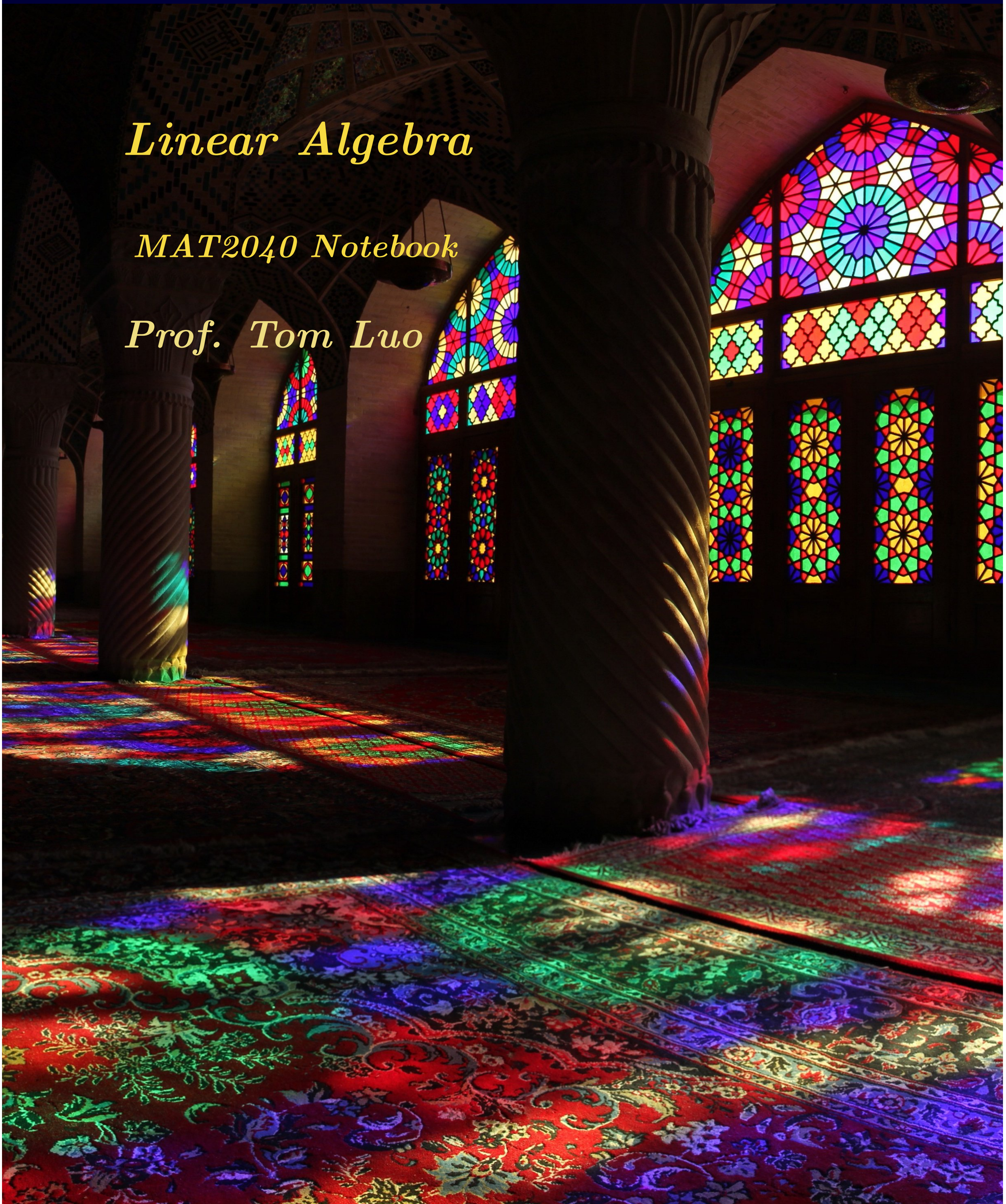
香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen



Linear Algebra

MAT2040 Notebook

Prof. Tom Luo

The First Edition

A FIRST COURSE

IN

LINEAR ALGEBRA

A FIRST COURSE

IN

LINEAR ALGEBRA

MAT2040 Notebook

Prof. Tom Luo

The Chinese University of Hongkong, Shenzhen

Prof. Ruoyu Sun

University of Illinois Urbana-Champaign

香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

Contents

Preface

This book is intended for the foundation course MAT2040, which is the first course on the linera algebra. It aims to cover basic linear algebra knowledge and its simple applications. This book was first written in 2017, and it is reviewed and revised in 2018. We have correct several mistakes shown in the previous book and modify some proofs a little bit to give readers better insights of linear algebra. During the modification, we also refer to many reading materials, which are also recommended for you:

