\begin{bmatrix} \mathbf{A_1} & \cdots & \mathbf{A_n} \end{bmatrix}}_{\substack{A \\ \Cap \\ \mathbb{C}^{m\times n}}} = \underbrace{\begin{bmatrix} \mathbf{q_1} & \cdots & \mathbf{q_n} & \mathbf{q_{n+1}} & \cdots & \mathbf{1_n} \end{bmatrix}}_{\substack{Q \\ \Cap \\ \mathbb{C}^{m\times m}}} \circ \underbrace{\begin{bmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & * \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}}_{\substack{R. \\ \Cap \\ \mathbb{C}^{m\times n}}}$$

Remark I.6.2. Reprise to Gram Schmidt.

We can first use Gram Schmidt to make a basis orthogonal, and normality is trivial as normalizing in the induced normed vector space is straightforward. ⌟

Since with orthonormal basis, we have $\mathbf{v} \in \operatorname{span}(\beta)$ as for all $\mathbf{A_i} \in \operatorname{span} \beta$ that:

$$\mathbf{A_i} = \sum_{j=1}^{n} (\mathbf{q_j^*} \mathbf{A_j}) \mathbf{q_j} = \sum_{j=1}^{i} (\mathbf{q_j^*} \mathbf{A_j}) \mathbf{q_j} + 0 + \cdots + 0 = (\mathbf{q_1^*} \mathbf{A_1}) \mathbf{q_1} + (\mathbf{q_2^*} \mathbf{A_2}) \mathbf{q_2} + \cdots + (\mathbf{q_i^*} \mathbf{A_i}) \mathbf{q_i}.$$

Using the block multiplication, we have:

$$A = \begin{bmatrix} \mathbf{A_1} & \cdots & \mathbf{A_n} \end{bmatrix} = \begin{bmatrix} \mathbf{q_1} & \cdots & \mathbf{q_n} \end{bmatrix} \begin{bmatrix} \mathbf{q_1^*A_1} & \mathbf{q_2^*A_2} & \mathbf{q_3^*A_3} & \cdots & \mathbf{q_n^*A_n} \\ 0 & \mathbf{q_2^*A_2} & \mathbf{q_3^*A_3} & \cdots & \mathbf{q_n^*A_n} \\ 0 & 0 & \mathbf{q_3^*A_3} & \cdots & \mathbf{q_n^*A_n} \\ 0 & 0 & 0 & \cdots & \mathbf{q_n^*A_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{q_n^*A_n} \end{bmatrix}$$

Here, we may write this as the pseudo code that:

```
input: A_1, ..., A_n
output: q_1, ..., q_n

for j = 1, ..., n:
   u_j := A_j
   for i = 1, ..., j - 1: % There is nothing for j = 1
      r_ij := q_i^* A_j
      u_j := u_j - r_ij q_j
      % The above line has risk of Catastrophic Cancellation when subtracting similar number.
      % There may be large rounding of error.
   end
   q_j := u_j / ||u_j|| % Marked Step
   r_jj := a_j^* A_j
```

Remark I.6.3. Case when rank is less.

Let A has rank $r < n$, we should be pick arbitrary unit vectors orthogonal to the other basis at the marked step to keep going with Gram Schmidt. ⌟

We get $\beta = \{\mathbf{q_1}, \cdots, \mathbf{q_n}\}$ as a orthonormal set:

- when A is full rank, $\operatorname{im} A = \operatorname{span}(\beta)$, and
- in any case $\operatorname{im} A \subset \operatorname{span}(\beta)$.

Hence we have:

$$A = \tilde{Q}\tilde{R} = \begin{bmatrix} \mathbf{q_1} & \cdots & \mathbf{q})\mathbf{n} \end{bmatrix} \tilde{R}.$$

- We want to find $\{\mathbf{q_{n+1}}, \cdots, \mathbf{1_n}\}$ as orthonormal set orthogonal to β, with
- $\tilde{R}$ having $m - n$ zero rows.

In particular, the (classical) Gram Schmidt will introduce large rounding error compounding up. Here, we can introduce the modified Gram Schmidt for computing purposes.

Proposition I.6.4. Modified Gram Schmidt.

Given $\{\mathbf{A_1}, \mathbf{A_2}, \cdots, \mathbf{A_n}\}$, columns of $A = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_n} \end{bmatrix}$

(i) Initialize $\mathbf{u_1} = \mathbf{A_1}$ and $\mathbf{q_1} = \mathbf{u_1}/\|\mathbf{u_1}\|$.

(ii) Have ${\mathbf{u_2}}^{(1)} = \mathbf{A_2}$, with:

$$\mathbf{u_2} = {\mathbf{u_2}}^{(2)} = P_{\mathbf{q_1}}{\mathbf{u_2}}^{(1)} = (\mathrm{Id} - \mathbf{q_1}{\mathbf{q_1}}^*){\mathbf{u_2}}^{(1)} = {\mathbf{u_2}}^{(1)} - \mathbf{q_1}(\mathbf{q_1^*}\mathbf{u_2^{(1)}}) = {\mathbf{u_2}}^{(1)} - (\mathbf{q_1^*}\mathbf{u_2^{(1)}})\mathbf{q_1}.$$

(j) For the j-th step, we have:

$$\begin{aligned} \mathbf{u_j^{(1)}} &= \mathbf{A_j}, \\ \mathbf{u_j^{(2)}} &= \mathbf{u_j^{(1)}} - (\mathbf{q_1^*}\mathbf{u_j^{(1)}})\mathbf{q_1}, \\ &\vdots \\ \mathbf{u_j} = \mathbf{u_j^{(j)}} &= \mathbf{u_{j-1}^{(j-1)}} - (\mathbf{q_{j-1}^*}\mathbf{u_j^{(j-1)}})\mathbf{q_{j-1}}. \end{aligned}$$

```
for i = 1, ..., n:    % first loop to initialize all u_i's
   u_i := A_i

for i = 1, ..., n:    % calculating each step
   r_ii := ||u_i||
   q_i := u_i / ||u_i||
   for j = i + 1, ..., n:
      r_ij := q_i^* u_j
      u_j := u_j - r_ij * q_i
```

Definition I.6.5. Computational Complexity.

The complexity is measured by FLOPs: each operation is considered to operate with a unit time. ⌟

Example I.6.6. Complexity of Inner Product.

Recall `r_ ij := q_i^*u_j` is an inner product. In general, we consider $\mathbf{x}^*\mathbf{y}$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^m$, hence there will be m multiplications and m subtractions, hence it has $2m$. ⌟

Hence, if we consider the second iterated loop with `j` with the above pseudo-code, we have the total number of operations being:

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n} 4m = 4m\sum_{i=1}^{n}\sum_{j=i+1}^{n} 1 = 4m \cdot \frac{n^2 - n}{2} \sim 2mn^2.$$

If $A \in \mathbb{C}^{m\times m}$ is Hermitian, *i.e.*, $A^* = A$, then we have the orthogonal diagonalization that $Q^*AQ = D$, where Q is unitary. However, we want to have the decomposition for the more general A, that is:

Definition I.6.7. Householder Triangularization.

If $A \in \mathbb{C}^{m\times n}$ in which $m \geqslant n$, and sometimes we require $\dim(\operatorname{im} A)A = n$, then we have orthogonal triangularization that:

$$Q^*A = \begin{bmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & * \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} = R,$$

where $Q \in \mathbb{C}^{m\times m}$ is unitary. (This is equivalent with $A = QQ^*A = Q$, which is just QR *decomposition*). ⌟

The idea is that let:

$$Q^* = Q_n Q_{n-1} \cdots Q_2 Q_1,$$

with each Q_k being unitary, we have:

$$Q = (Q^*)^* = Q_1^* Q_2^* \cdots Q_n^*.$$

Example I.6.8. A 5×3 Matrix with Householder Triangularization.

Let A be a 5 by 3 that:

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

so by 3 steps of each Q_1, Q_2, and Q_3 to obtain that:

$$A = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \xmapsto{Q_1} \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix} \xmapsto{Q_2} \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix} \xmapsto{Q_3} \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where Q_1 changes entirely the first column, Q_2 changes all except the first row for the second column, and Q_3 changes all except the first two rows for the third column. ⌟

Recall that unitary ($Q^* = Q^{-1}$) $\Longleftrightarrow$ $\|Q.\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x}$, which is isometry. Hence, the structure of Q_n is:

$$Q_k = \begin{bmatrix} \mathrm{Id} & 0 \\ 0 & F, \end{bmatrix}$$

so we have:

$$\begin{bmatrix} \mathrm{Id} & 0 \\ 0 & F \end{bmatrix} \cdot \begin{bmatrix} T & B \\ 0 & X \end{bmatrix} = \begin{bmatrix} T & B \\ 0 & FX \end{bmatrix}.$$

Here, we want Q_k to be unitary, that is having orthonormal columns. So we need columns of F to be orthonormal, so equivalently, F must be an isometry on $\mathbb{C}^{m-(k-1)}$.

Here, we take $\mathbf{x} = \begin{bmatrix} * \\ \vdots \\ * \end{bmatrix} \in \mathbb{C}^{m-(k-1)}$, so we have $F.\mathbf{x} = \begin{bmatrix} \|\mathbf{x}\| \\ \vdots \\ 0 \end{bmatrix}$, so we have:

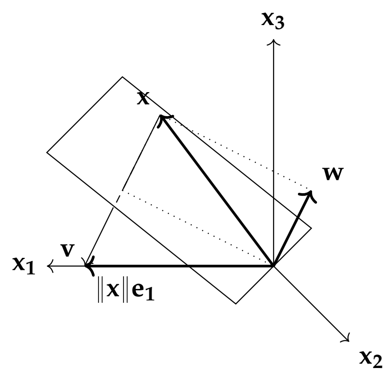


Figure I.12. Illustration of Householder reflection.

Hence, we have that $F.\mathbf{x} - x = \|\mathbf{x}\|\mathbf{e_1} - \mathbf{x}$, hence, we have:

$$\mathbf{w} = \text{orthogonal projection of } \mathbf{x} \text{ onto } \mathbf{v} = P_{\mathbf{v}}\mathbf{x} = \frac{\mathbf{v}\mathbf{v}^*}{\|\mathbf{v}\|^2}\mathbf{x},$$

so we have that $\mathbf{v} = -2\mathbf{w}$. Therefore, we have:

$$F.\mathbf{x} = \mathbf{x} - 2\mathbf{w} = \operatorname{Id}.\mathbf{x} - 2\frac{\mathbf{v}\mathbf{v}^*}{\|\mathbf{v}\|^2}\mathbf{x},$$

hence we have:

$$F = \operatorname{Id} - 2\frac{\mathbf{v}\mathbf{v}^*}{\|\mathbf{v}\|^2}.$$

Theorem I.6.9. Properties of Householder Reflection.

For the F being a Householder Reflection, we have the following properties:

(i) F is unitary,

(ii) F is Hermitian,

(iii) $F^{-1} = F$ or $F^2 = \operatorname{Id}$.

Remember that we have $Q_k = \begin{bmatrix} \operatorname{Id} & 0 \\ 0 & F \end{bmatrix}$, we have:

(i) Q_k is unitary,

(ii) Q_k Hermitian, and

(iii) Q_k is involuntary.

Remark I.6.10. Computational Complexity for Householder Triangularization.

We could haver defined:

$$F.\mathbf{x} = -\|\mathbf{x}\|\mathbf{e_i},$$

we choose $F.\mathbf{x}$ that is farthest away from $\mathbf{x}$, and the trick is choose $F.\mathbf{x} = -\text{sgn}(x_1)\|\mathbf{x}\|\mathbf{e_1}$. ⌟

So we have computational cost is $\sim 2mn^2 - \frac{2}{3}n^3$.

Example I.6.11. Householder QR Algorithm.

For the algorithm, we do:

```
for k = 1 ... n:
   x = A_k:m,k
   v_k = sgn(x_i)||x||e_1 + x
   v_k = v_k / ||v_k||
A_k:m,k:n = A_k:m,k:n - 2v_k(v_k^* A_k:m,k:n)
```

In particular, we have:

$$A^{(k)} = Q_k A^{(k-1)} = Q_k Q_{k-1} \cdots Q_1 A,$$

and:

$$A^{(n)} = R = Q_n Q_{n-1} \cdots Q_1 A.$$

We want to solve that $A.\mathbf{x} = \mathbf{b}$, which is equivalently $QR.\mathbf{x} = \mathbf{b}$, that is $R.\mathbf{x} = Q^*.\mathbf{b}$.
To calculate $Q^*\mathbf{b}$, we use:

```
for k = 1 ... n:
   b_k:m = b_k:m - 2v_kv_k^*b_k:m
```

And to compute $Q.\mathbf{u}$, we use:

```
for k = n ... 1:
   u_k:m = u_k:m' - 2v_kv_k^*u_k:m.
```

In particular $A_a : b, c : d$ meaning to obtain the ath to bth row and cth column to dth column of A. ⌟

Remark I.6.12. Operation Count for Householder QR Decomposition.

Operation count is dominated by innermost for loop, that is the last line in Householder algorithm.
Note that `A_k:m,j` has length $l = m - (k-1)$, lets do operator count in terms of l, that is:

- `A_k:m,j`−2`v_k(v_k`ˆ`*A_k:m,j)`.
- Dot Product: l multiplications and $l - 1$ additions.
- scalar multiplication: l multiplications.
- subtraction: l subtractions.

Hence, there are a total of $4l - 1$ flops. This meas that we do approximately 4 flops for every entry operated on.

Then to count the total number of entries. For the kth step, we have:

$$\begin{aligned}
\sum_{k=1}^{n} \big(m-(k-1)\big)\big(n-(k-1)\big) &= \sum_{k=1}^{n} \big(mn + (-m-n)(k-1) + (k-1)^2\big) \\
&= mn^2_(- m - n) \sum_{k=1}^{n} (k-1) + \sum_{k=1}^{n} (k-1)^2 \\
&= mn^2 + (-m-n)\frac{1}{2}n(n+1) + \frac{1}{2}n(n+1))(2n+1) \\
&\approx mn^2 - \frac{1}{2}mn^2 - \frac{1}{2}n^3 + \frac{1}{3}n^3 = \frac{1}{2}mn^2 - \frac{1}{6}n^3.
\end{aligned}$$

By multiplying 4, we have $2mn^2 - \frac{2}{3}n^3$. (Recall that this was $2mn^2$ for Gram-Schmidt process.) ⌟

II Applications with Computer Programming

II.1 `MATLAB` Preliminaries

The `MATLAB` programs can be helpful in conducting computations, and its embedded arrays allow linear algebra computations.

```
x = (-128:128)'/128;        % Create a column from -128 to 128 (inclusively) and normalize
A = [x.^0 x.^1 x.^2 x.^3]; % Create Matrix with each column being from -1 to 1 with 257 steps.
[Q,R] = qr(A,0);            % QR factorization

scale = Q(257,:);           % Calculate the scale
Q = Q*diag(1./scale);       % Modify Q via diagonal matrix
plot(x,Q);                  % Plot x against Q.
```

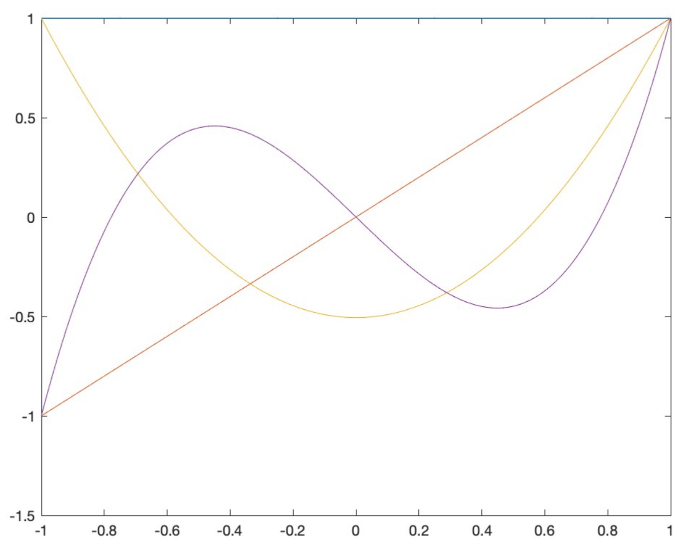


Figure II.1. `MATLAB` plot on the above code snippet.

Example II.1.1. Computation Error of Gram Schmidt.

First, we have the matrix as:

$$A - \begin{bmatrix} 0.70000 & 0.70711 \\ 0.70001 & 0.70711 \end{bmatrix}.$$

By keeping 5 digits, we have:

$$Q = \begin{bmatrix} 0.70710 & 1.0000 \\ 0.80811 & 0.0000 \end{bmatrix},$$

which is not very accurate. ⌟

Of course, we may implement the (traditional) Gram Schmidt algorithm through `MATLAB` function. This function on `MATLAB` takes in a matrix A and returns the factorization of Q and R.

```
function [Q,R] = clgs(A)
n = length(A);
R = zeros(n);
Q = zeros(n);
V = zeros(n);
  for j = 1:n
    V(:,j) = A(:,j);
    for i = 1:j-1
      R(i,j) = Q(:,i)'*A(:,j);
      V(:,j) = V(:,j) - R(i,j)*Q(:,i);
    end
  R(j,j) = norm(V(:,j),2);
  Q(:,j) = V(:,j)./R(j,j);
  end
end
```

II.2 Representation of Numbers

Consider the base 10 representation of 273, we have:

$$(273)_{10} = 2 \times 10^2 + 7 \times 10^1 + 3 \times 10^0.$$

Definition II.2.1. Base 2 Representation.

In base 2 representation, all numbers are represented by 0's and 1's. ⌟

For example, consider 100101 in binary, we have:

$$(100101)_2 = 1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 32 + 4 + 1 = 37.$$

Example II.2.2. Converting Base 10 to Base 2.

Consider $(156)_10$ and we want to make it base 2, we have:

$$\begin{aligned} 156 &= 78 \times 2 + 0, \\ 78 &= 39 \times 2 + 0, \\ 39 &= 19 \times 2 + 1, \\ 19 &= 9 \times 2 + 1, \\ 9 &= 4 \times 2 + 1, \\ 4 &= 2 \times 2 + 0, \\ 2 &= 1 \times 2 + 0, \\ 1 &= 0 \times 2 + 1. \end{aligned}$$

Hence, we have $(156)_{10} = (10011100)_2$. ⌟

Then, we consider the representation of floating point number. Without loss of generality, we represent numbers in $[0, 1)$.

Consider the base 10 number, we have:

$$(0.345)_{10} = 3 \times 10^{-1} + 4 \times 10^{-2} + 5 \times 10^{-3}.$$

For base 2 number, we consider:

$$(0.1101)_2 = 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-4} = 0.5 + 0.25 + 0.625 = (0.8125)_{10}.$$

Example II.2.3. Converting Base 10 Floating to Base 2.

Here, we consider converting base 10 floating points to base 2 floating points, namely $(0.1)_{10}$:

$$\begin{array}{l|l} 0.1 & 0. \\ 0.1 \times 2 = 0.2 < 1 & 0.0 \\ 0.2 \times 2 = 0.4 < 1 & 0.00 \\ 0.4 \times 2 = 0.8 < 1 & 0.000 \\ 0.8 \times 2 = 1.6 \geqslant 1 & 0.0001 \\ 0.6 \times 2 = 1.2 \geqslant 1 & 0.00011 \\ 0.2 \times 2 = 0.4 < 1 & 0.000110 \\ \vdots & \vdots \end{array}$$

Notice that there is a repeating pattern, so we have:

$$(0.1)_{10} = (0.0001100110011\cdots)_2 = (0.0\overline{0011})_2.$$

This is a repeating decimals. If we were to reconvert, we have:

$$\begin{aligned} \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^9 + \cdots &= \left[\left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^5\right]\left[1 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^8 + \cdots\right] \\ &= \left(\frac{1}{2}\right)^4 \left[1 + \frac{1}{2}\right] \frac{1}{1-(1/2)^4} = \frac{3}{2} \times \frac{1}{16} \times \frac{1}{15/16} = \frac{1}{10}. \end{aligned}$$

The above example can account for some issues with the floating point inaccuracies in the representation of numbers.

Definition II.2.4. `IEEE` Floating Point Representation.

For single precision, there are 32 bits, which is 4 bytes.

1 sign bit	`#e` 8 exponent bits	`#f` 23 mantissa bits

For double precision, there are 64 bits, which is 8 bytes.

1 sign bit	`#e` 11 exponent bits	`#f` 52 mantissa bits

Here we have the bits being stored in the continue memory locations. For converting to a base 10 number, we have:

$$N = (-1)^s(1+f)2^{e-127}.$$

Here, 127 is the bias for single precision, note that $1 + f \in [1.2)$.

Let $f = (0.m_1 \cdots m_{23})_2$, we have it as:

$$m_1\left(\frac{1}{2}\right)^1 + \cdots + m_{23}\left(\frac{1}{2}\right)^{23}.$$

- For $e = 127$, we just have $\pm(1.m_1 \cdots m_{23})_2$,

- For $e = 128$, we just have $\pm(1m_1m_2.\cdots m_{23})_2$, which has one less precision, that is 2^{-22}, and it is in $[2, 4)$,
- For $e = 129$, we just have $\pm(1m_1m_2m_3.\cdots m_{23})_2$, which has two less precision, and it is in $[4, 8)$.

For the smaller numbers, we have:

- For $e = 126$, we have $(0.1m_1 \cdots m_{23})_2$, so the precision is one more, that is 2^{-24}, the size of interval is $[1/2, 1)$.
- For $e = 127$, we have $(0.1m_1 \cdots m_{23})_2$, so the precision is two more, , the size of interval is $[1/4, 1/2)$.

Example II.2.5. Converting Base 10 to FP.

Consider converting from base 10 to FP, we have:

$$(15)_10 = (1111)_2,\ (0.1)_{10} = (0.0\overline{1111})_2,$$

thus we have:

$$(15.1)_{10} = 1111.0\overline{0011} = 1.1110\overline{0011} \times 2^3$$

Hence the exponent is $127 + 3 = 130$, and the mantissa being $1110\overline{0011}$, up to the correct number of digits:

1 sign bit	`#e` 8 exponent bits	`#f` 23 mantissa bits
`0`	`10000010`	`11100011001100110011001`

In this case, we truncated all the digits afterwards, causing imprecisions. ⌟

In particular, the catastrophic cancellation since we are not considering the digits afterwards.

Even with double precisions points, the we still have the base as:

$$(-1)^s(1 + f)\underbrace{2^{e-1023}}_{\text{bias}},$$

Definition II.2.6. Machine Representable Number.

A Machine representable number (MRN) is a number that can be represented exactly $f_p(x) = x$. ⌟

Let $f = (0.m_1m_2 \cdots m_{52})_2$, and we have the following case:

- For $e = 1023$, we have the values being in $[1, 2)$, with the coefficient of $(1/2)^{52} \simeq 2.220 \times 10^{-16}$.
- For $e = 1024$, we have the values being in $[2, 4)$, with the coefficient of $(1/2)^{51} \simeq 4.441 \times 10^{-16}$.
- For $e = 1025$, we have the values being in $[4, 8)$, with the coefficient of $(1/2)^{50} \simeq 8.882 \times 10^{-16}$.
- For $e = 1077$, we have the values being in $[2^{52}, 2^{53})$, with the coefficient of $(1/2)^0 = 1$.

For any generic e, we have the following:

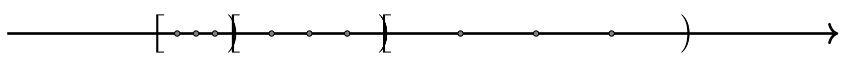
- Interval: $[2^{e-1023}, 2^{e-1022})$,
- Width of interval: $2^{e-1022} - 2^{e-1023} = 2^{e-1023}$,
- Step size: $2^{-52+e-1023} = 2^{e-1075}$, and

- Number of MRNs: 2^{-52}.

Example II.2.7. Largest and Smallest (Positive) MRN.

The largest MRN is $\underbrace{\texttt{111...1}}_{\text{52 ones}}\underbrace{\texttt{000...0}}_{\text{972 zeros}} \simeq 10^{304}$. Here, the step size is $10^{1024-52}$.
The smallest MRN is $0.\underbrace{\texttt{000...0}}_{\text{1022 zeros}}\texttt{1}\underbrace{\texttt{000...0}}_{\text{52 ones}} = 2^{-1023} \simeq 2.470 \times 10^{-324}$. (This is evaluated to zero on computers).
The next smallest (real smallest) is $2^{-1023} + 2^{-(1023+52)}$.

For each fixed e, it represents an interval, which got larger when e grows larger, with the numbers inside being uniformly distributed along each interval:



Remark II.2.8. Zero, Infinity, and NaN in Machine Representation.

For floating numbers, 0, $\pm\infty$, and NaN (not a number) cannot be represented conventionally, but they are distributed with a special slot.

- For 0, we have:

1 sign bit	`#e` 11 exponent bits	`#f` 52 mantissa bits
`0` or `1`	`000...0`	`000...0`

- For ∞, we have:

1 sign bit	`#e` 11 exponent bits	`#f` 52 mantissa bits
`0` or `1`	`111...1`	`000...0`

  For the $-\infty$, it makes the sign bit as negative.

- For NaN, we have:

1 sign bit	`#e` 11 exponent bits	`#f` 52 mantissa bits
`0` or `1`	`111...1`	`111...1`

Definition II.2.9. Machine Epsilon.

The machine epsilon, denoted $\epsilon_{\text{machine}}$, is the distance between 1 and the next larger MRN.

Example II.2.10. Machine Epsilon for 1.

For the double precision floating point, we represent the number 1 as:

1 sign bit	`#e` 11 exponent bits	`#f` 52 mantissa bits
`0`	`01111111111`	`000...0`

where the exponent is 1023 in base 10.
Therefore, the next MRN is:

1 sign bit	#e 11 exponent bits	#f 52 mantissa bits
0	01111111111	000...01

which is exactly $1 + 2^{-52}$, or the step size for $e = 1023$ is always 2^{-52}. Therefore:

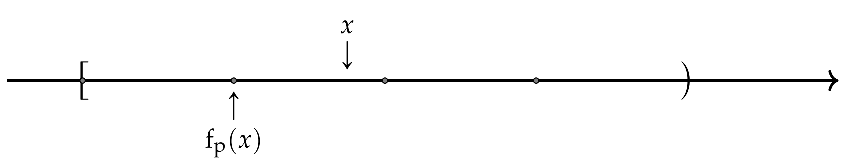
$$\epsilon_{\text{machine}} = 2^{-52}.$$

Remark II.2.11. Absolute and Relative Error.

Let x be a positive real number, we have:

- Let $\mathrm{f_p}(x)$ be the floating point representation of x.
- We assume that we are in double precision.

From x, you first determine e, and then we find the number on the interval (for simplicity, we just represent 4 points on the number line):



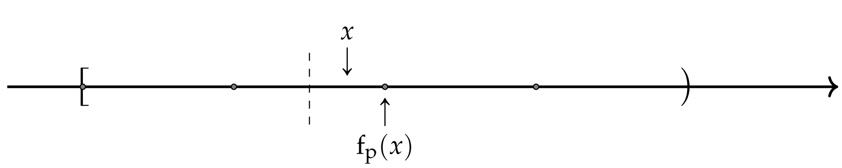
Suppose we have just truncation, we have:

- The absolute error is $|x - \mathrm{f_p}(x)| \leqslant 2^{-52+e-1024}$, which is the step size.
- The relative error is $\dfrac{|x - \mathrm{f_p}(x)|}{|\mathrm{f_p}(x)|} \leqslant \dfrac{2^{-52+e-1024}}{|(-1)^s(1+f)2^{e-1024}|} = \frac{2^{-52}}{1+f} \leqslant \epsilon_{\text{machine}}$.

The key conclusion is that the relative error is bounded above by the machine epsilon.
Alternatively, if we consider the numerator of the relative error as x, we have:

$$\frac{|x - \mathrm{f_p}(x)}{|x|} = \frac{|x - \mathrm{f_p}(x)|}{|\mathrm{f_p}(x)|} \cdot \frac{|\mathrm{f_p}(x)|}{|x|} \leqslant \epsilon_{\text{machine}}.$$

Then, we consider the rounding off, so we have:



Now, the absolute error is halved, and relative error is bounded by $\epsilon_{\text{machine}}/2$.

Here, we shall be concerned on how we can trust the computer.

For double precisions floating point, we have:

$$\epsilon_{\text{machine}} = 2^{-52} \simeq 2.22045 \times 10^{-16}.$$

So we can keep about 15 to 16 digits accurate

Example II.2.12. `MATLAB` Illustration on Numbers.

Below is a code segment of the illustration on `MATLAB`

```
>> 1.2345678901234567890123456789O
ans = 1.23456789012346

>> ans + 100000000
ans = 1.000000012345679e+008

>> ans - 100000000
ans = 1.234567890553165
```

Note that for the second operation, the operation truncates the digits after the first 9, and when the same number is subtracted, it results in some junk digits.
More specifically, look at the following code snippet:

```
>> 1.2345678901234567890 + 10000000 - 10000000
ans = 1.234567890553165

>> 1.2345678901234567890 + (10000000 - 10000000)
ans = 1.234567890123456
```

Here, we can observe that associativity does not hold over computer level computations. ⌟

This is caused by the similarity of significant digits, so only the first 15-th digits are exact, and from 16-th digits afterwards, they may be affected by round-off.

This is an example of the *Catastrophic Cancellation*, which occurs when computing $x - y$, where $x > y$ but $x \simeq y$.
There, $x - y$ can result in fewer significant digits than x and/or y.

Example II.2.13. Trick to Eliminate Subtraction.

When we compute:

$$y = \frac{\sqrt{x^2+4} - x}{2} \text{ for } x \text{ very large, and } x > 0.$$

For $x = 10^{10.5}$, `MATLAB` returns $y = 0$. However, we can have:

$$y = \frac{\sqrt{x^2+4} - x}{2} \cdot \frac{\sqrt{x^2+4} + x}{\sqrt{x^2+4} + x} = \frac{x^2+4-x^2}{2(\sqrt{x^2+4}+x)} = \frac{2}{\sqrt{x^2+4}+x},$$

and without subtractions inside, we have the output approximately as 3.162278×10^{-11}. ⌟

Remark II.2.14. Real Number and Floating Point.

Let $x \in \mathbb{R}$, and $\mathrm{f_p}(x)$ be the floating point representation of x, which is a MRN, and it is defined as:

$$\epsilon_{\text{machine}} = (-1)^s(1+f)2^{e-1023}.$$

The machine epsilon is determined as the absolute value of the difference between 1 and the next larger MRP, that is 2^{-52} for double precision floating point number.
The relative error in floating point approximating is:

$$\frac{|x - \mathrm{f_p}(x)|}{|x|}.$$

Whereas for truncation, we redefine:

$$\epsilon_{\text{machine}} = \frac{1}{2(1-2^{-52})}2^{-52}.$$

Thus, in both cases, we have:

$$\frac{|x - \mathrm{f_p}(x)|}{|x|} \leqslant \epsilon_{\text{machine}}.$$

⌟

Below, we introduce a "axiom" of floating point representations, but it turns out to be direct from the previous remark.

Proposition II.2.15. Property of Floating Point Representation.

For all $x \in \mathbb{R}$, there exists ε (positive or negative) with $|\varepsilon| \leqslant \epsilon_{\text{machine}}$ such that:

$$\mathrm{f_p}(x) = x(1+\varepsilon).$$

I.e., the relative distance between x and $\mathrm{f_p}(x)$ is always smaller than $\epsilon_{\text{machine}}$.

Proof. Without loss of generality, we let $x > 0$, since $x < 0$ is a similar case. We have:

$$\begin{aligned}
-\epsilon_{\text{machine}}x \leqslant \mathrm{f_p}(x) - x &\leqslant \epsilon_{\text{machine}}x, \\
x - \epsilon_{\text{machine}}x \leqslant \mathrm{f_p}(x) &\leqslant x + \epsilon_{\text{machine}}x, \\
x(1-\epsilon_{\text{machine}}) \leqslant \mathrm{f_p}(x) &\leqslant x(1+\epsilon_{\text{machine}}), \\
1 - \epsilon_{\text{machine}} \leqslant \frac{\mathrm{f_p}(x)}{x} &\leqslant 1 + \epsilon_{\text{machine}}.
\end{aligned}$$

Therefore, there exists ε with $|\varepsilon| \leqslant \epsilon_{\text{machine}}$ such that:

$$\frac{\mathrm{f_p}(x)}{x} = 1 + \varepsilon,$$

so we have $\mathrm{f_p}(x) = x(1+\varepsilon)$, as desired. □

Then, we think about the floating point arithmetic.

Remark II.2.16. Notation for Arithmetics.

Here, we have $\mathbb{R}$ denote the real numbers, and $\mathcal{F}$ denote the machine representable numbers in floating point with double precision. We think of $\mathrm{f_p} : \mathbb{R} \to \mathcal{F}$.
The four main arithmetic operations on $\mathbb{R}$ are $+$, $-$, $\times$, and $\div$.
The four main arithmetic operations on $\mathcal{F}$ are $\oplus$, $\ominus$, $\otimes$, and $\oslash$. ⌟

Here, let $*$ be the generic operator, since we have $\mathcal{F} \subset \mathbb{R}$, we defined the maps in $\mathcal{F}$ naturally pre-composing the inclusion $\iota : \mathcal{F} \hookrightarrow \mathbb{R}$ and post-compose with the $\mathrm{f_p}$. Therefore, for any $x', y' \in \mathcal{F} \subset \mathbb{R}$, we defined the operation as:

$$x' \circledast y' = \mathrm{f_p}\big(\iota(x') * \iota(y')\big) = \mathrm{f_p}(x' * y').$$

Here, by the property of floating point representation, we have:

$$x' \circledast y' = \mathrm{f_p}(x' * y') = (x' * y') \cdot (1 + \varepsilon),$$

where $|\varepsilon| \leqslant \epsilon_{\text{machine}}$.

Definition II.2.17. Big O Notation.

Given two functions of real-valued inputs, $a(t)$ and $b(t) > 0$, we have $a(t) = \mathcal{O}\big(b(t)\big)$ as $t \to 0$ when there exists $C > 0$ such that $|a(t)| \leqslant C \cdot b(t)$ in a neighborhood of $t = 0$, *i.e.*, there exists some $\delta > 0$ such that the statement holds for all $|t| < \delta$. ⌟

Here, we can have an example with some function.

Example II.2.18. Sine Function in Big O.

$\sin t = \mathcal{O}(|t|)$ as $t \to 0$, since we have $|\sin t| \leqslant |t|$ in $B_\delta(0)$ for some $\delta > 0$. ⌟

Remark II.2.19. Floating Point with Machine Epsilon.

If $x' = \mathrm{f_p}(x)$, then:

$$\frac{|x - x'|}{|x|} < 1 \cdot \epsilon_{\text{machine}},$$

thus, we have:

$$\frac{|x - x'|}{|x|} = \mathcal{O}(\epsilon_{\text{machine}}).$$

We say that the relative error is of the order of $\epsilon_{\text{machine}}$. ⌟

III Computational Methods

III.1 Least Square Approximations

Consider $A \in \mathbb{C}^{m\times n}$ with $A = \begin{bmatrix} \mathbf{A_1} & \mathbf{A_2} & \cdots & \mathbf{A_n} \end{bmatrix}$.

Remark III.1.1. Equivalence Conditions to Trivial Kernel.

For the above A, we have $\ker A = \{\mathbf{0}\} \iff \mathbf{A_1}, \cdots, \mathbf{A_n}$ are linearly independent $\iff \dim(\operatorname{im} A) = n$. In such case, we have $m \geqslant n$, that is, it must be square or having more rows. ⌟

Remark III.1.2. Existence and Uniqueness of Solutions to Linear Equation.

From linear algebra, with $A.\mathbf{x} = \mathbf{b}$ for matrix A and vectors $\mathbf{x}, \mathbf{b}$, we have:

(i) If $b \in \operatorname{im} A$, then there exists (at least) a solution to the linear equation.

(ii) If $\ker A = \{\mathbf{0}\}$, the uniqueness is guaranteed.

If $\mathbf{b} \notin \operatorname{im} A$, then for all $\mathbf{x}$, we have $A.\mathbf{x} \neq \mathbf{b}$. ⌟

Our goal is to choose $\mathbf{x} \in \mathbb{C}^n$ so that $A.\mathbf{x} \simeq \mathbf{b}$, *i.e.*, $A.\mathbf{x}$ is as close to $\mathbf{b}$ as possible.

This leads to the least square problem: Given a matrix $A \in \mathbb{C}^{m\times n}$ and a vector $\mathbf{b} \in \mathbb{C}^m$, we find $\mathbf{x}$ that minimizes the residue:

$$\epsilon = \|\mathbf{b} - A.\mathbf{x}\|_2$$

We are having the following assumptions:

(i) $m \geqslant n$, and

(ii) $\dim(\operatorname{im} A) = n \iff \ker A = \{\mathbf{0}\}$.

Recall that we have $\ker A = \ker(A^*A)$, so $\ker A = \{\mathbf{0}\}$ implies that A^*A is invertible.

For the first case, we consider $\mathbf{b} \in \operatorname{im} A$, so we have $A.\mathbf{x} = \mathbf{b}$ being consistent but overdetermined, *i.e.*, there are more equations than variables. The following is a illustration when $n = 2$:

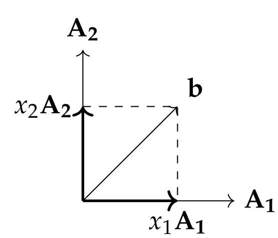


Figure III.1. Overdetermined equation in $n = 2$ plane for image.

Here, we then have $A.\mathbf{x} = \mathbf{b}$ implying that $A^*A.\mathbf{x} = A^*.\mathbf{b}$, in which A^*A is invertible, hence $\mathbf{x} = (A^*A)^{-1}A^*.\mathbf{b}$, in which we have:

$$A^+ = (A^*A)^{-1}A^* \in \mathbb{C}n \times m,$$

which is the *pseudo-inverse* of A.

Proposition III.1.3. Properties of Pseudo-inverse.

For a *pseudo-inverse* of A, denoted A^+, the following properties hold:

(i) $A^+A = (A^*A)^{-1}A^*A = \mathrm{Id}_n$,

(ii) $AA^+ = A(A^*A)^{-1}A^* = P$.

Then, we shall consider the other case, *i.e.*, if $\mathbf{b} \notin \operatorname{im} A$, then the minimizer of $\inf_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{b} - A.\mathbf{x}\|_2$ is still $\mathbf{x} - A^+.\mathbf{b}$, where $\mathbf{x}$ is the vector of coordinates of $\mathbf{y}$ with respect to $\mathbf{A_1}, \cdots, \mathbf{A_n}$.

- For a generic $\mathbf{x} \in \mathbb{C}^n$, we have $\mathbf{z} = A.\mathbf{x} \in \operatorname{im} A$, and
- For any $\mathbf{z} \in \operatorname{im} A = \operatorname{span}\{\mathbf{A_1}, \cdots, \mathbf{A_n}\}$, there exists $\mathbf{x} \in \mathbb{C}^n$ such that $\mathbf{z} = A.\mathbf{x}$.

Remember that we want to minimize $\inf_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{b} - A.\mathbf{x}\|_2$, which is equivalent to minimizing $\inf_{\mathbf{z}\in\operatorname{im} A} \|\mathbf{b} - \mathbf{z}\|_2$, which can be considered as follows:

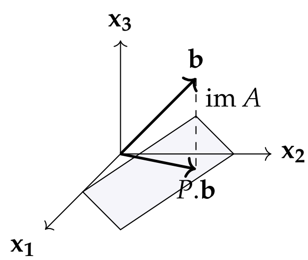


Figure III.2. The projection onto the image span.

Proposition III.1.4. Minimizer of the Approximation Problem.