- ENGG 5781 Course Notes by Prof. Wing-Kin (Ken) Ma, CUHK, Hongkong, China, http://www.ee.cuhk.edu.hk/∼wkma/engg5781
- Roger A. Horn and Charles R. Johnson, Matrix Analysis (Second Edition), Cambridge University Press, 2012.
- S. Boyd and L. Vandenberghe, Introduction to Applied Linear Algebra (Vectors, Matrices, and Least Squares), Cambridge University Press, 2018.

The whole book can cover a semester course in a 14week, each section in which corresponds to a 2-hour lecture. If you read the whole book, and work some mini-exercises, you will learn a lot. We hope you will get the insights on linear algebra and apply them in your own subject.

CUHK(SZ)
August 29, 2018

Acknowledgments

This book is from the MAT2040 in summer semester, 2017. It is revised in 2018 to correct some mistakes, and revise some proofs to give readers better insights on linear algebra.

CUHK(SZ)

Notations and Conventions

$\mathbb{R}^n$	n-dimensional real space
$\mathbb{C}^n$	n-dimensional complex space
$\mathbb{R}^{m\times n}$	set of all $m \times n$ real-valued matrices
$\mathbb{C}^{m\times n}$	set of all $m \times n$ complex-valued matrices
x_i	ith entry of column vector $\boldsymbol{x}$
a_{ij}	(i,j)th entry of matrix $\boldsymbol{A}$
$\boldsymbol{a}_i$	ith column of matrix $\boldsymbol{A}$
$\boldsymbol{a}_i^{\mathrm{T}}$	ith row of matrix $\boldsymbol{A}$
$\mathbb{S}^n$	set of all $n \times n$ real symmetric matrices, i.e., $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $a_{ij} = a_{ji}$ for all i,j
$\mathbb{H}^n$	set of all $n \times n$ complex Hermitian matrices, i.e., $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\bar{a}_{ij} = a_{ji}$ for all i,j
$\boldsymbol{A}^{\mathrm{T}}$	transpose of $\boldsymbol{A}$, i.e, $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{T}}$ means $b_{ji} = a_{ij}$ for all i,j
$\boldsymbol{A}^{\mathrm{H}}$	Hermitian transpose of $\boldsymbol{A}$, i.e, $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{H}}$ means $b_{ji} = \bar{a}_{ij}$ for all i,j
$\mathrm{trace}(\boldsymbol{A})$	sum of diagonal entries of square matrix $\boldsymbol{A}$
$\mathbf{1}$	A vector with all 1 entries
$\mathbf{0}$	either a vector of all zeros, or a matrix of all zeros
$\boldsymbol{e}_i$	a unit vector with the nonzero element at the ith entry
$\mathcal{C}(\boldsymbol{A})$	the column space of $\boldsymbol{A}$
$\mathcal{R}(\boldsymbol{A})$	the row space of $\boldsymbol{A}$
$\mathcal{N}(\boldsymbol{A})$	the null space of $\boldsymbol{A}$
$\mathrm{Proj}_{\mathcal{M}}(\boldsymbol{A})$	the projection of $\boldsymbol{A}$ onto the set $\mathcal{M}$

8.2. Thursday

Three ways for matrix decomposition are significant in linear alegbra:

$$\begin{cases} \text{LU (from } \textbf{Gaussian elimination}) \\ \text{QR (from } \textbf{Orthogonalization}) \\ \text{SVD (from } \textbf{eigenvalues and eigenvectors}) \end{cases}$$

We have learnt the first two decomposition. And the third way is increasingly significant in the information age.

In the last lecture we learnt that any real symmetric matrix adimits *diagonalization*, i.e., *eigendecomposition*. However, can we get some **universal** decomposition, i.e., Is there any decomposition that can be applied to all matrices?

The anwer is yes. The key idea behind is to do *symmetrization*. We have to consider $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

8.2.1. SVD: Singular Value Decomposition

Theorem 8.3 — SVD. Given any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, there exists a 3-tuple $(\boldsymbol{U}, \Sigma, \boldsymbol{V}) \in \mathbb{R}^{m\times m} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times n}$ such that

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}},$$

where $\boldsymbol{U}, \boldsymbol{V}$ are **orthogonal**, and Σ takes the form

$$\Sigma_{ij} = \begin{cases} \sigma_i, & i = j \\ 0, & i \neq j \end{cases},$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ and with $p = \min\{m, n\}$.