Let $A \in \mathbb{C}^{m\times n}$ with $m \geqslant n$, we have $\dim(\operatorname{im} A) = n$. We fix $\mathbf{b} \in \mathbb{C}^m$ arbitrarily. The minimizer of the problem is:

$$\inf_{\mathbf{z}\in\operatorname{im} A} \|\mathbf{b} - \mathbf{z}\|_2$$

is $\mathbf{z} = \mathbf{y}$, where $\mathbf{y} = P.\mathbf{b} = AA^+.\mathbf{b}$.
Equivalently, the minimizer of $\inf_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{b} - A.\mathbf{x}\|_2$ is $\mathbf{x} = A^+\mathbf{b}$.

Such situation can be applied onto linear regression in statistics.

Example III.1.5. Linear Regression Example.

Let data points in $\mathbb{R}^2$ be:

$$\{(x_1, y_1), \cdots, (x_m, y_m)\},$$

so the linear model can be considered as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} .$$

When $\beta_0, \beta_1, \beta_2$ are fixed, for each x_i, the i-th prediction is:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$

and the i-th residue is:

$$\epsilon_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2),$$

so the sum of the square errors is:

$$\mathrm{SSE}(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right]^2.$$

Here, we use the trick by:

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

Here, we have:

$$\mathbf{b} - A.\boldsymbol{\beta} = \begin{bmatrix} y_1 - (\beta_0 + \beta_1 x_1 + beta_2 x_2^2) \\ \vdots \\ y_m - (\beta_0 + \beta_1 x_m + beta_2 x_m^2) \end{bmatrix}.$$

Therefore, we have:

$$\|\mathbf{b} - A.\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right]^2 = \mathrm{SSE}(\beta_0, \beta_1, \beta_2),$$

so we are minimizing $\hat{\boldsymbol{\beta}} = A^+.\mathbf{b}$. ⌟

Theorem III.1.6. Projections Minimize the Least Square Problem.

Let $A \in \mathbb{C}^{m \times n}$, with $m \geqslant n$ have $\dim(\operatorname{im} A) = n$ (full rank). Let $\mathbf{b} \in \mathbb{R}^m$ ne arbitrary, the minimizer of the problem is:

$$\min_{\mathbf{z} \in \operatorname{im} A} \|\mathbf{b} - \mathbf{z}\|_2,$$

is $\mathbf{z} = \mathbf{y}$, where:

$$\mathbf{y} = P\mathbf{b} = AA^+\mathbf{b} = A(A^*A)^{-1}A^*\mathbf{b}.$$

Equivalently, the minimizer of the problem $\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{b} - A.\mathfrak{x}\|_2$ is $\mathbf{x} = A^+.\mathbf{b}$.

The full rank condition is helpful when trying to solve $A.\mathbf{x} = \mathbf{b}$, when $\mathbf{b} \notin \operatorname{im} A$, so for all $\mathbf{x}$, $A.\mathbf{x} \neq \mathbf{b}$.

Remark III.1.7. Computational Costs for Matrix Operations.

The computational cost to compute A^+ is:

- $C(A^*A) \sim 2mn^2$,
- $C\big((A^*A)^{-1}\big) \sim \frac{8}{3}n^2$, and
- $C\big((A^*A)^{-1}A^*\big) \sim 2mn^2$.

Hence, the total computation cost for the projection for least square problem is $4mn^2 + \frac{8}{3}n^3$. ⌟

Note that there are other ways to compute $\mathbf{x} = A^+.\mathbf{b}$ by:

(i) Cholesky factorization,

(ii) QR factorization, and

(iii) SVD.

In particular, we have the QR factorization as:

$$\begin{array}{ccccc} A & = & Q & \circ & R, \\ \Cap & & \Cap & & \Cap \\ \mathbb{C}^{m\times n} & & \mathbb{C}^{m\times m} & & \mathbb{C}^{m\times n} \end{array}$$

where R has the top n rows being upper triangular.

Moreover, we may use the reduced QR decomposition that $A = \tilde{Q}\tilde{R}$ where we reduce to the n columns for Q and R. However, we have:

$$\tilde{Q}^*\tilde{Q} = \begin{bmatrix} \mathbf{q_1^*} \\ \vdots \\ \mathbf{q_n^*} \end{bmatrix} \begin{bmatrix} \mathbf{q_1^*} & \cdots & \mathbf{q_n^*} \end{bmatrix} = \mathrm{Id}\,.$$

Also, we have $\operatorname{span}\{\mathbf{q_1}, \cdots, \mathbf{q_n}\} = \operatorname{span}\{\mathbf{A_1}, \cdots, \mathbf{A_n}\}$, so $P = \tilde{Q}(\tilde{Q}^*\tilde{Q})^{-1}\tilde{Q}^* = \tilde{Q}\tilde{Q}^*$.

We let $\mathbf{y}$ be the orthogonal projection of $\mathbf{b}$ onto $\operatorname{im} A$, that is:

$$\mathbf{y} = P\mathbf{b} = \tilde{Q}\tilde{Q}^*\mathbf{b},$$

and let $\mathbf{x}$ be the coordinates of $\mathbf{y}$ with respect to $\{\mathbf{A_1}, \cdots, \mathbf{A_n}\}$, so:

$$\mathbf{y} = \sum_{i=1}^{n} x_i \mathbf{A_i} = A.\mathbf{x}.$$

Therefore, we have $A.\mathbf{x} = \tilde{Q}\tilde{Q}^*\mathbf{b}$, that is $\tilde{Q}\tilde{R}\mathbf{x} = \tilde{Q}\tilde{Q}^*\mathbf{b}$, by post-compose $\tilde{Q}^*$, we can get $\tilde{R}.\mathbf{x} = \tilde{Q}^*\mathbf{b}$, which is easy to compute.

Remark III.1.8. Algorithm and Cost of the Minimization Process.

The inputs are A and $\mathbf{b}$, and the output is $\mathbf{x}$, where $\mathbf{x} = A^+.\mathbf{b}$, we do the following:

(i) compute the reduced QR decomposition of A (with the householder triangularization, its complexity is $2mn^2 - \frac{2}{3}n^3$),

(ii) compute $\mathbf{c} = \tilde{Q}^*\mathbf{b}$ (there are n dot products in $\mathbb{C}^m$, so a total of about $2mn$), and then

(iii) solve $\tilde{R}.\mathbf{x} = \tilde{Q}^*.\mathbf{b}$ for $\mathbf{x}$ (for this part, note that the number of FLOP each block substitution is $1 + 3 + 5 + \cdots$, that is n^2).

Therefore, the computational cost is $\sim 2mn^2 - \frac{2}{3}n^3$. ⌟

III.2 Conditioning and Condition Number

Definition III.2.1. Ill Conditioned.

A problem is "ill conditioned' when a small variation of data causes large variation of solution. ⌟

Example III.2.2. Ill Conditioned Problem.

Consider $A.\mathbf{x} = \mathbf{b}$, it is either:

(i) $A \in \mathbb{C}^{m\times m}$ is invertible, then $\mathbf{x} = A^{-1}.\mathbf{b}$, or

(ii) $A \in \mathbb{C}^{m\times n}$, $\dim(\operatorname{im} A) = n$, and $\mathbf{b} \in \operatorname{im} A$, so $\mathbf{x} = A^{+}.\mathbf{b}$.

Suppose we compute a solution $\tilde{\mathbf{x}}$ (not quite correct), and the error is:

$$\mathbf{e} = \delta\mathbf{x} = \mathbf{x} - \tilde{\mathbf{x}},$$

which is the difference between the real and computed solution to $A.\mathbf{x} = \mathbf{b}$. In reality, this cannot be computed since we do not have access to the real solution (that is why we are computing it.)
Hence, the goal is to find an upper bound for relative size of error, *i.e.*:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}.$$

What we can compute is the residue:

$$\mathbf{r} = \delta\mathbf{b} = \mathbf{b} - A.\tilde{\mathbf{x}} = A.\mathbf{x} - A.\tilde{\mathbf{x}} = A(\mathbf{x} - \tilde{\mathbf{x}}) = A\delta\mathbf{x}.$$

Therefore, we have $\delta\mathbf{x} = A^{-1}\delta\mathbf{b}$, hence its norm is:

$$\|\delta\mathbf{x}\| = \|A^{-1}\delta\mathbf{b}\| \leqslant \|A^{-1}\| \cdot \|\delta b\|.$$

Then, since we have $\mathbf{b} = A.\mathbf{x}$, we have:

$$\|\mathbf{b}\| = \|A.\mathbf{x}\| \leqslant \|A\| \cdot \|x\|,$$

Therefore, we may obtain the upper bound of the residual as:

$$\underbrace{\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}}_{\text{relative size of error}} \leqslant \underbrace{\|A\| \cdot \|A^{-1}\|}_{\text{conditional number of the matrix } A} \cdot \underbrace{\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}}_{relativesizeofresidual}.$$

⌟

Definition III.2.3. Condition Number of Matrix.

Let A be an invertible matrix, we have the condition number of A as:

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|.$$

⌟

The large condition number means ill-conditioned, and small condition number means well conditioned.
If A^{-1} is hard to compute, then one eigenvalue is $\simeq 0$, which implies large $\kappa(A)$.

When $\|\bullet\| = \|\bullet\|_2$, we have:

$$\kappa(A) = \frac{\sigma_1}{\sigma_m}.$$

Here are some motivation of the condition number of a square, invertible matrix A:

$$\kappa(A) = \underbrace{\|A\| \cdot \|A^{-1}\|}_{\textbf{norm induced by vector norm}}.$$

For example, when having the 2×2 case, we ant to solve that $A.\mathbf{x} = \mathbf{b}$, and the solution is $\mathbf{x} = A^{-1}.\mathbf{b}$, where as the inverse is:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies A^{-1} = \frac{1}{ad - bc} \underbrace{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}_{\mathrm{adj}(A)}.$$

Note that when A is close to singularity, that is, $\det A$ is close to zero, then we have catastrophic cancellation, which happens when subtracting 2 numbers very close to each other. *This is a large round-off error*.

In the $m \times m$ case, we de the inverse as:

$$\left[\begin{array}{c|c} A & \mathrm{Id}_n \end{array}\right] \xrightarrow{\text{RREF}} \left[\begin{array}{c|c} \mathrm{Id}_n & A^{-1} \end{array}\right]$$

The gist is that when A is close to singularity, A^{-1} cannot be computed accurately, so $\mathbf{x} = A^{-1}\mathbf{b}$ cannot be computed accurately.

Let $\tilde{\mathbf{x}}$ be the estimate of the solution $\mathbf{x}$, we can compute the residual as:

$$\delta\mathbf{b} = \mathbf{b} - A.\tilde{\mathbf{x}} = A.\mathbf{x} - A.\tilde{\mathbf{x}} = A.(\underbrace{\mathbf{x} - \tilde{\mathbf{x}}}_{\delta\mathbf{x} \text{ error}}).$$

Remark III.2.4. Computation of $\delta\mathbf{x}$.

Here, we have $\delta\mathbf{x} = A^{-1}\delta\mathbf{b}$, since we cannot compute $\delta\mathbf{x}$ accurately. ⌟

The general rule for numerical linear algebra is to avoid computing $\det A$ and A^{-1}, because they are:

(i) computationally intensive, and

(ii) computationally inaccurate when A is close to singularity.

Example III.2.5. Pseudo Inverse Case.

If $A \in \mathbb{C}^{m \times n}$ such that $m \geqslant n$, with full rank and $\mathbf{b} \in \mathbb{C}^m$, then the minimizer of $\|\mathbf{b} - A.\mathbf{x}\|_2$ is:

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

We avoided inverting A^*A by computing $\mathbf{x}$ via QR decomposition by solving $\tilde{R}.\mathbf{x} = \tilde{Q}^*.\mathbf{b}$. ⌟

Later we will see techniques to compute:

- the eigenvalues of square matrices, and
- the SVD of any matrix,

that avoid determinant and inverse.

We have proven that:

$$\underbrace{\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}}_{\substack{\text{relative size of error} \\ \text{(cannot compute)}}} \leqslant \underbrace{\|A\| \cdot \|A^{-1}\|}_{\substack{\kappa(A)\text{conditional number} \\ \text{(typically can computed} \\ \text{without } A^{-1})}} \cdot \underbrace{\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}}_{\substack{\text{relative size of residual} \\ \text{(can compute)}}}.$$

Remark III.2.6. Conditional Number based on Choice of Vector Norm.

$\kappa(A)$ depends on the choice of the vector norm. For example, choose $\|\bullet\|_2$ as a 2-norm, we have:

$$\kappa_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2.$$

Theorem III.2.7. Conditional Number as Fraction of Singular Value.

For any invertible $A \in \mathbb{C}^{m\times m}$:

$$\kappa(A) = \frac{\sigma_1}{\sigma_m},$$

where σ_1 and σ_m are, respectively, the largest and smallest singular value of A.

Proof. For any $A \in \mathbb{C}^{m\times m}$, we have $\|A\|_2 = \sigma_1$, that is the largest eigenvalue.
When writing SVD as:

$$A = U\Sigma V^* = U \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m \end{bmatrix} V^*.$$

Then:

$$A^{-1} = (V^*)^{-1}\Sigma^{-1}U^{-1} = \overline{V} \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_m \end{bmatrix} (\overline{U})^* = U_1 \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_m \end{bmatrix} V_1^*.$$

Note that since $V^{-1} = V^*$, then $V = \overline{V}$, so their complex conjugates are the same, now, we may invert the order of the rows to obtain that:

$$A^{-1} = U_2 \begin{bmatrix} 1/\sigma_m & 0 & \cdots & 0 \\ 0 & 1/\sigma_{m-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_1 \end{bmatrix} V_2^*,$$

hence we have the largest singular value is $1/\sigma_m$. The conditional number follows as $\sigma_1 \cdot \frac{1}{\sigma_m} = \frac{\sigma_1}{\sigma_m}$. □

In particular, the fraction is called the eccentricity of ellipsoid with semi-axes $\sigma_1, \cdots, \sigma_m$.

Example III.2.8. Conditional Number for Almost Singular Matrices.

In `MATLAB`, the command for conditional number is `cond(A)`, which is $\kappa_2(A)$. Here, we let:

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

- When $\alpha = 1$, we have $\kappa_2(A) \approx 6.85410$,
- When $\alpha = 10^{-5}$, we have $\kappa_2(A) \approx 4.00002 \times 10^5$,
- When $\alpha = 10^{-12}$, we have $\kappa_2(A) \approx 3.99949 \times 10^{12}$.

Theorem III.2.9. Conditional Number of Matrix Operations.

For any matrix norm induced by a vector norm, we have:

(i) For any nonzero constant $c \in \mathbb{C}$, we have that:

$$\kappa(cA) = \kappa(A).$$

(ii) $\kappa(\mathrm{Id}) = 1$, and

(iii) for any invertible A, $\kappa(A) \geqslant 1$.

Proof. Recall that $\|\,\mathrm{Id}\,\| = 1$ and $\|A^{-1}\| \geqslant 1/\|A\|$ for any norm.

(i) Note that $(cA)^{-1} = 1/c \cdot A^{-1}$, so $\kappa(cA) = \|cA\| \cdot \|1/c \cdot A^{-1}\| = |c|/|c| \cdot \|A\| \cdot \|A^{-1}\| = \kappa(A)$.

(ii) $\kappa(\mathrm{Id}) = \|\,\mathrm{Id}\,\| \cdot \|\,\mathrm{Id}^{-1}\,\| = 1 \cdot 1 = 1$.

(iii) $\kappa(A) = \|A\| \cdot \|A^{-1}\| \geqslant \|A\|/\|A\| = 1$, so $\kappa(A) \geqslant 1$. □

III.3 Stability

For the mathematical problem, we let it be defined as $f : B \to X$, where B is the set of possible data and X is the set of solutions.

Example III.3.1. Basic Linear Algebra Problem.

We are trying to solve that $A.\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{C}^{m\times m}$ is invertible.
The solution is:

$$\mathbf{x} = f(\mathbf{b}) = A^{-1}.\mathbf{b}.$$

⌟

Typically (especially when the condition number $\kappa(A)$ is large), we can compute only an approximated version of $\mathbf{x}$, denoted $\tilde{\mathbf{x}}$, through an algorithm (such as QR decomposition).

Definition III.3.2. Algorithm.

Here, we define an algorithm as:

$$\tilde{f} : B \to X,$$

in which the computed solution $\tilde{\mathbf{x}} = \tilde{f}(\mathbf{b})$ is by the actual data. ⌟

A "stable" algorithm is one which computes solutions $\tilde{\mathbf{x}}$ which are approximately equal to the exact solution for slightly perturbed data:

$$\tilde{f}(\mathbf{b}) \simeq f(\mathbf{b} + \delta\mathbf{b}).$$

Definition III.3.3. Backward Stable.

Given a problem f, an algorithm $\tilde{f}$ is called backward stable if for all set of data $\mathbf{b} \in B$, there exists a data perturbation $\delta\mathbf{b}$, where:

$$\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \mathcal{O}(\epsilon_{\text{machine}}) \text{ such that } \tilde{f}(\mathbf{b}) = f(\mathbf{b} + \delta\mathbf{b}).$$

This stability is called backward since instead of looking at the *forward error*, we have:

$$\delta \mathbf{x} = \tilde{f}(\mathbf{b}) - f(\mathbf{b}),$$

we look backwards to see what input could have produced the computed result $\tilde{f}(\mathbf{b})$ exactly.

This definition holds , the data perturbation required to explain the computed solution is relatively small (relative to the problem's data $\mathbf{b}$), *i.e.*, the algorithm is numerically robust (stable), to relatively small perturbations.

Example III.3.4. Subtraction is Backwards Stable.

Suppose the mathematical problem is:

$$f : \mathbb{C}^2 \to \mathbb{C}, \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \mapsto b_1 - b_2.$$

The algorithm is as follows:

```
input: b = [b_1, b_2]
output: ~x = ~f(b_1, b_2) = fp(b_1) (-) fp(b_2) = fp(b1' - b2')
```

Then, there exists ϵ_1 with $|\epsilon_1| \leqslant \epsilon_{\text{machine}}$ such that $b_1' = \mathrm{f_p}(b_1) = b_1(1+\epsilon_1)$. There exists ϵ_2 with $|\epsilon_2| \leqslant \epsilon_{\text{machine}}$ such that $b_2' = \mathrm{f_p}(b_2) = b_2(1+\epsilon_2)$. Moreover, there exists ϵ_3 with $|\epsilon_3| \leqslant \epsilon_{\text{machine}}$ such that $b_3' = \mathrm{f_p}(b_3) = b_3(1+\epsilon_3)$.
Hence, we have:

$$\begin{aligned}(b_1' - b_2')(1+\epsilon_3) &= \big[b_1(1+\epsilon_1) - b_2(1+\epsilon_2)\big](1+\epsilon_3) = b_1(1+\epsilon_1)(1+\epsilon_3) - b_2(1+\epsilon_2)(1+\epsilon_3) \\ &= b_1 + \underbrace{b_1(\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3)}_{\delta b_1} - \big[b_2 + \underbrace{b_2(\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3)}_{\delta b_2}\big] = (b_1+\delta b_1) - (b_2+\delta b_2) \\ &= f(b_1+\delta b_1 + b_2 + \delta b_2) = f(\mathbf{b}+\delta\mathbf{b}).\end{aligned}$$

Just to note since for the $\mathcal{O}$, we are letting it $\to 0$, not infinity, so we have:

$$|\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3| \leqslant |\epsilon_1| + |\epsilon_3| + |\epsilon_1|\cdot|\epsilon_3| \leqslant 2\epsilon_{\text{machine}} + \epsilon_{\text{machine}}^2 = \mathcal{O}(\epsilon_{\text{machine}}).$$

Hence, we have the algorithm backwards stable.

Here, we consider such operation in norm notations as well:

$$\delta\mathbf{b} = \begin{bmatrix} \delta b_1 \\ \delta b_2 \end{bmatrix} = b_1\epsilon_4 + b_2\epsilon_5.$$

Hence the square of the norm is:

$$\|\delta\mathbf{b}\|^2 = |b_1|^2|\epsilon_4|^2 + |b_2|^2|\epsilon_5|^2 = |b_1|^2\mathcal{O}(\epsilon_{\text{machine}}^2) + |b_2|^2\mathcal{O}(\epsilon_{\text{machine}}^2) = \big(|b_1|^2+|b_2|^2\big)\mathcal{O}(\epsilon_{\text{machine}}^2),$$

which leads to that:

$$\frac{\|\delta\mathbf{b}\|^2}{\|\mathbf{b}\|^2} = \mathcal{O}(\epsilon_{\text{machine}}^2),$$

so the single power norm in $\mathcal{O}(\epsilon_{\text{machine}})$.

This makes us recall the **catastrophic cancellation**, that is when we have $b_1 \simeq b_2$, we have:

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|\tilde{f}(\mathbf{b}) - f(\mathbf{b})}{\|f(\mathbf{b})\|}$$

be large. However, the backwards stability means that there exists points in the neighborhood such that the differences can be similar. Here, we have b_1 and b_2 close enough to have some point in the neighborhood to be backwards stable.

Definition III.3.5. Stability.

Given a problem f, an algorithm $\tilde{f}$ is stable if for all $\mathbf{b} \in B$, there exists a perturbation of data $\delta\mathbf{b}$ with:

$$\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \mathcal{O}(\epsilon_{\text{machine}}),$$

such that:

$$\frac{\|\tilde{f}(\mathbf{b}) - f(\mathbf{b} + \delta\mathbf{b})}{\|f(\mathbf{b} + \delta\mathbf{b})\|} = \mathcal{O}(\epsilon_{\text{machine}}).$$

⌟

The stability is a weaker statement than backwards stability.

Proposition III.3.6. Backwards Stability $\Longrightarrow$ Stability.

A backwards stability algorithm is stable. The *converse* is not necessarily true.

This relaxation is necessary, as there are algorithms that are stable but not backward stable.

Example III.3.7. Stable but not Backward Stable Problem.

Consider the computation of the outer product between 2 vectors $\mathbf{a}, \mathbf{b} \in \mathbb{C}^m$m the mathematical problem is:

$$f : \mathbb{C}^m \times \mathbb{C}^m \to \mathbb{C}^{m \times m},$$
$$(\mathbf{a}, \mathbf{b}) \mapsto A = \mathbf{a}\mathbf{b}^*.$$

The algorithm is to:

```
input: a,b in C^m
a' := fp(a) =: (a1', a2', ..., am')
b' := fp(b) =: (b1', b2', ..., bm')
~A := [~Aij]
~Aij := ai' (x) comp_conj(bj') := fp(ai' * comp_conj(bj'))
```

Note that:

$$\tilde{f}(\mathbf{a}, \mathbf{b}) = \underbrace{\tilde{A}}_{\text{not rank 1 matrix}} \neq \underbrace{(\mathbf{a} + \delta\mathbf{a})(\mathbf{b} + \delta\mathbf{b})}_{\text{rank 1 matrix}} = f(\mathbf{a} + \delta\mathbf{a}, \mathbf{b} + \delta\mathbf{b}).$$

Hence it is not backwards stable. It can be shown that the algorithm is stable, and we leave this as an exercise to the readers. ⌟

It is worth-noting that our conclusion on stability is independent from the choice of vector norm.

Definition III.3.8. Equivalent Norms.

Let X be a normed vector space, any two vector norms $\|\bullet\|_\alpha$ and $\|\bullet\|_\beta$ in X are equivalent if there exist $C_1, C_2 > 0$ such that for all $\mathbf{x} \in X$, we have:

$$C_1\|\mathbf{x}\|_\beta \leqslant \|\mathbf{x}\|_\alpha \leqslant C_2\|\mathbf{x}\|_\beta.$$

⌟

Proposition III.3.9. All Finite Dimensional Norms are Equivalent.

Let X be a finite dimensional normed vector space, all norms are equivalent.

Example III.3.10. $\|\bullet\|_2$ **and** $\|\bullet\|_\infty$ **for are Equivalent.**

One can prove that:

$$\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_1 \leqslant \sqrt{m}\|\mathbf{x}\|.$$

⌟

There are some consequences of the equivalence norms.

Proposition III.3.11. Squeeze Theorem.

Suppose X is a finite dimensional normed vector space, and let $\|\bullet\|_\alpha$ and $\|\bullet\|_\beta$ be any norms in X. Let a sequence $\{x_n\}_{n=1}^\infty \subset X$ be such that $\|x_n\|_\alpha \to 0$ as $n \to \infty$, then $\|x_n\|_\beta \to 0$ as $n \to \infty$.

Proof. This is naturally by the above inequality:

$$0 \leqslant \|x_n\|_\beta \leqslant \frac{1}{C_1}\|x_n\|_\alpha \to 0,$$

hence, we naturally have $\|x\|_\beta \to 0$ as $n \to \infty$. □

Another consequence is on the backward stability:

Proposition III.3.12. Backward Stability.

Suppose we have $\mathbf{b} \in X$, which is a finite dimensional normed vector space, and $\|\bullet\|_\alpha, \|\bullet\|_\beta$ are two norms in X:

$$\frac{\|\delta\mathbf{b}\|_\beta}{\|\mathbf{b}\|_\beta} = \mathcal{O}(\epsilon_{\text{machine}}).$$

Then, we have:

$$\frac{\|\delta\mathbf{b}\|_\alpha}{\|\mathbf{b}\|_\alpha} = \mathcal{O}(\epsilon_{\text{machine}}).$$

Proof. From the definition, we have:

$$\|\delta\mathbf{b}\|_\alpha \leqslant C_2\|\delta\mathbf{b}\|_\beta \text{ and } \frac{1}{\|\mathbf{b}\|_\alpha} \leqslant \frac{1}{C_1}\cdot\frac{1}{\|\mathbf{b}\|_\beta}.$$

Therefore, we must have:

$$\frac{\|\delta\mathbf{b}\|_\alpha}{\|\mathbf{b}\|_\alpha} \leqslant \frac{C_2}{C_1}\cdot\frac{\|\delta\mathbf{b}\|_\beta}{\|\mathbf{b}\|_\beta} \leqslant \frac{C_2}{C_1}C\cdot\epsilon_{\text{machine}},$$

thus $\|\delta\mathbf{b}\|_\alpha/\|\mathbf{b}\|_\alpha = \mathcal{O}(\epsilon_{\text{machine}})$. □

Then, we can think of the conditional number for squared and invertible matrices, that is:

$$\kappa = \|A\|\cdot\|A^{-1}\|.$$

Here, think of the problem of solving $A.\mathbf{x} = \mathbf{b}$, the approximation solution is $\tilde{\mathbf{x}}$. Here, we have:

$$\begin{aligned} &\text{Error} && \delta\mathbf{x} = \mathbf{x} - \tilde{\mathbf{x}}, \\ &\text{Residual} && \delta\mathbf{b} = \mathbf{b} - A.\tilde{\mathbf{x}}. \end{aligned}$$

Remark III.3.13. General Condition Number.

For any mathematical problem and algorithm for $f : B \to X$ as $\tilde{f} : B \to X$, we can define the condition number as:

$$\kappa = \sup_{\delta\mathbf{b}\in B} \frac{\frac{\|f(\mathbf{b})-\tilde{f}(\mathbf{b})\|}{\|f(\mathbf{b})\|}}{\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}}.$$

I.e., K is the smallest number that is larger than all possible ratios. ⌟

Hence:

$$\frac{\|f(\mathbf{b})-\tilde{f}(\mathbf{b})\|}{\|f(\mathbf{b})\|} \leqslant \kappa\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \text{ for all } \delta\mathbf{b}.$$

Therefore, we note that:

$$\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \text{ is small when } \tilde{f} \text{ is backward stable or stable,}$$

hence since κ could be large, stability causes the left-hand-side to be large.

Again, for the same example, we could have a backwards stable, but the relative error is large due to catastrophic cancellation.

Recall that for solving the problem $A.\mathbf{x} = \mathbf{b}$ for a $A \in \mathbb{C}^{m\times n}$, we have the QR decomposition based method, that is:

(i) Let $A = QR$, which is the QR decomposition of A, where Q is orthogonal and R is upper triangular. (Cost: $\sim \frac{2}{3}mn^2 - \frac{2}{3}n^3$).

(ii) Compute $Q^{-1}.\mathbf{b} = Q^{*}.\mathbf{b}$. (Cost: $\sim 2mn$).

(iii) Solve $R.\mathbf{x} = Q^{*}.\mathbf{b}$ with the upper triangular system. (Cost: $\sim n^2$).

Remark III.3.14. QR Decomposition Based Method is Stable.

The QR decomposition method, *i.e.*, the above three steps are backward stable. ⌟

Theorem III.3.15. Backward Stability of QR Decomposition.

Let $A = QR$ be the QR decomposition of A, and let $\tilde{Q}\tilde{R}$ be the QR decomposition computed by Householder triangularization. This algorithm is backward stable in the same sense that the computed solution $\tilde{\mathbf{x}}$ has the property:

$$\|(A+\delta A)\tilde{\mathbf{x}} - \mathbf{b}\| = \min, \qquad \frac{\|\delta A\|}{\|A\|} = \mathcal{O}(\epsilon_{\text{machine}})$$

for some $\delta A \in \mathbb{C}^{m\times n}$.

III.4 Stability and Gaussian Elimination

Then, we consider the Gauss Elimination.

Remark III.4.1. LU Decomposition.

We may use Gauss elimination to form the LU decomposition, that is for $A \in \mathbb{C}^{m\times m}$, we have $A = L \circ U$, where $L = \begin{bmatrix} * & 0 & \cdots & 0 \\ * & * & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ * & * & \cdots & * \end{bmatrix}$ and $U = \begin{bmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & * \end{bmatrix}$. ⌟

Here, the algorithm lies as follows to solve $A.\mathbf{x} = \mathbf{b}$.

(i) Compute $A = LU$. (Cost: $\sim \frac{2}{3}m^3$).

(ii) Solve $L.\mathbf{y} = \mathbf{b}$ for the lower triangular system for $\mathbf{y}$. (Cost: $\sim m^2$).

(iii) Solve $U.\mathbf{x} = \mathbf{y}$ for the upper triangular system for $\mathbf{x}$. (Cost: $\sim m^2$).

Note that $L.\mathbf{y} = \mathbf{b}$ implies that $L(U.\mathbf{x}) = \mathbf{b}$, so $A.\mathbf{x} = \mathbf{b}$ so $\mathbf{x}$ solves $A.\mathbf{x} = \mathbf{b}$.

Example III.4.2. UL is Simpler than QR.

The price is it might be instable (could be corrected). Here, we consider:

$$\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} . \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Here, we can consider the solutions trivially as:

$$y_1 = \frac{b_1}{a_{11}},$$
$$y_2 = \frac{b_2 = a_{21}y_1}{a_{22}},$$
$$y_3 = \frac{b_3 - a_{31}y_1 - a_{32}y_2}{a_{33}}.$$

⌟

In particular, we consider $L_{m-1}\cdots L_2L_1A = U$, where L_i is the i-th set of elementary row operation, hence for $A = LU$, we have:

$$L = (L_{m-1}\cdots L_2L_1)^{-1} = L_1^{-1}L_2^{-1}\cdots L_{m-1}^{-1}.$$

Example III.4.3. Generic 4-by-4 Matrix with Gaussian Elimination.

Consider the steps as follows:

$$\underset{A}{\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}} \xrightarrow{L_1} \underset{L_1A}{\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix}} \xrightarrow{L_2} \underset{L_2L_1A}{\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix}} \xrightarrow{L_3} \underset{L_3L_2L_1U}{\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix}}.$$

⌟

More trivially for the 2-by-2 case, a matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is simply by $R_2 \leftarrow R_2 + kR_1$, that is $\begin{bmatrix} a & b \\ c+ka & d+kb \end{bmatrix}$, where $k = -c/a$ when $a \neq 0$ to achieve upper triangularization.

Then, we consider:

$$L_1 = \begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix},$$

whose inverse is then:

$$L_1^{-1} = \begin{bmatrix} 1 & 0 \\ -k & 1 \end{bmatrix},$$

which is also upper triangular.

For the 3-by-3 case, let:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ h & i & j \end{bmatrix},$$

we do the row operation to obtain that:

$$L_1 A = \begin{bmatrix} a & b & c \\ d + k_1 a & e + k_1 b & f + k_1 c \\ g + k_2 a & h + k_2 b & i + k_2 c \end{bmatrix},$$

where $k_1 = -d/a$ and $k_2 = -g/a$. Hence, we write:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ k_1 & 1 & 0 \\ k_2 & 0 & 1 \end{bmatrix},$$

where the inverse is:

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -k_1 & 1 & 0 \\ -k_2 & 0 & 1 \end{bmatrix}.$$

Here, we let the second table to be:

$$L_1 A = \begin{bmatrix} a & b & c \\ 0 & \alpha & \beta \\ 0 & \gamma & \delta \end{bmatrix}.$$

Here, we use the similar process for row operation. Here we have:

$$L_2(L_1 A) = \begin{bmatrix} a & b & c \\ 0 & \alpha & \beta \\ 0 & \gamma + k_3 \alpha & \delta + k_3 \beta \end{bmatrix},$$

where $k_3 = -\gamma + \alpha$, and we have:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & k_3 & 1 \end{bmatrix},$$

where the inverse is:

$$L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -k_3 & 1 \end{bmatrix}.$$

Here, the inverses are:

$$L = L_1^{-1}L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -k_1 & 1 & 0 \\ -k_2 & 0 & 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -k_3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -k_1 & 1 & 0 \\ -k_2 & -k_3 & 1 \end{bmatrix}.$$

There, we may observe that $A = LU = L_1^{-1}L_2^{-2}\cdots L_{m-1}^{-1}$ gives the Gauss Elimination process.

In the generic case, we consider the matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}.$$

Consider the first transformation, we have:

$$\ell_{i1} = \frac{a_{i1}}{a_{11}} \text{ for all } i = 2, 3, \cdots, m.$$

It is noteworthy to mention that if a_{11} is zero, we want to shuffle the rows. In fact, we would like to rearrange in a manner that has the largest coefficient at the top. But anyways, we have:

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{m1} & 0 & \cdots & 1 \end{bmatrix}.$$

For the second step, we assume that we have:

$$L_1A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ 0 & ax22 & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_{m2} & \cdots & x_{mm} \end{bmatrix}.$$

Hence, we consider that:

$$\ell_{i2} = \frac{x_{i2}}{x_{22}}.$$

$$L_2^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ell_{m2} & \cdots & 1 \end{bmatrix}.$$

Remark III.4.4. Decomposition exactly as the product of the entries.

We have the matrix entry exactly as:

$$L_{m-1}\cdots L_2L_1 = \mathrm{Id}_m + \sum_{i=1}^{m}(L_i - \mathrm{Id}_m).$$

⌟

Now, we consider the Pseudo code for the code segment as:

```
input: A in C(m*m)
output: L in C(m*m), U in C(m*m)
for j = 1, 2, ..., m-1:
```

```
    for j = 1, 2, ..., m-1:
        l[i,j] = n[i,j] / u[j,j]
        u[i,j:m] = u[i,j:m] - l[i,j] * u[j,j:m] % (*) Computationally intense
    end
end
```

Consider the computational complexity, we have the `(*)` intensive, that is:

- It multiplies a scalar by a vector of length ℓ.
- It subtracts 2 vectors of length ℓ.

Hence, there are 2ℓ flops.
At the jth step, we have $\ell = m - j + 1$.

Consider that there are less operations needed overtime, we consider the number of operations as:

$$\sum_{j=1}^{m-1} 2\ell(m-j) = 2\sum_{j=1}^{m-1}(m-j)^2 = 2\sum_{j=1}^{m-1} j^2 = 2 \cdot \frac{(m-1)m\big(2(m-1)+1\big)}{6} \sim \frac{2}{3}m^3.$$

Remark III.4.5. Problem with LU Decomposition.

The LU Decomposition incurs the following issues:

- We may have division by zero, such as $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$, where $\ell_{21} = \frac{1}{0}$.
- Stability issues: For $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$. It is invertible, and we compute $\kappa(A)$ with respect to the 2-norm, that is σ_2/σ_1. Consider a diagonalizable matrix, we have $\sigma_1 = |\lambda_1|$ and $\sigma_2 = |\lambda_2|$.
  For this case, we have the eigenvalues are $\frac{1\pm\sqrt{5}}{2}$, so we have $\sigma_1 = \frac{1+\sqrt{5}}{2}$ and $\sigma_2 = \frac{\sqrt{5}-1}{2}$, we have:
  $$\frac{\sigma_1}{\sigma_2} \simeq 2.618.$$

We consider the following example, then:

Example III.4.6. Unstable for Gauss Elimination.

Let's define $\delta A = \begin{bmatrix} 10^{-20} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \epsilon & 0 \\ 0 & 0 \end{bmatrix}$ and here we apply B as:

$$B = A + \delta A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix},$$

and so by the Gauss Elimination, we have:

$$L = \begin{bmatrix} 1 & 0 \\ \ell_{21} & 1 \end{bmatrix} \text{ where } \ell_{21} = \frac{1}{10^{-20}} = 10^{20}.$$

Thus, we have:

$$L = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 10^{-20} & 1 \\ 0 & 1 - 10^{20} \end{bmatrix}.$$

Here, we assume that $\mathrm{f_p}(10^20) = 10^{20}$ and $\mathrm{f_p}(10^{-20}) = 10^{-20}$, then we have machine representation as:

$$\tilde{L} = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix} \text{ and } \tilde{U} = \begin{bmatrix} 10^{-20} & 1 \\ 0 & -10^{20} \end{bmatrix}.$$

Thus, we have the matrix multiplication as:

$$\tilde{L}\tilde{U} = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix} \begin{bmatrix} 10^{-20} & 1 \\ 0 & -10^{20} \end{bmatrix} = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 0 \end{bmatrix}.$$

We may observe that this is very far from the initial A, since we had a large deviation with $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$.

Note that this is *not* a Catastrophic cancellation, but rather a *roundoff error*, hence it is unstable. ⌟

The solution to the problem is *Partial Pivoting*.

Remark III.4.7. Partial Pivoting.

For the above example, we may exchange the rows in B, that is:

$$B = \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix} \longrightarrow C = \begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix}.$$

Thus, we have $R_2 \leftarrow R_2 - \epsilon R_1$ with $L_1 = \begin{bmatrix} 1 & 0 \\ -\epsilon & 1 \end{bmatrix}$ and thus $L = L_1^{-1} = \begin{bmatrix} 1 & 0 \\ \epsilon 1 & \end{bmatrix}$, so we have $L_1 C = \begin{bmatrix} 1 & 1 \\ 0 & 1-\epsilon \end{bmatrix}$, which is upper triangular.

Again, if we assume $\epsilon = 10^{-20}$, we have the floating point approximation as:

$$\tilde{L} = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix}, \text{ and so } \tilde{U} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \text{ and thus } \tilde{L}\tilde{U} = \begin{bmatrix} 1 & 1+\epsilon \\ \epsilon & 1 \end{bmatrix} \simeq \begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix} = C.$$

Note that the above example is a little coincident, since the error almost cancels out. ⌟

More in general, suppose we start from $A \in \mathbb{C}^{m\times m}$, at jth step of Gauss Elimination, we have:

$$\begin{array}{c} \overbrace{\qquad\qquad\qquad}^{j-1}\ \overbrace{\qquad}^{m-j+1} \\ \begin{bmatrix} * & * & \cdots & * & \\ 0 & * & \cdots & * & \\ \vdots & \vdots & \ddots & \vdots & * \\ 0 & 0 & \cdots & * & \\ & & 0 & & x_{jk} \end{bmatrix} \end{array}$$

Here, we choose the x_{jk} parts as pivot, which is the largest entry in absolute value.

- We switch corresponding row with jth row.
- Then perform Gauss Elimination, which results in:
  - keep track of all the row switchings.
  - with partial pivoting, stability improves.

Aside, in complete pivoting, at jth step, you choose the largest entry (in absolute value) in entire block.

- The pro is that we have an improvement in stability, although it is too small to justify computational cost.
- The cons is we have to compute the values of $(m-j+1)^2$ numbers, and we keep track of all raw switches and column switches.

Typically, partial pivoting is good enough.

Definition III.4.8. Transposition.

Suppose we want to switch the ith and jth row of a matrix, the matrix can be acted by a *transposition*, denoted σ_{ij}.

Example III.4.9. Switching 2nd and 3rd Row.