R

- If $\boldsymbol{V} = \boldsymbol{U}$, this decomposition is exactly **eigen-decomposition**.

- Specifically speaking,
  - $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ such that its columns are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$
  - $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ such that its columns are eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$
  - $\Sigma \in \mathbb{R}^{m\times n}$ looks like a diagonal matrix, i.e., it has the form

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \text{ if } m \geq n$$
$$\Sigma = \begin{pmatrix} \sigma_1 & & & 0 & \dots & 0 \\ & \ddots & & \vdots & \ddots & \vdots \\ & & \sigma_m & 0 & \dots & 0 \end{pmatrix} \text{ if } m < n.$$

with $\sigma_i = \sqrt{\lambda_i}$ for $i = 1,2,\dots,\min\{m,n\}$, where *$\lambda_i$'s are eigenvalues of* $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ *(if $m < n$) or* $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$. *(if $m \geq n$)*)

Definition 8.7 [SVD] The above decomposition is called the **singular value decomposition (SVD)**

- σ_i is called the ith **singular value**
- The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are called the ith **left and right singular vectors**, respectively.
- $(\sigma_i, \boldsymbol{u}_i)$ are the eigen-pairs of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$; $(\sigma_i, \boldsymbol{v}_i)$ are the eigen-pairs of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ for $i = 1,2,\dots,\min\{m,n\}$.
- The following notations may be used to denote the singular values of $\boldsymbol{A}$:

$$\sigma_{\max}(\boldsymbol{A}) \geq \sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_p(\boldsymbol{A}) = \sigma_{\min}(\boldsymbol{A})$$

The proof for the SVD decomposition is constructive. To see the insights of the proof, let's study the case $m = n$ first, then we extend the proof for general case:

Proposition 8.2 SVD always exists for any **real square nonsingular** matrix.

Proof. For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, you may verify that $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ is PD, thus it admits the eigen-decomposition:

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{U}\Sigma\boldsymbol{U}^{\mathrm{T}}, \quad \text{with } \lambda_1 \geq \cdots \geq \lambda_n > 0. \tag{8.8}$$

We define $\Sigma := \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_m})$ and $\boldsymbol{V} := \boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}\Sigma^{-1}$.

You may verify that $\boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{A}$ and $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}$, i.e., $\boldsymbol{V}$ is orthogonal. The proof is complete. ■

Proposition 8.3 SVD always exists for any **real** matrix.

Proof.

- Firstly, consider the matrix product $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$. Since $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} \in \mathbb{S}^m$ and $\boldsymbol{A} \succeq 0$, we can decompose $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ as

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{U}\Sigma\boldsymbol{U}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^{\mathrm{T}} & \boldsymbol{U}_2^{\mathrm{T}} \end{bmatrix} = \boldsymbol{U}_1\tilde{\Sigma}\boldsymbol{U}_1^{\mathrm{T}} \tag{8.9}$$

  where:

  - we assume that the eigenvalues are ordered, i.e.,

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \quad \text{and} \quad \lambda_{r+1} = \cdots = \lambda_p = 0$$

    with r being the number of nonzero eigenvalues
  - $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ denotes an orthogonal matrix, and its columns are the corresponding eigenvectors
  - We partition $\boldsymbol{U}$ as

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix}, \quad \boldsymbol{U}_1 \in \mathbb{R}^{m \times r}, \boldsymbol{U}_2 \in \mathbb{R}^{m \times (m-r)},$$

    and $\tilde{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$.

- Secondly, we show that

$$\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A} = 0 \tag{8.10}$$

Since $\boldsymbol{U}$ is orthogonal, we obtain:

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1^{\mathrm{T}} \\ \boldsymbol{U}_2^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1 & \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_2 \\ \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_1 & \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2 \end{bmatrix} = \boldsymbol{I} \implies \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_1 = \mathbf{0}.$$

Substituting Eq.(8.9) into $\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A})^{\mathrm{T}}$, we obtain:

$$\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A})^{\mathrm{T}} = (\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_1)\tilde{\Sigma}\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_2 = \mathbf{0} \tag{8.11}$$

By Eq.(8.11) and the simple result that $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} = \mathbf{0}$ implies $\boldsymbol{B} = \mathbf{0}$ (write $\boldsymbol{B}$ into column vectors form to verify it), we conclude that $\boldsymbol{U}_2\boldsymbol{A} = \mathbf{0}$