Suppose we are switching 2nd and 3rd rows of a 3-by-3 matrix, we have:

$$\sigma_{2\,3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Proposition III.4.10. Transposition as Matrix.

σ_{ij} can be represented as the matrix:

$$\sigma_{ij} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix},$$

where the ith column & rows and the jth column & rows.

Remark III.4.11. Postcompose with Transposition.

For any matrix A, $\sigma_{ij}A$ switches on ith and jth row, and $A\sigma_{ij}$ switches on ith and jth column. This is also called *simple permutation*.

Proposition III.4.12. Transposition has Order 2.

$\sigma_{ij}^2 = \mathrm{Id}$, or equivalently $\sigma_{ij}^{-1} = \sigma_{ij}$. We call such action *involutary*.

Technically, in one step, we can perform a generic permutation, that is permuting different rows. Here, we may denote it as:

$$(r_1\ r_2\ \cdots\ r_n).$$

Proposition III.4.13. Permutations can be Composed of Transposition.

Each permutation can be achieved by a sequence of tranpositions.

Example III.4.14. Decomposing Permutation.

Here, we consider the permutation (1 2 3), we may write it as $(2\ 3) \circ (1\ 2)$. ⌟

Proposition III.4.15. Permutations are Orthogonal.

For any permutation of rows of the corresponding matrix σ is an orthogonal matrix, *i.e.*:

$$\sigma^{-1} = \sigma^{\mathsf{T}}, \text{ or } \sigma^{\mathsf{T}}\sigma = \mathrm{Id}\,.$$

Proof. Since the columns of σ are the shuffled canonical basis, hence it is a *orthonormal basis*, and thus σ is orthonormal. □

Example III.4.16. Gauss Elimination with Pivoting.

Consider $B = \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix}$, we want to have the 1 as the leading entry, so we have:

$$\sigma_{1\ 2} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ so that } \sigma_{1\ 2}B = \begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix}.$$

Hence, we then have $L_1 = \begin{bmatrix} 1 & 0 \\ -\epsilon & 1 \end{bmatrix}$, so that:

$$L_1\sigma_{1\ 2}B = \begin{bmatrix} 1 & 1 \\ 0 & 1-\epsilon \end{bmatrix} = U,$$

which leads to $\sigma_{1\ 2}B = LU$, where $L = L_1^{-1} = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix}$. ⌟

More in general, for a m-by-m matrix A, we have:

$$L_{m-1}\sigma_{m-1}\cdots L_2\sigma_2 L_1\sigma_1 A = U.$$

Recall that matrix multiplications do not necessarily commute, so the goal is to manipulate the above equation to get $PA = LU$.

Example III.4.17. Pivoting for 3-by-3 Matrix.

Consider $A = \begin{bmatrix} -2 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -8 & 4 \end{bmatrix}$, so we want the second row to be the first row, hence we have:

$$\sigma_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which gives that:

$$\sigma_1 A = \begin{bmatrix} 6 & -6 & 7 \\ -2 & 2 & -1 \\ 3 & -8 & 4 \end{bmatrix}$$

Then, the pivot is $\ell_{21} = -2/6 = -1/3$ and $\ell_{31} = 3/6 = 1/2$, so we have:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\ell_{21} & 1 & 0 \\ -\ell_{31} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix}.$$

Then, by computation, we have:

$$L_1\sigma_1 A = \begin{bmatrix} 6 & -6 & 7 \\ 0 & 0 & 4/3 \\ 0 & -5 & 1/2 \end{bmatrix}$$

Then, σ_2 is switching the second and third row, then we have:

$$\sigma_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Now, we have:

$$\sigma_2 L_1 \sigma_1 A = \begin{bmatrix} 6 & -6 & 7 \\ 0 & -5 & 1/2 \\ 0 & 0 & 4/3 \end{bmatrix} = U.$$

Note that here we have $L_2 = \mathrm{Id}$, where we are lucky. ⌟

Here, we have:

$$\sigma_2 L_1 \sigma_1 A = U.$$

Note that σ_2 and L_1 does not commute, so we want to define some $L_1'\sigma_2 L_1 \sigma_2$, then we have:

$$L_1'\sigma_2\sigma_1 A = \sigma_2 L_1 \sigma_2 \sigma_2 \sigma_1 A = \sigma_2 L_1 \sigma_1 A = U.$$

Then $L_1'\sigma_2\sigma_1 A = U$, and we have $\sigma_2\sigma_1 A = (L_1')^{-1}U$.
Here, we consider $L_1 \mapsto L_1' = \sigma_2 L_1 \sigma_2$, which is:

$$L_1' = \underbrace{\sigma_2}_{\text{permuting rows 2 and 3}} L_1 \underbrace{\sigma_2}_{\text{permuting columns 2 and 3}} = \begin{bmatrix} 1 & 0 & 0 \\ -\ell_{31} & 1 & 0 \\ -\ell_{21} & 0 & 1 \end{bmatrix}.$$

Here, we have the permutation of the columns cleaning up the permutation of the rows.
Then, we have:

$$(L_1')^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{31} & 1 & 0 \\ \ell_{21} & 0 & 1 \end{bmatrix}.$$

Thus, we will be $L_1'\sigma_2\sigma_1 A = U$ and $PA = LU$, where $\sigma = \sigma_2\sigma_1$ for permutation, $L = (L_1')^{-1}$ for lower triangular matrix, and U and the upper triangular matrix.

In the general m-by-m case, we can define:

$$\begin{aligned}
L_{m-1}' &= L_{m-1}, \\
L_{m-2}' &= \sigma_{m-1} L_{m-2} \sigma_{m-1}, \\
L_{m-3}' &= \sigma_{m-1}\sigma_{m-2} L_{m-3} \sigma_{m-2}\sigma_{m-1}, \\
&\vdots \\
L_2' &= (\sigma_{m-1}\cdots\sigma_4\sigma_3) L_2 (\sigma_3\sigma_4\cdots\sigma_{m-1}), \\
L_1' &= (\sigma_{m-1}\cdots\sigma_3\sigma_2) L_1 (\sigma_2\sigma_3\cdots\sigma_{m-1}).
\end{aligned}$$

Now, we just compute that:

$$L_{m-1}'L_{m-2}'\cdots L_2'L_1'\sigma_{m-1}\sigma_{m-2}\cdots\sigma_2\sigma_1 A = L_{m-1}\sigma_{m-1}L_{m-2}\sigma_{m-2}L_{m-3}\cdots\sigma_3 L_2\sigma_2 L_1\sigma_1 A.$$

Remark III.4.18. Conjugation Action of Permutation is Exchanging ℓ_i and ℓ_j.

Consider matrix L, then the action $\sigma_i \circlearrowleft L$, it is simply exchanging the ℓ_i and ℓ_j entry and all the other are the same. ⌟

Hence, we have the computational complexity being $\sim \frac{2}{3}m^3$.

III.5 Hermitian Matrices and Quadratic Forms

Definition III.5.1. Hermitian Matrix.

A matrix $A \in \mathbb{C}^{m\times m}$ is Hermitian if $A^* = A$. ⌟

Here, we have $A = \left[a_{ij}\right]$ such that $\overline{a_{ji}} = a_{ij}$.
In particular, since for $i = j$, then $\overline{a_{ii}} = a_{ii}$, so the diagonal entries are real.

Proposition III.5.2. Hermitian Matrix has Real Eigenvalues.

If A^*A, then A has real eigenvalues.

Proof. Suppose that λ is an eigenvalue of A, *i.e.*, there exists $\mathbf{x} \neq \mathbf{0}$ such that $A.\mathbf{x} = \lambda\mathbf{x}$.
Then $\overline{\lambda}\|\mathbf{x}\|_2^2 = \overline{\lambda}\mathbf{x}^*\mathbf{x} = (\lambda\mathbf{x})^*\mathbf{x} = (A.\mathbf{x})^*\mathbf{x} = \mathbf{x}^*A^*\mathbf{x} = \mathbf{x}^*A\mathbf{x} = \mathbf{x}^*(\lambda\mathbf{x}) = \lambda\mathbf{x}^*\mathbf{x} = \lambda\|\mathbf{x}\|_2^2$.
Hence, we have $\overline{\lambda}\|\mathbf{x}\|^2 = \lambda\|\mathbf{x}\|^2$, so $\overline{\lambda} = \lambda$ since $\|\mathbf{x}\| \neq 0$. Thus λ is real. □

The eigenvectors for eigenvalue λ are the nonzero solutions $\mathbf{x}$ to $A.\mathbf{x} = \lambda\mathbf{x}$, which is equivalent to $A.\mathbf{x} - \lambda\mathbf{x} = \mathbf{0}$, which is equivalently $A.\mathbf{x} - \lambda\,\mathrm{Id}\,.\mathbf{x} = \mathbf{0}$, that is $(A - \lambda\,\mathrm{Id}).\mathbf{x} = \mathbf{0}$.
Hence, the eigenspace associated to λ is:

$$E_\lambda = \ker A - \lambda\,\mathrm{Id} = \{\mathbf{x} : (A - \lambda\,\mathrm{Id})\mathbf{x} = \mathbf{0}\} \subset \mathbb{C}.$$

Theorem III.5.3. Distinct Eigenspaces are Orthogonal.

If $A^* = A$ and λ and μ are distinct eigenvalues of A, then $E_\mu \perp E_\lambda$, *i.e.*, for all $\mathbf{u} \in E_\mu$ and $\mathbf{v} \in E_\lambda$, we have $\mathbf{v}^*\mathbf{u} = 0$.

Proof. Consider $\mathbf{u}^*A\mathbf{v}$ being a scalar $z \in \mathbb{C}$, so $z^* = \overline{z}$, then:

$$(\mathbf{u}^*A.\mathbf{v})^* = \mathbf{u}^*A\mathbf{v}.$$

Then, we have:

$$\begin{aligned}(\mathbf{u}^*A\mathbf{v})^* &= \mathbf{v}^*A^*\mathbf{u} = \mathbf{v}^*A\mathbf{u} = \mathbf{v}^*(\mu\mathbf{u}) \\ &= \overline{\mathbf{u}^*A\mathbf{v}} = \overline{\mathbf{u}^*(\lambda\mathbf{v})} = \overline{\lambda\mathbf{u}^*\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{u}^*\mathbf{v}} = \lambda\mathbf{v}^*\mathbf{u}.\end{aligned}$$

Hence, $\lambda\mathbf{v}^*\mathbf{u} = \mu\mathbf{v}^*\mathbf{u}$, then $(\lambda - \mu)\mathbf{v}^*\mathbf{u} = 0$, thus for $\lambda \neq \mu$, we have $\mathbf{v}^*\mathbf{u} = 0$. □

Theorem III.5.4. Spectral Theorem.

If $A^* = A$, then A is unitary diagonalizable, *i.e.*, there exists a unitary matrix $Q \in \mathbb{C}^{m\times n}$ $(Q^{-1} = Q^*)$ such that:

$$Q^*AQ = D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix},$$

where $Q = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots \mathbf{v_m} \end{bmatrix}$, where $\beta = \{\mathbf{v_1}, \cdots, \mathbf{v_m}\}$ is an *orthonormal* basis of $\mathbb{C}^m$ made of eigenvectors of A.

Definition III.5.5. Bilinear Form.

A bilinear form is a function $b : \mathbb{C}^m \times \mathbb{C}^m \to \mathbb{C}$ of the type:

$$b(\mathbf{x}, \mathbf{y}) = \mathbf{x}^* A\mathbf{y} \text{ where } A \in \mathbb{C}^{m\times n},$$

which has the following properties:

- $b(\mathbf{x_1} + \mathbf{x_2}, \mathbf{y}) = b(\mathbf{x_1}, \mathbf{y}) + b(\mathbf{x_2}, \mathbf{y})$.
- $b(\mathbf{x}, \mathbf{y_1} + \mathbf{y_2}) = b(\mathbf{x}, \mathbf{y_1}) + b(\mathbf{x}, \mathbf{y_2})$.
- $b(k\mathbf{x}, \mathbf{y}) = \overline{k}b(\mathbf{x}, y)$, and
- $b(\mathbf{x}, k\mathbf{y}) = kb(\mathbf{x}, \mathbf{y})$.

⌟

Definition III.5.6. Quadratic Form.

A quadratic form is a function $q : \mathbb{C}^m \to \mathbb{C}^m$ of the type:

$$q(\mathbf{x}) = \mathbf{x}^* A\mathbf{x} \text{ where } A \text{ is a matrix of the quadratic form.}$$

Typically, we assume $A^* = A$. ⌟

Example III.5.7. Computation of Quadratic Form.

Let $A = \begin{bmatrix} a & \overline{b} \\ b & c \end{bmatrix}$, where $a, c \in \mathbb{R}$, then we have:

$$\begin{aligned} q(\mathbf{x}) &= \begin{bmatrix} \overline{x_1} & \overline{x_2} \end{bmatrix} \begin{bmatrix} a & \overline{b} \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \overline{x_1} & \overline{x_2} \end{bmatrix} \begin{bmatrix} ax_1 + \overline{b}x_2 \\ bx_1 + cx_2 \end{bmatrix} \\ &= a\overline{x_1}x_1 + \overline{b}\overline{x_1}x_2 + bx_1\overline{x_2} + c\overline{x_2}x_2 = a|x_1|^2 + 2\Re(\overline{b}\overline{x_1}x_2) + c|x_2|^2 \in \mathbb{R}. \end{aligned}$$

⌟

In general, when A is Hermitian, then:

$$\overline{q(\mathbf{x})} = \overline{\mathbf{x}^* A\mathbf{x}} = (\mathbf{x}^* A\mathbf{x})^* = \mathbf{x}^* A^* xx = \mathbf{x}^* A\mathbf{x} = q(\mathbf{x}).$$

Hence $q(\mathbf{x}) \in \mathbb{R}$ when $A^* = A$.

Definition III.5.8. Skew-Hermitian.

A matrix $B \in \mathbb{C}^{m\times m}$ is *skew*-Hermitian if $B^* = -B$. ⌟

Proposition III.5.9. Decomposition of Matrix to Hermitians.

For any $A \in \mathbb{C}^{m\times m}$ may be decomposed as:

$$A = A_H + A_S,$$

where A_H is hermitian and A_S is skewed-hermitian.

Proof. We construct that:

$$\begin{aligned} A_H &= \frac{1}{2}(A + A^*), \\ A_S &= \frac{1}{2}(A - A^*). \end{aligned}$$

□

Suppose B is skew-symmetric, then:

$$-\mathbf{x}^* B\mathbf{x} = \mathbf{x}^* B^* \mathbf{x} = (\mathbf{x}^* B\mathbf{x})^* = \overline{\mathbf{x}^* B\mathbf{x}},$$

this we know that $\mathbf{x}^* B\mathbf{x} \in \mathrm{i}\mathbb{R}$, *i.e.*, purely imaginary.

Here, for a generic matrix A, we have:

$$q(\mathbf{x}) = \mathbf{x}^* A\mathbf{x} = \mathbf{x}^*(A_H + A_S)\mathbf{x} = \underbrace{\mathbf{x}^* A_H \mathbf{x}}_{\in\mathbb{R}} + \underbrace{\mathbf{x}^* A_S \mathbf{x}}_{\in \mathrm{i}\mathbb{R}}.$$

Definition III.5.10. Positive Definite.

Given a Hermitian matrix $A \in \mathbb{C}^{m\times m}$, we call it *positive definite* if $\mathbf{x}^* A\mathbf{x} > 0$ for all $\mathbf{x} \in (\mathbb{C}^m)^\times$, *i.e.*, $\mathbf{x} \neq \mathbf{0}$. In particular, we call the matrix as *Hermitian Positive Definite* (HPD). ⌟

Proposition III.5.11. Characterization of Positive Definiteness.

Suppose that A is hermitian, then:

(i) Eigenvalues of A are all real.

(ii) A is unitary diagonalizable, *i.e.*, we can find an orthonormal basis:

$$\beta = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}\}$$

Proof of (ii). If we consider:

$$Q = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_m} \end{bmatrix}$$

It is unitary ($Q^{-1} = Q^*$) and:

$$Q^* A Q = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}.$$

Here, we rewrite it as $A = Q\Lambda Q^*$, and we recall that for any $\mathbf{b} \in \mathbb{C}^m$, we have $Q^*.\mathbf{b}$ as the vector of coordinates of $\mathbf{b}$ with respect to β, namely $[\mathbf{b}]_\beta = (c_1, \cdots, c_m)$ with property that:

$$c_1\mathbf{v_1} + \cdots + c_m\mathbf{v_m} = \mathbf{b}.$$

Hence, we have $Q.[\mathbf{b}]_\beta = \mathbf{b}$, and thus $[\mathbf{b}]_\beta = Q^{-1}.\mathbf{b} = Q^*.\mathbf{b}$.
Then, we can rewrite the quadratic form $q(\mathbf{x}) = \mathbf{x}^*A\mathbf{x}$ in terms of the coordinates of $\mathbf{x}$ with respect to β for all $\mathbf{x} \in \mathbb{C}^m$, so we denote $[\mathbf{x}]_\beta = Q^*\mathbf{x}$.
Then, we consider $A = Q\Lambda Q^*$, so we have the quadratic form as:

$$q(\mathbf{x}) = \mathbf{x}^*A\mathbf{x} = \mathbf{x}^*(Q\Lambda Q^*)\mathbf{x} = (\mathbf{x}^*Q)\Lambda(Q^*\mathbf{x}) = (Q^*\mathbf{x})^*\Lambda(Q^*\mathbf{x}) = [\mathbf{x}]_\beta^*\Lambda[\mathbf{x}]_\beta.$$

Then, the above equation becomes:

$$\begin{aligned} q(\mathbf{x}) &= \begin{bmatrix} \overline{c_1} & \overline{c_2} & \cdots & \overline{c_m} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} \\ &= \lambda_1\overline{c_1}c_1 + \lambda_2\overline{c_2}c_2 + \cdots + \lambda_m\overline{c_m}c_m = \lambda_1|c_1|^2 + \lambda_2|c_2|^2 + \cdots + \lambda_m|c_m|^2. \end{aligned}$$

Since the eigenvalues are real, where $[\mathbf{x}]_\beta = (c_1, \cdots, c_m)$. □

Theorem III.5.12. Positive Definiteness $\iff$ Positive Eigenvalues.

Let $A \in \mathbb{C}^{m\times m}$ be hermitian. A is positive definite if and only if all of its eigenvalues are strictly positive.

Proof. ($\Longrightarrow$:) We choose $\mathbf{x} = \mathbf{v_1}$, which is the first vector of β, we have:

$$\mathbf{x} = \mathbf{v_1} = 1\mathbf{v_1} + 0\mathbf{v_2} + \cdots + 0\mathbf{v_m},$$

so we have:

$$[\mathbf{x}]_beta = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

which implies that $q(\mathbf{x}) = \lambda_1 \cdot |1|^2$. Hence $\lambda_1 > 0$. Since this applies for any canonical vector in $\mathbb{C}^m$, we can choose $\mathbf{x} = \mathbf{v_i}$, and we similarly obtain that $\lambda_i > 0$.
($\Longleftarrow$:) Let $\mathbf{x} \in (\mathbb{C}^m)^\times$ be arbitrary. We consider $[\mathbf{x}]_\beta = (c_1, \cdots, c_m)$, at least one of the c_i's is nonzero, otherwise $\mathbf{x} = \mathbf{0}$. So at least one of terms in $q(\mathbf{x})$ is strictly positive, and since all terms are nonnegative, then $q(\mathbf{x}) > 0$. □

Example III.5.13. Using Quadratic Form to Find Positivity.

Show that $5x^2 - 4xy + 2y^2 > 0$ for all $x, y \in \mathbb{R}$ and $(x, y) \neq \mathbf{0}$. Thus, we have:

$$q(\mathbf{x}) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 5 & -2 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

We note that $A = A^\intercal$ and it is real, si $A = A^*$ is hermitian. Thus:

$$\begin{cases} \lambda_1\lambda_2 = \det A = 10 - 4 = 6, \\ \lambda_1 + \lambda_2 = \operatorname{Tr} A = 5 + 2 = 7. \end{cases}$$

Thus, we have $\lambda_1 = 6$ and $\lambda_2 = 1$, so $q(\mathbf{x}) > 0$ for all $(x, y) \neq 0$. ⌟

Remark III.5.14. Positive Eigenvalues for Generic Matrices.

For a generic matrix A with strictly positive eigenvalues, we know that:

$$\operatorname{Tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_m > 0.$$

Therefore $a_{1\ 1} + a_{2\ 2} + \cdots + a_{m\ m} > 0$. ⌟

However, if we add hermitian assumption, then we have the following theorem.

Theorem III.5.15. HPD $\Longrightarrow$ Positive Diagonal.

If A is Hermitian Positive Definite, then $a_{ii} > 0$, *i.e.*, the diagonal entries are strictly positive.

Proof. Suppose $\mathbf{x}^* A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, we choose $\mathbf{x} = (1, 0, \cdots, 0)$, so we have:

$$\mathbf{x}^* A\mathbf{x} = a_{11} > 0 \text{ by HPD.}$$

Similarly, we can choose $\mathbf{x}$ as any canonical vector, so $a_{22} > 0$. □

Example III.5.16. Hermitian is Necessary Condition.

We let $A = \begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}$ is HPD, so we have $a_{11} > 0$ and $a_{22} > 0$.

We need to have hermitian condition, otherwise let $C = \begin{bmatrix} -1 & 4 \\ -2 & 5 \end{bmatrix}$, which is not hermitian. Here, we have $\lambda_1\lambda_2 = 3$ and $\lambda_1 + \lambda_2 = 4$, that is $\lambda_1 = 3$ and $\lambda_2 = 1$, so it has positive eigenvalues, and the trace is positive, but the diagonal is not strictly positive. ⌟

Proposition III.5.17. Positive Definite $\Longleftrightarrow$ Positive Eigenvalues for Hermitian Matrix.

Let A be Hermitian, then A is positive definite if and only if all eigenvalues are strictly positive.

Theorem III.5.18. HPD $\Longrightarrow$ Full Rank Conjugation is HPD.

If A is hermitian positive definite, then for any matrix $X \in \mathbb{C}^{m\times n}$ with $m \geqslant n$ and full rank ($\dim(\operatorname{im} X) = n$), $X^* AX \in \mathbb{C}^{n\times n}$ is also hermitian positive definite.

Proof. To show it is hermitian, since $A^* = A$:

$$(X^* AX)^* = X^* A^* (X^*)^* = X^* AX$$

To show it is positive definite, since X has rank n, so its columns are linearly independent. Therefore, if $\mathbf{x} \neq 0$, we have $X.\mathbf{x} \neq \mathbf{0}$. Thus:

$$\mathbf{x}^* (X^* AX)\mathbf{x} = (X.\mathbf{x})^* A(X.\mathbf{x}) > 0,$$

so $X^* AX$ is HPD. □

A consequence of the theorem is that the *principal submatrices* of A are HPD, *i.e.*, the matrices from A by eliminating the same rows and columns with same indices.

Example III.5.19. Eliminating First Row and Column.

Suppose $A \in \mathbb{C}^{3\times 3}$ and we want to remove the first row and column. We use:

$$X = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

where we now have:

$$X^* A X = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}.$$

⌟

III.6 Cholesky Decomposition

Definition III.6.1. Cholesky Decomposition.

A Cholesky decomposition of a matrix $A \in \mathbb{C}^{m\times m}$ is a factorization $A = R^* R$, where R is upper triangular with positive entries on the main diagonal.
Particularly, we have:

$$A = \begin{bmatrix} * & 0 & \cdots & 0 \\ * & * & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \end{bmatrix} \begin{bmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & * \end{bmatrix}$$

⌟

Theorem III.6.2. HPD $\Longleftrightarrow$ Existence of Cholesky Decomposition.

Suppose $A \in \mathbb{C}^{m\times m}$, then A is hermitian positive definite if and only A has a Cholesky decomposition. Moreover, the decomposition has to be unique, and when A is real, so is R.

Proof. ($\Longleftarrow$:) Let R^*R be hermitian, then $(R^*R)^* = R^*(R^*)^* = R^*R$, then we have:

$$A = R^*R = R^* \operatorname{Id} R \text{ which is HPD.}$$

($\Longrightarrow$:) We denote A as:

$$A = \begin{bmatrix} a_{11} & \mathbf{w}^* \\ \mathbf{w} & K \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & & & & \\ \ell_{31} & & & & \\ \vdots & & \operatorname{Id}_{m-1} & & \\ \ell_{m1} & & & & \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{w}^* \\ 0 & \\ 0 & K_1 \\ \vdots & \\ 0 & \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{w}/a_{11} & \operatorname{Id}_{m-1} \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{w}^* \\ \mathbf{0} & K_1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{w}/a_{11} & \operatorname{Id}_{m-1} \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{0} \\ \mathbf{0} & K_1 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{w}^*/a_{11} \\ \mathbf{0} & \operatorname{Id}_{m-1} \end{bmatrix}.$$

Here, we can form the matrix as:

$$\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{w}/a_{11} & \operatorname{Id}_{m-1} \end{bmatrix} \begin{bmatrix} 1/\sqrt{a_{11}} & \mathbf{0} \\ \mathbf{0} & \operatorname{Id}_{m-1} \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{0} \\ \mathbf{0} & K_1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{a_{11}} & \mathbf{0} \\ \mathbf{0} & \operatorname{Id}_{m-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{w}^*/a_{11} \\ \mathbf{0} & \operatorname{Id}_{m-1} \end{bmatrix}.$$

Then, we can repeat the steps, as:

$$A = R_!^* R_2^* \begin{bmatrix} \mathrm{Id}_2 & 0 \\ 0 & K_2 \end{bmatrix} R_2 R_1.$$

And following that, we can have after m steps that:

$$A = \underbrace{R_1^* R_2^* \cdots R_m^*}_{R^*} \underbrace{R_1 R_2 \cdots R_m}_{R}.$$

□

Example III.6.3. Cholesky Decomposition of Matrix.

Let $A = \begin{bmatrix} 4 & 6 \\ 6 & 13 \end{bmatrix}$. Since it is real and symmetric, it is hermitian.
Here, we have:

$$\begin{cases} \lambda_1 \lambda_2 = \det A = 16, \\ \lambda_1 + \lambda_2 = \operatorname{Tr} A = 17. \end{cases}$$

Hence, we have $\lambda_1 = 16$ and $\lambda_2 = 1$, and they are both positive. Hence, we have:

$$A = R^* R = \begin{bmatrix} r_{11} & 0 \\ r_{12} & r_{22} \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 6 & 13 \end{bmatrix}.$$

Here, we have $r_{11} - 2$ and $r_{22} = 2$, with $R = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix}$. ⌟

Remark III.6.4. Computational Complexity for.

The computational complexity of Cholesky decomposition is $\sim \frac{1}{3} m^3$.
Recall that the computational complexity for LU decomposition is $\sum \frac{2}{3} m^3$. ⌟

Example III.6.5. Manual Computation of 3-by-3 Cholesky Decomposition.

Suppose we have a Hermitian matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \overline{a_{12}} & a_{22} & a_{23} \\ \overline{a_{13}} & \overline{a_{23}} & a_{33} \end{bmatrix}.$$

Note that we may ignore the lower part of the matrix as it has information contained in the upper half. This is due to A being *Hermitian*.
The we consider the decomposition as:

$$A = \begin{bmatrix} r_{11} & 0 & 0 \\ \overline{r_{12}} & r_{22} & 0 \\ \overline{r_{13}} & \overline{r_{23}} & r_{33} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = R^* R.$$

Here, we have 6 unknowns for the system of equation. However, the solution is unique up to the diagonal being positive real numbers.

- By $r_{11}^2 = a_{11}$, we have $r_{11} = \sqrt{a_{11}}$.
- By $r_{11} r_{12} = a_{12}$, so we have $r_{12} = a_{12}/r_{11}$.
- By $a_{22} = \overline{r_{12}} r_{12} + r_{22}^2 = |r_{12}|^2 + r_{22}^2$, so we have $r_{22} = \sqrt{a_{22} - |r_{12}|^2}$.
- By $a_{13} = r_{11} r_{13}$, we have $r_{13} = a_{13}/r_{11}$.

- By $a_{23} = \overline{r_{12}} r_{13} + r_{22} r_{23}$, so $r_{23} = (a_{23} - \overline{r_{12}} r_{13})/r_{22}$.
- By $a_{33} = \overline{r_{13}} r_{13} + \overline{r_{23}} r_{23} + r_{33}^2$, thus $r_{33} = \sqrt{a_{33} - |r_{13}|^2 - |r_{23}|^2}$.

Here, note that for r_{11}, r_{22}, and r_{33}, the solution is not unique as we could have the negative values. However, given that A is Hermitian, must have them being positive, so the solution is unique. ⌟

We develop the algorithm based on the proof:

$$A = (R_m^* \cdots R_2^* R_1^*) \,\mathrm{Id}(R_1 R_2 \cdots R_m).$$

- The complexity is $\sim \frac{1}{3} m^3$.
- This is backward stable, meaning that it produce matrix $\tilde{R}$ (upper triangular with positive $r_{ii} > 0$) such that:

  $$\delta A = A - \tilde{R}^* \tilde{R},$$

  and it has property that:

  $$\frac{\|\delta A\|}{\|A\|} = \mathcal{O}(\epsilon_{\text{machine}}).$$

Remark III.6.6. Ill Conditioned Matrix.

$\tilde{R}$ may have significant relative error if A is ill condition, *i.e.*, $\delta R = R - \tilde{R}$, we will have:

$$\frac{\|\delta R\|}{\|R\|} = \mathcal{O}\big(\kappa(A) \cdot \epsilon_{\text{machine}}\big).$$

⌟

Hence, the algorithm lies as follows:

(i) Compute the Cholesky: $A = R^* R$. $(\sim \frac{1}{3} m^3)$.

(ii) Solve $R^* \mathbf{y} = \mathbf{b}$ for $\mathbf{y}$ as lower triangular. $(\sim m^2)$.

(iii) Solve $R\mathbf{x} = \mathbf{b}$ for xx as the upper triangular. $(\sim m^2)$.

Hence the total complexity is $\sim \frac{1}{3} m^3$.

Example III.6.7. Application of Cholesky Decomposition.

The Cholesky decomposition can be applied to the following cases:

- Solving $A\mathbf{x} = \mathbf{b}$, when A is HPD, that is:

  $$A\mathbf{x} = \mathbf{b} \longrightarrow R^* \underbrace{R\mathbf{x}}_{\mathbf{y}} = \mathbf{b}.$$

- Solving the least square approximation. Consider $A \in \mathbb{C}^{m \times m}$, where $m \geqslant n$, the $\dim(\operatorname{im} A) = n$ (by Rank-Nullity $\ker A = \{0\}$). Here, we have the columns being linearly independent.
  We want to solve $A.\mathbf{x} = \mathbf{b}$. Here we discuss by cases:

  (i) For $\mathbf{b} \in \operatorname{im} A$, so $A.\mathbf{x} = \mathbf{b}$ has a solution and it is unique by linearly independence of columns.
  Recall that we showed $\ker A = \ker(A^* A)$, so $\ker(A^* A) = \{\mathbf{0}\}$ and so is invertible.
  Thus $A.\mathbf{x} = \mathbf{b}$ implies that $(A^* A)\mathbf{x} = A^* \mathbf{b}$, so $\mathbf{x} = (A^* A)^{-1} A^* \mathbf{b} = A^+ \mathbf{b}$. Thus $\mathbf{x} = A^+ \mathbf{b}$.

  (ii) Consider $\mathbf{b} \notin \operatorname{im} A$, we want to find $\mathbf{x} \in \mathbb{C}^n$ that minimizes $\|\mathbf{b} - A\mathbf{x}\|_2$, and the solution is still:

  $$\mathbf{x} = A^+ \mathbf{b},$$

  where we have $\mathbf{y} = A.\mathbf{x} = x_1 \mathbf{A_1} + \cdots + x_n \mathbf{A_n}$ is a vector in $\operatorname{im} A$ closest to $\mathbf{b}$.

Note that A is not HPD, but A^*A is HPD, we may check:

- Hermitian: $(A^*A)^* = A^*A$.
- $\ker(A^*A) = \ker A = \{\mathbf{0}\}$. Hence A^*A is not singular.
- Suppose λ is an eigenvalue of A^*A, we have:

$$A^*A\mathbf{x} = \lambda\mathbf{x}, \text{ where } \mathbf{x} \neq \mathbf{0}.$$

  Thus, we have $(A\mathbf{x})^*A\mathbf{x} = \lambda\mathbf{x}^*\mathbf{x}$, that is $\lambda = \|A\mathbf{x}\|^2/\|\mathbf{x}\|^2 \geqslant 0$, so $\|A.\mathbf{x}\|^2 = \lambda\|\mathbf{x}\|^2$, and thus $\lambda > 0$.

Thus having A^*A being HPD implies that it has Cholesky decomposition, that is $A^*A = R^*R$, and in both cases, we have $\mathbf{x} = A^+\mathbf{b} \iff \mathbf{x} = (A^*A)^{-1}A^*\mathbf{b}$, and $\mathbf{x}$ is the unique solution to $(A^*A)\mathbf{x} = A^*\mathbf{b}$.

Now, the problem is, how do we solve the equivalence without inverting the matrix, that is:

$$A^*A.\mathbf{x} = A^*\mathbf{b} \iff R^* \underbrace{R.\mathbf{x}}_{\mathbf{z}} = A^*\mathbf{b}.$$

⌟

Here we consider the algorithm with input A and $\mathbf{b}$ as:

(i) Compute A^*A. $(\sim mn^2)$.

(ii) Cholesky: $A^*A = R^*R$. $(\sim \frac{1}{3}n^2)$.

(iii) Compute $A^*\mathbf{b}$.

(iv) Solve $R^*\mathbf{z} = A^*\mathbf{b}$ for $\mathbf{z}$, which is upper triangular.

(v) Solve $R\mathbf{x} = \mathbf{z}$ for $\mathbf{x}$, which is lower triangular.

Recall that the first two steps are with the most significant cost, then the total cost is $\sim \frac{1}{3}n^3 + mn^2$. Assuming that $m = n$, we have the complexity as $\frac{4}{3}n^3$.

IV Numerical Computation of Eigenvalues

IV.1 Eigenspace, Reprise

For this part, we want to compute eigenvalues numerically without using the characteristic polynomial.

Definition IV.1.1. Eigenvalue.

$\lambda \in \mathbb{C}$ is an eigenvalue of $A \in \mathbb{C}^{m\times m}$ if there exists $\mathbf{x} \in \mathbb{C}^m$ such that $\mathbf{x} \neq 0$, in which $A.\mathbf{x} = \lambda\mathbf{x}$. ⌟

Here, we note that this it compatible with the original definitionăthat:

$$A.\mathbf{x} = \lambda\mathbf{x} \iff A.\mathbf{x} - \lambda\mathbf{x} = \mathbf{0} \iff A.\mathbf{x} - \lambda\,\mathrm{Id}\,.\mathbf{x} = \mathbf{0} \iff (A - \lambda\,\mathrm{Id})\mathbf{x} = \mathbf{0}.$$

Hence, $A - \lambda\,\mathrm{Id}$ must be singular, that is: λ is an eigenvalue of $A \iff \det(A - \lambda\,\mathrm{Id}) = 0$.

Definition IV.1.2. Characteristic Polynomial.

The characteristic polynomial of $A \in \mathbb{C}^{m\times m}$ is:

$$p_A(t) = \det(A - t\,\mathrm{Id}) = (-1)^n t^n + p_{n-1}t^{n-1} + \cdots + p_1 t + p_0,$$

where the eigenvalues are the roots of $p_A(t)$. ⌟

If $\lambda_1, \lambda_2, \cdots, \lambda_p$ are the p distinct roots of $p_A(t)$, we can write:

$$p_A(t) = (-1)^n (t-\lambda_1)^{n_1}(t-\lambda_2)^{n_2}\cdots(t-\lambda_p)^{n_p},$$

where n_i is the algebraic multiplicity of λ_i.

Definition IV.1.3. Spectrum.

The spectrum of $A \in \mathbb{C}^{m\times m}$ is the set of distinct eigenvalues of A, denoted:

$$\sigma(A) = \{\lambda_1, \lambda_2, \cdots, \lambda_p\}$$

⌟

Definition IV.1.4. Eigenspace.

Given a matrix $A \in \mathbb{C}^{m\times m}$ and λ is a eigenvalue of A, then the eigenspace of λ is:

$$E_\lambda = \{\mathbf{x} : (A - \lambda\,\mathrm{Id})\mathbf{x} = \mathbf{0}\} = \ker(A - \lambda\,\mathrm{Id}),$$

which is the set if all eigenvalues of A associated to λ and $\mathbf{0}$. In particular, we defined the dimension of the eigenspace being the geometric multiplicity of λ. ⌟

Theorem IV.1.5. Property of Geometric Multiplicity.

For any $\lambda_i \in \sigma(A)$, we have $1 \leqslant g_i \leqslant n_i$.

Definition IV.1.6. Non-defective Matrix.

A matrix $A \in \mathbb{C}^{m\times m}$ is called non-defective if for all λ_i, the geometric multiplicity is the same as algebraic multiplicity of λ_i. It is defective if it is non-defective. ⌟

Theorem IV.1.7. Non-defective $\iff$ Diagonalizable.

Let $A \in \mathbb{C}^{m\times m}$, A is non-defective if and only if A is diagonalizable, *i.e.*, there exists nonsingular $S \in \mathbb{C}^{m\times m}$

such that:

$$S^{-1}AS = D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}.$$

Note that we may also rewrite the matrix as:

$$A = SDS^{-1},$$

which is a type of eigenvalue revealing factorization.

A special case is the Hermitian matrices, that is $A^* = A$.

Theorem IV.1.8. Hermitian $\Longrightarrow$ Unitarily Diagonalizable.

Let $A \in \mathbb{C}^{m\times m}$ be Hermitian, then A is unitarily diagonalizable, *i.e.*, there exists a matrix $Q \in \mathbb{C}^{m\times m}$, unitary ($Q^* = Q^{-1}$) such that $Q^*AQ = D$

Definition IV.1.9. Normal Matrix.

A matrix $A \in \mathbb{C}^{m\times m}$ is normal if $A^*A = AA^*$. This aligns with the definition of a normal subgroup, that is, $A^*A(A^*)^{-1} = A$.

Theorem IV.1.10. Normal $\Longleftrightarrow$ Unitarily Diagonalizable.

Let $A \in \mathbb{C}^{m\times m}$, then A is normal if and only if A is unitarily diagonalizable.

Remark IV.1.11. Inclusions of Matrices.

Hermitian Matrices $\subsetneq$ Normal Matrices $=$ Orthogonally Diagonalizable Matrices.

IV.2 Schur Triangularization

Theorem IV.2.1. Schur Lemma.

For any matrix $A \in \mathbb{C}^{m\times m}$, there exists:

- an upper triangular matrix T,
- a unitary matrix Q,

such that $A = QTQ^*$. This is called a Schur factorization.