- Thirdly, we construct the following matrices:

$$\widehat{\Sigma} = \tilde{\Sigma}^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}), \quad \boldsymbol{V}_1 = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}_1\widehat{\Sigma}^{-1} \in \mathbb{R}^{n\times r}.$$

Combining it with Eq.(8.9), we can verify that $\boldsymbol{V}_1^{\mathrm{T}}\boldsymbol{V}_1 = \boldsymbol{I}$. Furthermore, there exists a matrix $\boldsymbol{V}_2 \in \mathbb{R}^{n\times(n-r)}$ such that $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix}$ is orthogonal. Moreover, we can verify that

$$\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_1 = \widehat{\Sigma}, \quad \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_2 = \mathbf{0} \tag{8.12}$$

- Fourthly, consider the matrix product $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}$. From Eq.(8.12) and Eq.(8.10), we have

$$\begin{aligned} \boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V} &= \begin{bmatrix} \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_1 & \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_2 \\ \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_1 & \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_2 \end{bmatrix} \\ &= \begin{bmatrix} \widehat{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} := \Sigma \end{aligned}$$

By multiplying the above equation on the left by $\boldsymbol{U}$ and on the right by $\boldsymbol{V}^{\mathrm{T}}$, we obtain the desired result $A = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}$. The proof is complete.

■

8.2.2. Remark on SVD decomposition

8.2.2.1. Remark 1: Different Ways of Writing out SVD

Definition 8.8 [Paritioned form of SVD] let r be the number of nonzero singular values, and note that $\sigma_1 \geq \cdots \geq \sigma_r > 0, \sigma_{r+1} = \cdots = \sigma_p = 0$. Then from the standard form, we derive the **partitioned form of SVD**:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^{\mathrm{T}} \\ \boldsymbol{V}_2^{\mathrm{T}} \end{bmatrix} \tag{8.13}$$

where:

- $\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$
- $\boldsymbol{U}_1 = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r \end{bmatrix} \in \mathbb{R}^{m \times r}, \boldsymbol{U}_2 = \begin{bmatrix} \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_m \end{bmatrix} \in \mathbb{R}^{m \times (m-r)}$
- $\boldsymbol{V}_1 = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_r \end{bmatrix} \in \mathbb{R}^{n \times r}, \boldsymbol{V}_2 = \begin{bmatrix} \boldsymbol{v}_{r+1} & \cdots & \boldsymbol{v}_n \end{bmatrix} \in \mathbb{R}^{n \times (n-r)}$

Note that $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1, \boldsymbol{V}_2$ are semi-orthogonal, i.e., they all have orthonormal columns. ■

Definition 8.9 [Thin SVD] We can re-write Eq.(8.13) as the **thin form of SVD**:

$$\boldsymbol{A} = \boldsymbol{U}_1 \tilde{\Sigma} \boldsymbol{V}_1^{\mathrm{T}} \tag{8.14}$$

■

Definition 8.10 [Outer-product form] By expanding the Eq.(8.14), we derive the **outer-product form of SVD**:

$$\boldsymbol{A} = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} \tag{8.15}$$

8.2.2.2. Remark 2: SVD and Eigen-decomposition

The eigenvalues for $A^{\mathrm{T}}A$ and AA^{T} are the same for first p terms.

Proposition 8.4 Suppose A admits the SVD $A = U\Sigma V^{\mathrm{T}}$, then we have:

$$AA^{\mathrm{T}} = UD_1U^{\mathrm{T}}, \qquad D_1 = \Sigma\Sigma^{\mathrm{T}} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2, \underbrace{0, \dots, 0}_{m-p \text{ zeros}}) \tag{8.16}$$

$$A^{\mathrm{T}}A = VD_2V^{\mathrm{T}}, \qquad D_2 = \Sigma^{\mathrm{T}}\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2, \underbrace{0, \dots, 0}_{n-p \text{ zeros}}) \tag{8.17}$$

Proof. Just apply the SVD form and the orthogonality of U and V. ■

8.2.2.3. Remark 3: SVD and Subspace

We are curious about how many singular values of A are nonzero.

Proposition 8.5 The following properties hold:

1. $\mathcal{C}(A) = \mathcal{C}(U_1), \mathcal{C}(A)^{\perp} = \mathcal{C}(U_2)$;
2. $\mathcal{C}(A^{\mathrm{T}}) = \mathcal{C}(V_1), \mathcal{C}(A^{\mathrm{T}})^{\perp} = \mathcal{N}(A) = \mathcal{C}(V_2)$;
3. $\operatorname{rank}(A) = r$, i.e., the number of nonzero singular values.

Proof. The above properties are easily seen to be true using SVD. Also, you should apply the definition for column space and null space. You should verify these properties by yourself.

$$\mathcal{C}(A) = \{y \in \mathbb{R}^m \mid y = Ax, x \in \mathbb{R}^n\} \tag{8.18a}$$

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \tag{8.18b}$$

For the third part of proposition(8.5), since $\operatorname{rank}(A) = \dim(\mathcal{C}(A)) = \dim(\mathcal{C}(U_1))$, and U_1 has r orthonormal columns, we derive that $\dim(\mathcal{C}(U_1)) = r = \operatorname{rank}(A)$. ■

For the SVD decomposition

$$A = U\Sigma V^{\mathrm{T}},$$

we can convert it into the following two forms:

$$AV = U\Sigma V^{\mathrm{T}} V = U\Sigma$$

$$A = U\Sigma V^{\mathrm{T}} \implies A^{\mathrm{T}} = V\Sigma U^{\mathrm{T}} \implies A^{\mathrm{T}} U = V\Sigma U^{\mathrm{T}} U = V\Sigma.$$

If we write it into vector forms, we obtain:

$$\begin{cases} A\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j \\ A^{\mathrm{T}} \boldsymbol{u}_j = \sigma_j \boldsymbol{v}_j \end{cases}, \quad j = 1, 2, \ldots, r. \tag{8.19}$$

The columns of U ($\boldsymbol{u}_j$) are called the **left singular vector** of A; the columns of V ($\boldsymbol{v}_j$) are called the **right singular vector** of A; σ_j is called the **singular value**.

We can easily understand the proposition(8.5) and Eq.(8.19) by the following graph:

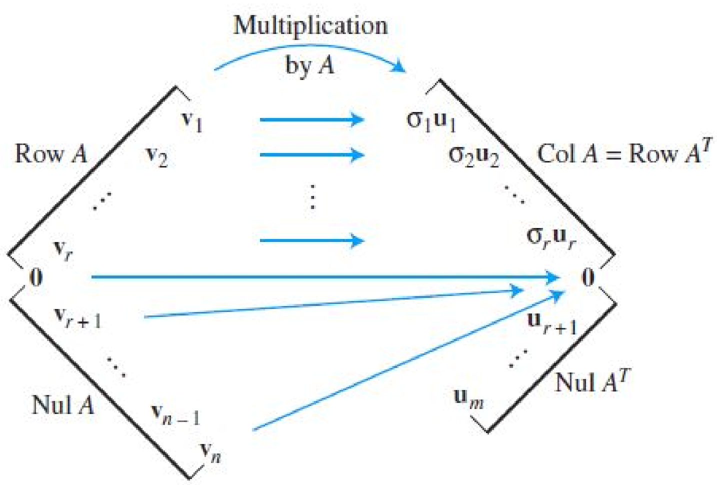


Figure 8.2: The fundamental spaces and the action of A.

Explanation:

- When $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$ are multiplied by A, they are converted into $\{\sigma_1\boldsymbol{u}_1, \ldots, \sigma_r\boldsymbol{u}_r\}$; when $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$ are multiplied by A, they are converted into $\mathbf{0}$.

- The first r columns of V forms the basis for the row space of A, i.e., $\mathcal{C}(V_1) = \mathcal{C}(A^{\mathrm{T}})$.

- The last $n - r$ columns of V forms the basis for the null space of A, i.e., $\mathcal{C}(V_2) =$

$\mathcal{N}(\boldsymbol{A})$.

- The first r columns of $\boldsymbol{U}$ forms the basis for the column space of $\boldsymbol{A}$, i.e., $\mathcal{C}(\boldsymbol{U}_1) = \mathcal{C}(\boldsymbol{A})$.
- The last $m-r$ columns of $\boldsymbol{U}$ forms the basis for the null space of $\boldsymbol{A}^{\mathrm{T}}$, i.e., $\mathcal{C}(\boldsymbol{U}_2) = \mathcal{N}(\boldsymbol{A}^{\mathrm{T}})$

Recall the outer-product form of SVD,

$$\boldsymbol{A} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^{\mathrm{T}}$$

where $r = \operatorname{rank}(\boldsymbol{A}) =$ number of nonzero singular values, which is the third meaning for the rank:

R Up till now, $\operatorname{rank}(\boldsymbol{A})$ has three meanings:

- $\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{row}(\boldsymbol{A}))$
- $\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{col}(\boldsymbol{A}))$
- $\operatorname{rank}(\boldsymbol{A}) =$ number of nonzero singular values of $\boldsymbol{A}$.