Proof. For $m = 2$. Let $A \in \mathbb{C}^{m\times m}$, it has at least one eigenvalue λ, then there exists $\mathbf{x} \neq \mathbf{0}$ such that $A.\mathbf{x} = \lambda\mathbf{x}$. We normalize $\mathbf{x}$ as $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$. By denoting $\mathbf{u} = (a, b)$, we choose a vector $\mathbf{v}$ such that $\mathbf{v} \perp \mathbf{u}$, and

$\mathbf{v}$ is unitary vector. (For example, we let $\mathbf{v} = (-\overline{b}, \overline{a})$, so that $\mathbf{v}^*\mathbf{u} = 0$.)
Therefore, let $\mathbb{Q} = \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix}$, it has orthonormal columns, so Q is unitary.
Now, we have:

$$Q^*AQ = \begin{bmatrix} \mathbf{u}^* \\ \mathbf{v}^* \end{bmatrix} A \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^*A\mathbf{u} & \mathbf{u}^*A\mathbf{v} \\ \mathbf{v}^*A\mathbf{u}\mathbf{v}^*A\mathbf{v} & \end{bmatrix},$$

which implies that:

$$\mathbf{u}^*(A.\mathbf{u}) = \mathbf{u}^*(\lambda\mathbf{u}) = \lambda\mathbf{u}^*\mathbf{u} = \lambda,$$
$$\mathbf{v}^*(A.\mathbf{v}) - \mathbf{v}^*(\lambda\mathbf{u}) = \lambda\mathbf{v}^*\mathbf{u} = 0.$$

Now, we have:

$$Q^*AQ = \begin{bmatrix} \lambda & * \\ 0 & * \end{bmatrix} = T,$$

which is upper triangular, so we have $A = QTQ^*$.
For generic m, let $A \in \mathbb{C}^{m\times m}$ having at least one eigenvalue λ, there exists $\mathbf{x} \neq 0$ such that $A.\mathbf{x} = \lambda\mathbf{x}$, and we normalize it to $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$.
Now, we compute $\mathbf{u}$ to an orthonormal basis, that is:

$$\beta = \{\mathbf{u}, \mathbf{v_1}, \cdots, \mathbf{v_m}\} \subset \mathbb{C}^m$$

Here, we want to solve $\mathbf{u}^*\mathbf{v} = 0$, where $\dim(\operatorname{im} u^*) = 1$, $\dim(\ker \mathbf{u}^*) = m-1$, so we can use the Gram-Schmidt to orthogonalize $m-1$ linearly independent solutions and normalize them.
Now, we define $Q = \begin{bmatrix} \mathbf{u} & \mathbf{v_2} & \cdots & \mathbf{v_m} \end{bmatrix} = \begin{bmatrix} \mathbf{u} & V \end{bmatrix}$, so we compute:

$$Q^*AQ = \begin{bmatrix} \mathbf{u}^* \\ V^* \end{bmatrix} A \begin{bmatrix} \mathbf{u} & V \end{bmatrix} = \begin{bmatrix} \mathbf{u}^*A\mathbf{u} & \mathbf{u}^*AV \\ V^*A\mathbf{u} & V^*AV \end{bmatrix} = \begin{bmatrix} \lambda & \mathbf{w}^* \\ \mathbf{0} & B \end{bmatrix}.$$

Now we have:

- $\mathbf{u}^*A\mathbf{u} = \lambda$, and
- $V^*(A\mathbf{u}) = \lambda V^*\mathbf{u} = \lambda \begin{bmatrix} \mathbf{v_i}^*\mathbf{u} \end{bmatrix} = \mathbf{0}$.

By the induction hypothesis, we have $B = U^*RU$, where R is upper triangular, and U is unitary. So $R = U^*BU$.
Now, we will have:

$$Q_2 = \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U \end{bmatrix}.$$

Now, we have:

$$Q_2^*(Q_1^*AQ_1)Q_2 = \begin{bmatrix} \lambda & * \\ \mathbf{0} & R \end{bmatrix} = T.$$

Thus, we have shown that $A = QTQ^*$. □

Such decomposition is called a *Schur triangularization*.

Remark IV.2.2. Eigenvalues of Upper Triangular Matrix.

Let the upper triangular matrix T be:

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ 0 & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{mm} \end{bmatrix},$$

the eigenvalues of T is $t_{11}, t_{22}, \cdots, t_{mm}$, which is exactly the eigenvalues of A, *i.e.*, T and A have the same spectrum.

Thus, we consider *Schur triangularization* as an eigenvalue-revealing factorization.

Theorem IV.2.3. Hermitian $\Longrightarrow$ Diagonal Decomposition.

If A is Hermitian, then for the Schur triangularization $A = QTQ^*$, T is diagonal.

Proof. From the decomposition $T = Q^*AQ$, so $T^* = (Q^*AQ)^* = Q^*A^*(Q^*)^* = Q^*AQ = T$, so $T^* = T$, *i.e.*, T is Hermitian.

But, also note that T is triangular, then:

$$T^* = \begin{bmatrix} \overline{t_{11}} & 0 & \cdots & 0 \\ \overline{t_{12}} & \overline{t_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \overline{t_{1m}} & \overline{t_{2m}} & \cdots & \overline{t_{mm}} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ 0 & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{mm} \end{bmatrix} = T.$$

Thus, we must have all non-diagonal entries being zero, and the diagonal being real. $\square$

A consequence of the above statement is that for Hermitian matrices, the *Schur triangularization* is just a *unitary diagonalization*, since $Q^{-1} = Q^*$.

IV.3 Obstructions to Finding Eigenvalues

Now, we have learned three eigenvalue revealing factorization.

- A is *non-defective* $\Longleftrightarrow$ A is diagonalizable, *i.e.*, there exists nonsingular S such that $A = S\Lambda S^{-1}$.
- A is *normal* ($A^*A = AA^*$) $\Longleftrightarrow$ A is unitarily diagonalizable, *i.e.*, there exists unitary Q such that $A = Q\Lambda Q^*$.
- Any $A \in \mathbb{C}^{m\times m}$ has a Schur triangularization, *i.e.*, there exists upper triangular T such that there exists unitary Q in which $A = QTQ^*$.

Remark IV.3.1. Traditional Way to Compute Eigenvalues.

Let a matrix $A \in \mathbb{C}^{m\times m}$, we do the following steps:

(i) Compute the characteristic polynomial:

$$p_A(t) = \det(A - t\,\mathrm{Id}). \qquad \text{(Cost: } \mathcal{O}(m^3).)$$

(ii) Find the roots of the characteristic polynomial $p_A(t)$. (Unstable process.)

Note that for the second step, for the roots of a degree 2 polynomial, we can use a quadratic formula, where $\lambda_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. For degree 3 and 4 polynomials, there are respectively *cubic formula* and *quartic formula*, which are very complex. Although being very complex, we can find the roots within a finite number of operations (*i.e.*, $+, -, \times, \div, \sqrt[k]{(-)}$).

Theorem IV.3.2. Abel-Ruffini Theorem.

There is no "quintic formula."
For any $m \geqslant 5$, there exists a polynomial $p(t)$ of degree m that has a root r that cannot be represented computed using a finite number of operation (*i.e.*, $+, -, \times, \div, \sqrt[k]{(-)}$).

This theorem is a direct consequence of Galois correspondence and the Galois correspondence group being unsolvable.

A consequence is that when $m \geqslant 5$, the roots of $p(t)$ can only (in general) be approximated with an infinite number of operations, which is called *iterative methods*.

However, finding roots of polynomials with iterative methods is *unstable*, so to estimate eigenvalues of A, we will avoid finding roots of characteristic polynomials.

A fact is that if $A \in \mathbb{C}^{m \times m}$ with $m \geqslant 5$, then all eigenvalue solvers are iterative.

Remark IV.3.3. General Scheme to Find Eigenvalues.

For a given $A \in]C^{m \times m}$, we will deploy iterative methods to find an approximation of Schur factorization (eigenvalue revealing), so we find a sequence $\{Q_i\}_{i=1}^{\infty}$ of unitary matrices such that:

$$Q_n^* \cdots Q_2^* Q_1^* A (Q_1 Q_2 \cdots Q_n) \xrightarrow{n \to \infty} T,$$

which is upper triangular, and Q_n is unitary. (T will be diagonal if A is Hermitian).
In particular, we consider the scheme as two phases, in which:

$$\underset{A}{\begin{bmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \cdots & * \end{bmatrix}} \xrightarrow{\text{Phase 1}} \underset{H}{\begin{bmatrix} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & \cdots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & * & * \end{bmatrix}} \xrightarrow{\text{Phase 2}} \underset{T}{\begin{bmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & * \end{bmatrix}}$$

Here, phase 1 turns the matrix A into Hessenberg form, that is having zeros below the first subdiagonal.

- For phase 1, it has finite number of steps, *i.e.*, $\mathcal{O}(m^3)$ flops.
- For phase 2, there needs to be finitely many steps, that is:
  - There needs to be $\mathcal{O}(m)$ iterations to achieve convergence to T.
  - Each iteration needs $\mathcal{O}(m^2)$ flops.

So the total is $\mathcal{O}(m)\mathcal{O}(m^2) = \mathcal{O}(m^3)$ flops.

If we apply Phase 2 directly to A, the the number of flops is $\mathcal{O}(m^4)$ or higher.

Remark IV.3.4. Special Hermitian Case of Decomposition.

When A is Hermitian, then H and T are also Hermitian, so H is:

$$H = \begin{bmatrix} * & * & 0 & \cdots & 0 \\ * & * & * & \cdots & 0 \\ 0 & * & * & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & * \end{bmatrix},$$

and T is diagonal. H has nonzero entries on the diagonal, first sub-diagonal and first super-diagonal.

Our scheme is as follows:

(i) We have:

$$A \xmapsto{\rho(U_1,\bullet)} H = U_1 A U_1^*,$$

where this is the conjugation action of U_1 and U_1 is unitary, so it is a similarity transformation. Here, we let $U_1 = Q_1 Q_2 \cdots Q_{m-2}$, where Q_j is unitary matrix that performs a Householder reflection, which is normally used for QR factorization, that is:

$$A = QR,$$

which is not a similarity transformation (which is different from $A = U_1^* H U$).

(a) For the first step of QR factorization, we have $A \in \mathbb{C}^{m\times m}$ as:

$$\begin{bmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \cdots & * \end{bmatrix} \xmapsto{Q_1\cdot\bullet} \begin{bmatrix} \star & \star & \star & \cdots & \star \\ 0 & \star & \star & \cdots & \star \\ 0 & \star & \star & \cdots & \star \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \star & \star & \cdots & \star \end{bmatrix},$$

in which the first column got transformed into zero entries except for the first row.

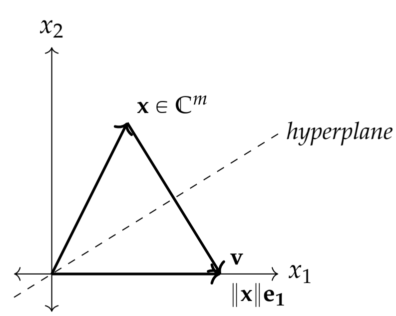


Figure IV.1. Decomposition as a reflection in QR factorization.

Here, we define $\mathbf{v} = \|\mathbf{x}\|\mathbf{e_1} - \mathbf{x}$, so:

$$F = \mathrm{Id} - 2\frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}},$$

and we let $Q_1 = F$.

Proposition IV.3.5. Properties of First Step in Factorization.

Here are the properties of F and Q_1:

- $F^* = F$, which is Hermitian,
- $F^{-1} = F^*$, which is unitary, thus
- $F^{-1} = F$, which is involutory.

(b) In the second step of QR, we have the step as:

$$\begin{bmatrix} \star & \star & \star & \cdots & \star \\ 0 & \star & \star & \cdots & \star \\ 0 & \star & \star & \cdots & \star \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \star & \star & \cdots & \star \end{bmatrix} \xmapsto{Q_2 \cdot \bullet} = \begin{bmatrix} \star & \star & \star & \cdots & \star \\ 0 & \clubsuit & \clubsuit & \cdots & \clubsuit \\ 0 & 0 & \clubsuit & \cdots & \clubsuit \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \clubsuit & \cdots & \clubsuit \end{bmatrix},$$

Here, we define $\mathbf{v} = \|\mathbf{x}\|\mathbf{e_1} - \mathbf{x}$, with

$$F = \mathrm{Id}_{m-1} - 2\frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}},$$

so we have:

$$Q_2 = \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & F \end{bmatrix},$$

so Q_2 is hermitian, unitary, and involutory.

(c) Then, we repeat such step for all k, that is having:

$$Q_k = \begin{bmatrix} \mathrm{Id}_{k-1} & \mathbf{0}^* \\ \mathbf{0} & F \end{bmatrix},$$

which is hermitian, unitary, and involutory.

Now, we get after $m-1$ steps to get $Q_{m-1}Q_{m-2}\cdots Q_2Q_1A = R$, so we have:

$$(Q_{m-1}\cdots Q_2Q_1)^[-1] = Q_1^{-1}Q_2^{-1}\cdots Q_{m-1}^{-1} = Q_1Q_2\cdots Q_{m-1}.$$

Now, we let $A = QR$ where $Q = Q_1\cdots Q_{m-1}$, which is not eigenvalue revealing.

(ii) Here, we make the first attempt to compute the Hessenberg form, which is to:

- Perform first steps of QR, that is:
$$A \mapsto Q_1A,$$
- We compute $A \mapsto Q_1AQ_1$, which is similarity transformation, which preserves eigenvalues. In particular, $(Q_1A)Q_1$ performs on columns of Q_1A the same operations that Q_1 operated on the rows of A.
Here, we may not necessarily have Q_1AQ_1 to have zeros (except for the first row) to be preserves. So our work is undone.

Remark IV.3.6. Equivalent with Abel-Ruffini Theorem.

Note that this is reasonable since if we could have get the eigenvalues of a matrix within finitely many steps, that is equivalently solving for zeros of polynomials of arbitrarily large degree, which is a contradiction to Abel-Ruffini Theorem.

(iii) Thus, we start over with another attempt, here, we use the last $m-2$ entries for the $\mathbf{x}$ in householder reflection, where we let:

$$\begin{bmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \cdots & * \end{bmatrix} \xmapsto{Q_1} \bullet \begin{bmatrix} * & * & * & \cdots & * \\ \star & \star & \star & \cdots & \star \\ 0 & \star & \star & \cdots & \star \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \star & \star & \cdots & \star \end{bmatrix} \xmapsto{\bullet Q_1} \begin{bmatrix} * & \clubsuit & \clubsuit & \cdots & \clubsuit \\ \star & \clubsuit & \clubsuit & \cdots & \clubsuit \\ 0 & \clubsuit & \clubsuit & \cdots & \clubsuit \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \clubsuit & \clubsuit & \cdots & \clubsuit \end{bmatrix}$$

Note that the zeros are preserves.

Then, we do the same thing with Q_2 with the last $m-3$ entries on the second column as $\mathbf{x}$ in the Householder reflection. Thus, we have:

$$\begin{bmatrix} * & \clubsuit & \clubsuit & \cdots & \clubsuit \\ \star & \clubsuit & \clubsuit & \cdots & \clubsuit \\ 0 & \clubsuit & \clubsuit & \cdots & \clubsuit \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \clubsuit & \clubsuit & \cdots & \clubsuit \end{bmatrix} \xmapsto{Q_2 \bullet Q_2} \begin{bmatrix} * & \clubsuit & \heartsuit & \cdots & \heartsuit \\ \star & \clubsuit & \heartsuit & \cdots & \heartsuit \\ 0 & \heartsuit & \heartsuit & \cdots & \heartsuit \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \heartsuit & \cdots & \heartsuit \end{bmatrix} \xmapsto{Q_3 \bullet Q_3} \cdots \xmapsto{Q_{m-2} \bullet Q_{m-2}} \begin{bmatrix} * & \clubsuit & \heartsuit & \cdots & ? \\ \star & \clubsuit & \heartsuit & \cdots & ? \\ 0 & \heartsuit & \heartsuit & \cdots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & ? \end{bmatrix},$$

in which our desired result is in Hessenberg form, which is exactly:

$$A = \underbrace{(Q_1 Q_2 \cdots Q_{m-2})}_{U_1} H \underbrace{(Q_{m-2}) \cdots Q_2 Q_2}_{U_1^*}.$$

Proposition IV.3.7. Hermitian $\implies$ Hessenberg is Tridiagonal.

If A is Hermitian, then H is Tridiagonal, *i.e.*, having only diagonal, sub-diagonal, and super-diagonal possibly nonzero.

Proof. Here, we have $H = U_1^* A U_1$, then $H^* = U_1^* A^* U_1 = U_1^* A U_1 = H$, so H is Hermitian implies that it is upper triangular, so H is Tridiagonal. $\square$

Here, we make the simplifying assumption that $A = A^*$, that is to computer the singular values of B by computing eigenvalues of $A = B^* B$, which is Hermitian, *i.e.*, $A^* = A$.

Therefore, the consequences are that:

- Eigenvalues are red, and
- A has unitary diagonalization, *i.e.*, there exists unitary Q ($Q^* = Q^{-1}$) as the orthonormal eigenvectors of A.

Definition IV.3.8. Rayleigh Quotient.

Let A be Hermitian, the Rayleigh quotient of A is the function $r : \mathbb{C}^m \backslash \{\mathbf{0}\} \to \mathbb{C}$, defined as:

$$r(\mathbf{x}) = \frac{\mathbf{x}^* A \mathbf{x}}{\mathbf{x}^* \mathbf{x}},$$

which is the quadratic form over $\|\mathbf{x}\|^2$. ⌟

Proposition IV.3.9. Rayleigh Quotient of Eigenvector is its Eigenvalue.

Suppose $\mathbf{v} \neq \mathbf{0}$ is an eigenvector of A associated to eigenvalue λ, *i.e.*, $A.\mathbf{v} = \lambda\mathbf{v}$, then:

$$r(\mathbf{v}) = \frac{\mathbf{v}^* A\mathbf{v}}{\mathbf{v}^*\mathbf{v}} = \frac{\mathbf{v}^*(\lambda\mathbf{v})}{\mathbf{v}^*\mathbf{v}} = \lambda.$$

With this property, we can find the eigenvalue as long as we can find the eigenvectors.

Remark IV.3.10. Geometric Interpretation of Rayleigh Quotient.

Let $A \in \mathbb{C}^{m\times m}$ be Hermitian, and fix any nonzero vector $\mathbf{x} \in \mathbb{C}^m$. Consider the function $f : \mathbb{R} \to \mathbb{R}$ as:

$$f(\alpha) = \|A\mathbf{x} - \alpha\mathbf{x}\|_2.$$

We want to minimize this w.r.t α. If $\mathbf{x}$ is an eigenvector, and for $\alpha = \lambda$, which is the eigenvalue, we have

$$f(\lambda) = \|A.\mathbf{v} - \lambda\mathbf{v}\|_2 = 0,$$

which is the least value for a normed vector space. However, we want to think of $\mathbf{x}$ when it is not the eigenvector. ⌟

Theorem IV.3.11. Minimizer for f is Rayleigh Quotient.

Let $A \in \mathbb{C}^{m\times m}$ be Hermitian. Then, for any $\mathbf{x} \neq \mathbf{0}$, the minimizer of $f(\alpha) = \|A.\mathbf{x} - \alpha\mathbf{x}\|_2$ is $\alpha = r(\mathbf{x})$, which is the eigenvalue of $\mathbf{x}$ when $\mathbf{x}$ is eigenvector.

Proof. Equivalently, $f(\alpha)$ is minimizing:

$$\begin{aligned}
\big(f(\alpha)\big)^2 &= \|A.\mathbf{x} - \alpha\mathbf{x}\|_2^2 = (A.\mathbf{x} - \alpha\mathbf{x})^*(A\mathbf{x} - \alpha\mathbf{x}) \\
&= \mathbf{x}^* A^* A\mathbf{x} - \alpha\mathbf{x}^* A\mathbf{x} - \alpha\mathbf{x}^* A^*\mathbf{x} + \alpha^2\mathbf{x}^*\mathbf{x} \\
&= \mathbf{x}^* A^2 - 2\alpha\mathbf{x}^* A\mathbf{x} + \alpha^2\mathbf{x}^*\mathbf{x}.
\end{aligned}$$

Then, we take the derivatives as:

$$\frac{d}{d\alpha}\left[\big(f(x)\big)^2\right] = -2\mathbf{x}^* A\mathbf{x} + 2\alpha\mathbf{x}^*\mathbf{x} = 0,$$

which solves as:

$$\alpha = \frac{\mathbf{x}^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = r(\mathbf{x}),$$

as desired. □

We have seen that if $\mathbf{x} = \mathbf{v}$, then $r(\mathbf{v}) = \lambda$, and $\mathbf{v}$ is the stationary point of $r(\mathbf{x})$, so that:

$$\nabla r(\mathbf{v}) = \mathbf{0}.$$

Another assumption us that $A \in \mathbb{R}^{m\times m}$ so that for $\mathbf{x} \in \mathbb{R}^m$, we have $A^* = A$ implying that $A^\intercal = A$.

Theorem IV.3.12. Gradient of Rayleigh Quotient.

The gradient $\nabla r(\mathbf{x}) = \begin{bmatrix} \frac{\partial r}{\partial x_1}(\mathbf{x}) \\ \frac{\partial r}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial r}{\partial x_m}(\mathbf{x}) \end{bmatrix}$, where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$ given by:

$$\nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^*\mathbf{x}}(A\mathbf{x} - r()\mathbf{x}) \in \mathbb{C}^m.$$

Proposition IV.3.13. Formal Derivatives.

With the same construction, we have:

(i) $\dfrac{\partial}{\partial x_i}(\mathbf{x}^*\mathbf{x}) = 2x$, and

(ii) $\dfrac{\partial}{\partial x_i}(\mathbf{x}^* A\mathbf{x}) = 2(A_{\mathbf{x}})_i$, that is the i-th entry of vector $A.\mathbf{x}$.

Proof. Recall from calculus, we have:

$$\frac{d}{dx}\left[\frac{f}{g}\right] = \frac{f'g - fg'}{g^2}.$$

Then, the partial derivative is:

$$\begin{aligned}
\frac{\partial}{\partial x_i} r(\mathbf{x}) &= \frac{\partial}{\partial x_i}\frac{\mathbf{x}^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \frac{\frac{\partial}{\partial x_i}(\mathbf{x}^* A\mathbf{x}) \cdot \mathbf{x}^*\mathbf{x} - (\mathbf{x}^* A\mathbf{x})\frac{\partial}{\partial x_i}(\mathbf{x}^*\mathbf{x})}{(\mathbf{x}^*\mathbf{x})^2} \\
&= \frac{2(A.\mathbf{x})_i\mathbf{x}^*\mathbf{x} - (\mathbf{x}^* A\mathbf{x})2x_i}{(\mathbf{x}^*\mathbf{x})^2} \\
&= \frac{2}{\mathbf{x}^*\mathbf{x}}\left[(A\mathbf{x})_i - \frac{\mathbf{x}^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}}x_i\right] \\
&= \frac{2}{\mathbf{x}^*\mathbf{x}}\left[(A.\mathbf{x}) - \big(r(\mathbf{x})\mathbf{x}\big)_i\right] \\
&= \frac{2}{\mathbf{x}^*\mathbf{x}}\big(A\mathbf{x} = r(\mathbf{x})\mathbf{x}\big)_i.
\end{aligned}$$

Hence, we have:

$$\nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^*\mathbf{x}}\big(A\mathbf{x} - r(\mathbf{x})\mathbf{x}\big),$$

which completes the proof. □

Note that if $\mathbf{x} = \mathbf{v}$, we have:

$$\nabla \mathbf{r}(\mathbf{v}) = \frac{2}{\mathbf{v}^*\mathbf{v}}\big(A\mathbf{v} - r(\mathbf{v})\mathbf{v}\big) = \mathbf{0},$$

and thus $\mathbf{v}$ is a stationary point of $r(\mathbf{x})$.

Proposition IV.3.14. Eigenvectors are Critical Points.

$\mathbf{v}$ is an eigenvector of A if and only if $\nabla r(\mathbf{v}) = \mathbf{0}$.

Proof. ($\Longrightarrow$:) Assume that $\mathbf{x} = \mathbf{v}$, then we have:

$$r(\mathbf{v}) = \frac{2}{\mathbf{v}^*\mathbf{v}}(A\mathbf{v} - \underbrace{r(\mathbf{v})}_{\lambda}\mathbf{v}) = \mathbf{0}.$$

($\Longleftarrow$:) Assume that $\nabla r(\mathbf{v}) = \mathbf{0}$, then $A.\mathbf{v} - r(\mathbf{v})\mathbf{v} = \mathbf{0}$, so $A.\mathbf{v} - r(\mathbf{v})\mathbf{v} = \mathbf{0}$, thus $A.\mathbf{v} = r(\mathbf{v})\mathbf{v}$, so $\mathbf{v}$ is eigenvector of A. $\square$

Recall for a Taylor series, we have:

$$f(x) \sim f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \text{higher order terms.}$$

For $r : (\mathbb{R}^m)^\times \to \mathbb{R}$, we have:

$$r(\mathbf{x}) = r() + \left[\nabla r(\mathbf{x_0})\right]^\mathsf{T}(\mathbf{x} - \mathbf{x_0}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_0})^\mathsf{T}\mathrm{H}_r(\mathbf{x_0})(\mathbf{x} - \mathbf{x_0}) + \text{higher order terms,}$$

where we have:

$$\mathrm{H}_r(\mathbf{x_0}) = \left[\tfrac{\partial r}{\partial x_i \partial y_j}(\mathbf{x_0})\right]$$

Here, we choose $\mathbf{x_0} = \mathbf{v}$, which is one eigenvector of A associated to eigenvalue λ, we have $r(\mathbf{v}) = \lambda$, $\nabla r(\mathbf{v}) = \mathbf{0}$. For $\mathbf{x}$ near $\mathbf{v}$, we have:

$$r(\mathbf{x}) \simeq \lambda + \frac{1}{2}(\mathbf{x} - \mathbf{v})^\mathsf{T}\mathrm{H}_r(\mathbf{v})(\mathbf{x} - \mathbf{v}),$$

i.e.:

$$r(\mathbf{x}) - \lambda \simeq \frac{1}{2}(\mathbf{x} - \mathbf{v})^\mathsf{T}\mathrm{H}_r(\mathbf{v})(\mathbf{x} - \mathbf{v}).$$

Recall that if $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$, and $\|\bullet\|$ is a matrix norm induced by a vector norm, then:

$$\|AB\| \leqslant \|A\| \cdot \|B\|.$$

Another fact is that if $\mathbf{x} \in \mathbb{C}^{m\times 1}$ (or $\mathbf{y} \in \mathbb{C}^{1\times m}$) and $\|\bullet\|_M$ is a matrix norm induced by a vector norm $\|\bullet\|_V$, we can interpret $\mathbf{x}$ or $\mathbf{y}$ as a matrix, and we will have that the $\|\mathbf{x}\|_M$ equal to the vector norm $\|xx\|_V$ that induced the matrix norm.

Sketch of proof. Suppose $\|\bullet\|_M$ is the matrix norm induced by the vector 2 norm $\|\bullet\|_2$, and suppose $\mathbf{x} \in \mathbb{C}^+m \times 1$, we compute the matrix norm of $\mathbf{x}$ as:

$$\|\mathbf{x}\|_M = \sup_{\substack{\mathbf{u}\in\mathbb{C}^1 \\ \|\mathbf{u}\|_2 = 1}} \|\mathbf{x}u\|_2 = \max_{\substack{u\in\mathbb{C} \\ |u|=1}} |u|\|\mathbf{x}\|_2 = \|\mathbf{x}\|_2.$$

$\square$

Hence, we get that:

$$\begin{aligned} |r(\mathbf{x}) - \lambda| &\simeq \frac{1}{2}\big|\underbrace{(\mathbf{x} - \mathbf{v})^\mathsf{T}\mathrm{H}_r(\mathbf{v})(\mathbf{x} - \mathbf{v})}_{\mathbb{C}^{1\times 1}}\big| \\ &= \frac{1}{2}\|(\mathbf{x} - \mathbf{v})^\mathsf{T}\mathrm{H}_r(\mathbf{v})(\mathbf{x} - \mathbf{v})\| \\ &\leqslant \frac{1}{2}\underbrace{\|(\mathbf{x} - \mathbf{v})^\mathsf{T}\|}_{\|\mathbf{x}-\mathbf{v}\|_2} \cdot \underbrace{\|\mathrm{H}_r(\mathbf{v})\|}_{\text{does not depend on } \mathbf{x}} \cdot \underbrace{\|(\mathbf{x} - \mathbf{v})\|}_{\|\mathbf{x}-\mathbf{v}\|_2}. \end{aligned}$$

Therefore, we obtain that:

$$|r(\mathbf{x}) - \lambda| \leqslant C\|\mathbf{x} - \mathbf{v}\|_2^2,$$

i.e., $|r(\mathbf{x}) - \lambda| = \mathcal{O}(\|\mathbf{x} - \mathbf{v}\|_2^2)$.

Therefore, the Rayleigh quotient is a *quadratically accurate* estimate of λ.

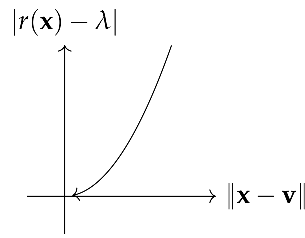


Figure IV.2. Convergence of Rayleigh quotient is quadratic.

IV.4 Power Method

Here, we first introduce another assumption. Consider A with m distinct eigenvalues, we order them in decreasing order:

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_m|.$$

The spectral gap ratio is $\frac{|\lambda_1|}{|\lambda_2|}$. (This only accounts for the first two.)

Here, we consider the **power method**

```
input: A
       vec_v^0 in C^m       % initial guess for eigenvector of lambda_1 and its norm is 1
while True:
       vec_w = A*vec_v^(k-1)
       vec_v^(u) = vec_w / ||vec_w||
       lambda^(k) = r(vec_v^(u)) = (vec_v^(k)^T*A*vec_V^(k)) / ((vec_v^(u))*vec_v^(u))
       % Termination condition
```

For the algorithm, we have the termination condition of:

$$|\lambda^{(k)} - \lambda^{(k-1)}| < \epsilon_1 \text{ or } \left| \frac{\lambda^{(k)} - \lambda^{(k-1)}}{\lambda^{(k-1)}} \right| < \epsilon_2.$$

Suppose that $\beta = \{\mathbf{q_1}, \mathbf{q_2}, \cdots, \mathbf{q_m}\}$ is orthonormal. Let the eigenbasis of $\mathbb{C}^m$ for A, we can write $\mathbf{v}^{(0)}$ as:

$$\mathbf{v}^{(0)} = c_1\mathbf{q_1} + c_2\mathbf{q_2} + \cdots + c_m\mathbf{q_m}.$$

In the first step, we have:

$$\begin{aligned}
\mathbf{w} &= A\mathbf{v}^{(0)} = c_1 A.\mathbf{q_1} + c_2 A.\mathbf{q_2} + \cdots + c_m A.\mathbf{q_m} = c_1\lambda_1\mathbf{q_1} + c_2\lambda_2\mathbf{q_2} + \cdots + c_m\lambda_m\mathbf{q_m}, \\
\mathbf{v}^{(1)} &= \frac{\mathbf{w}}{\|\mathbf{w}\|} = d_1\left(c_1\lambda_1\mathbf{q_1} + c_2\lambda_2\mathbf{q_2} + \cdots + c_m\lambda_m\mathbf{q_m}\right), \\
\lambda^{(1)} &= r(\mathbf{v}^{(1)}).
\end{aligned}$$

Then, for the second step, we have:

$$\begin{aligned}
\mathbf{w} &= A\mathbf{v}^{(0)} = d_1(c_1\lambda_1^2\mathbf{q_1} + c_2\lambda_2^2\mathbf{q_2} + \cdots + c_m\lambda_m^2\mathbf{q_m}), \\
\mathbf{v}^{(2)} &= \frac{\mathbf{w}}{\|\mathbf{w}\|} = d_2(c_1\lambda_1^2\mathbf{q_1} + c_2\lambda_2^2\mathbf{q_2} + \cdots + c_m\lambda_m^2\mathbf{q_m}), \\
\lambda^{(2)} &= r(\mathbf{v}^{(2)}).
\end{aligned}$$

In the k-th step, we have:

$$\mathbf{v}^{(k)} = d_k(c_1\lambda_1^k\mathbf{q_1} + c_2\lambda_2^k\mathbf{q_2} + \cdots + c_m\lambda_m^k\mathbf{q_m}),$$
$$\lambda^{(k)} = r(\mathbf{v}^{(k)}).$$

Here, we may rewrite the vector as:

$$\mathbf{v}^{(k)} = d_k\lambda_1^k\Bigg(c_1\mathbf{q_1} + \underbrace{c_2\left(\frac{\lambda_1}{\lambda_2}^k\mathbf{q_2}\right) + \cdots + c_m\left(\frac{\lambda_m}{\lambda_1}\right)^k\mathbf{q_m}}_{\to 0 \text{ as } k\to\infty}\Bigg).$$

Thus, as k is arbitrarily large, we have:

$$\mathbf{v}^{(k)} \simeq (d_k\lambda_1^k c_1)\mathbf{q_1}.$$

Since $\|\mathbf{q_1}\|_2 = 1$ and $\|\mathbf{v}^{(k)}\|_2 = 1$, we have $|b_k\lambda_1^k c_1| = 1$, so we have:

$$\text{either } \mathbf{v}^{(k)} \text{ to } \mathbf{q_1} \text{ or } \mathbf{v}^{(m)} \to -\mathbf{q_1}.$$

Remark IV.4.1. Computation Complexity of Power Method.

The *most costly* operations are with:

$$\mathbf{w}^{(k)} = A\mathbf{v}^{(k-1)} \text{ and } \lambda^{(k)} = r\big(\mathbf{v}^{(k)}\big)A\mathbf{v}^{(k)}.$$

It would be cheaper if A is in *Heisenberg* or *Tri-diagonal* form. ⌟

Proposition IV.4.2. Rate of Convergence for Power Method.

The rate of convergence for the power method is dependent on λ_2/λ_1, namely:

$$\|\mathbf{v}^{(k)} - (\pm\mathbf{q_1})\| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right),$$
$$|\lambda^{(k)} - \lambda| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right).$$

Here, we can also obtain the following, more formal, result.

Theorem IV.4.3. Eigenvectors and Eigenvalues for Manipulations in Matrices.

Let $A \in \mathbb{C}^{m\times m}$ and let $\mathbf{v}$ be an eigenvalue of A associated with eigenvalue λ.

(i) For any $\mu \in \mathbb{C}$, we have that $\lambda - \mu$ being an eigenvalue of $A - \mu\,\mathrm{Id}$ with same eigenvector $\mathbf{v}$.

(ii) If A is nonsingular, then $1/\lambda$ is an eigenvalue of A^{-1} with the same of eigenvector $\mathbf{v}$.

Proof. (i) Notice that:

$$(A - \mu\,\mathrm{Id}).\mathbf{v} = A.\mathbf{v} - \mu\,\mathrm{Id}\,.\mathbf{v} = \lambda\mathbf{v} - \mu\mathbf{v} = (\lambda - \mu)\mathbf{v}.$$

(ii) Here, $\lambda \neq 0$ since A is nonsingular, thus:

$$A.\mathbf{v} = \lambda\mathbf{v} \text{ which implies that } A^{-1}A.\mathbf{v} = A^{-1}.(\lambda\mathbf{v}),$$

thus, we have that:

$$\mathbf{v} = \lambda A^{-1}\mathbf{v},$$

thus we result in $A^{-1}.\mathbf{v} = \frac{1}{\lambda}\mathbf{v}$, as desired. □

As an immediate consequence of the previous theorem, we have the following result.

Proposition IV.4.4. Corollary on Manipulation of Matrices.

If λ is an eigenvalue of A with eigenvector, and $\mu \in \mathbb{C}$ is arbitrary such that $A - \mu\,\mathrm{Id}$ being nonsingular, then $\frac{1}{\lambda-\mu}$ is an eigenvalue of $(A - \mu\,\mathrm{Id})^{-1}$ with the same eigenvector $\mathbf{v}$.

In particular, the above corollary allows us to modify the gaps of the eigenvalues while preserving eigenvectors, it helps to improve the power method.

Suppose eigenvalues of A are:

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_m|, \text{ and } \frac{|\lambda_2|}{|\lambda_1|} \lessapprox 1.$$

Graphically, we would have:

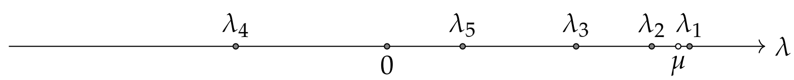


Figure IV.3. Eigenvalues of A on the axis with μ close to λ_1.

When μ is closer to λ_1 than λ_2, we have:

$$|\lambda_1 - \mu| \ll |\lambda_2 - \mu| < |\lambda_j - \mu| \text{ for all } j \geqslant 3.$$

Therefore, we have the inversion as:

$$\frac{1}{|\lambda_1 - \mu|} \gg \frac{1}{|\lambda_2 - \mu|} > \frac{1}{|\lambda_j - \mu|} \text{ for all } j \geqslant 3.$$

Here, the eigenvalues of $(A - \mu\,\mathrm{Id})^{-1}$ are:

$$\frac{1}{\lambda_1 - \mu}, \frac{1}{\lambda_2 - \mu}, \cdots, \frac{1}{\lambda - \mu},$$

with the same eigenvectors (of A) as:

$$\mathbf{q_1}, \mathbf{q_2}, \cdots, \mathbf{q_m}.$$

This becomes the **inverse iteration algorithm**.

```
input: A
       vec_v^0
       mu
for k = 1, 2, 3, ...:
       vec_w = (A - mu*Id)^(-1) vec_v^(k-1)       % inverse matrix here, we avoid this
       vec_v^(k) = vec_w / ||vec_w||
       lambda^(k) = r(vec_v^(k)) = (vec_v^(k))^T A vec_v^(k)
```

Remark IV.4.5. Avoid doing Inverses.

Especially for the noted step, there is invert of a matrix, so we instead solve instead:

$$(A - \mu\,\mathrm{Id})\mathbf{w} = \mathbf{v}^{(k-1)} \text{ for } \mathbf{w},$$

for example using LU decomposition. ⌟

Suppose now that we have:

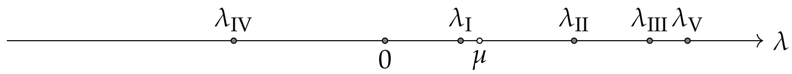


Figure IV.4. Eigenvalues of A on the axis with μ elsewhere.

In this case, we have:

$$|\lambda_{\text{I}} - \mu| < |\lambda_{\text{II}} - \mu| < |\lambda_i - \mu| \text{ for all other eigenvalues } \lambda_i.$$

Hence, we have:

$$\frac{1}{|\lambda_{\text{I}} - \mu|} > \frac{1}{|\lambda_{\text{II}} - \mu|} > \frac{1}{|\lambda_i - \mu|} \text{ for all other eigenvalues } \lambda_i.$$

Hence, the dominant eigenvalue of $(A - \mu \, \text{Id})^{-1}$ is:

$$\frac{1}{\lambda_{\text{I}} - \mu}.$$

Then, we apply inverse iteration algorithm, and we will get:

$$\mathbf{v}^{(k)} \xrightarrow{k \to \infty} \mathbf{q}_{\mathbf{I}} \qquad \text{and} \qquad \lambda^{(k)} \xrightarrow{k \to \infty} \lambda_{\text{I}}.$$

Iteration-wise, we have:

$$\mathbf{v}^{(k)} = d_k \lambda_{\text{I}} \left[c_{\text{I}} \mathbf{q}_{\mathbf{I}} + c_{\text{II}} \left(\frac{\lambda_{\text{II}}}{\lambda_{\text{I}}} \right)^k \mathbf{q}_{\mathbf{II}} + \cdots \right].$$

Proposition IV.4.6. Rate of Convergence for Inverse Iteration Algorithm.

The rate of convergence for the modified algorithm is:

$$\|\mathbf{v}^{(k)} - (\pm \mathbf{q}_{\mathbf{I}})\|_2 = \mathcal{O}\left(\left| \frac{\lambda_{\text{II}}}{\lambda_{\text{I}}} \right|^k \right)$$

$$|\lambda^{(k)} - \lambda_{\text{I}}| = \mathcal{O}\left(\left| \frac{\lambda_{\text{II}}}{\lambda_{\text{I}}} \right|^{2k} \right).$$

To make this faster, we want to update μ so it is closer to λ_{I}, *i.e.*, we can change $\mu^{(k)} := \lambda^{(k)}$. This leads to **Raileigh Quotient Iteration Algorithm**.

```
input: A in C(m*m)
       vec_v^(0)
       mu
for k = 1, 2, 3, ...:
       solve (A - mu*Id)vec_w = vec_v^(k-1) for vec_w
       vec_v^(u) = vec_w / ||vec_w||
       lambda^(k) = r(vec_v^(k))^* Avec_v^(k)
       mu = lambda^(k)
```

Within machine precision, this algorithm is *clearly* faster.