R However, $\operatorname{rank}(\boldsymbol{A}) \neq$ number of nonzero eigenvalues. Let me raise a counterexample:

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

then eigenvalues are $\lambda_1 = \lambda_2 = 0$, and $\operatorname{rank}(\boldsymbol{A}) = 1$.

R Also, note that many properties can be easily proved by **thin** or **outer-product** form of SVD. For example, $\operatorname{rank}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A})$. If you have no ideas of a proof in exam, you may try SVD.

8.2.2.4. Compact SVD

Due to the outer-product form of SVD, i.e., any matrix with rank r can be factorized into

$$
\begin{aligned}
\boldsymbol{A} &= \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}} \\
&= \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_r \end{bmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_r^{\mathrm{T}} \end{bmatrix},
\end{aligned}
$$

we obtain the following corollary:

Corollary 8.2 Every rank r matrix can be written as the sum of r rank 1 matrices. Moreover, these matrices could be perpendicular!

What's the meaning of perpendicular?

Definition 8.11 [perpendicular for matrix] For two real $n \times n$ matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, they are said to be **perpendicular** (**orthogonal**) if the inner product between $\boldsymbol{A}$ and $\boldsymbol{B}$ is zero:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{trace}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}) = \sum_{i,j=1}^{n} A_{ij}B_{ij} = 0.$$

Decompose $\boldsymbol{A} := \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. If we set $\boldsymbol{A}_i = \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} \sigma_i$, let's show $\boldsymbol{A}_i$'s are perpendicular:

$$
\begin{aligned}
\langle \boldsymbol{A}_i, \boldsymbol{A}_j \rangle &= \operatorname{trace}(\boldsymbol{A}_j^{\mathrm{T}} \boldsymbol{A}_i) \\
&= \operatorname{trace}(\sigma_i\sigma_j \boldsymbol{v}_j \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}) = \sigma_i\sigma_j \operatorname{trace}(\boldsymbol{v}_j \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}) \\
&= \sigma_i\sigma_j \operatorname{trace}(\boldsymbol{v}_j (\boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{u}_i) \boldsymbol{v}_i^{\mathrm{T}}) = \sigma_i\sigma_j \operatorname{trace}(\boldsymbol{v}_j \boldsymbol{0} \boldsymbol{v}_i^{\mathrm{T}}) \\
&= 0.
\end{aligned}
$$

How many rank 1 matrices do we need to pick to construct matrix $\boldsymbol{A}$? In fact, this number has no upper bound. For example, if we obtain

$$\boldsymbol{A} = \boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}$$

Then we can always decompose any rank 1 matrix into 2 rank 1 matrices:

$$\boldsymbol{A} = \boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}} + \frac{1}{2}\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}} + \frac{1}{2}\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}.$$

But this number has a lower bound, that is rank. In other words, rank$(\boldsymbol{A})$ = smallest number of rank 1 matrices with sum $\boldsymbol{A}$.

8.2.3. Best Low-Rank Approximation

Given matrix $\boldsymbol{A}$. What is the *best rank k approximation*? In other words, given matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, what is the optimal solution for the optimization:

$$\begin{aligned}
\min_{\boldsymbol{Z}} \quad & \|\boldsymbol{A} - \boldsymbol{Z}\|_F^2 \\
\text{s.t.} \quad & \operatorname{rank}(\boldsymbol{Z}) = k \\
& \boldsymbol{Z} \in \mathbb{R}^{m\times n}
\end{aligned}$$

Firstly let's introduce the definition for Frobenius norm:

Definition 8.12 [Frobenius norm] The Frobenius norm for $m \times n$ matrix $\boldsymbol{A}$ is given by

$$\|\boldsymbol{A}\|_F = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A}\rangle} = \sqrt{\operatorname{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}.$$

Theorem 8.4 Suppose the SVD for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is given by

$$\boldsymbol{A} = \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}} + \cdots + \sigma_r\boldsymbol{u}_r\boldsymbol{v}_r^{\mathrm{T}}.$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

Then the best rank $k(k \leq r)$ approximation of $\boldsymbol{A}$ is

$$\boldsymbol{A}_k = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \cdots + \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^{\mathrm{T}}.$$

For example, $\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}}$ is the best rank 1 approximation of $\boldsymbol{A}$.

8.2.3.1. Analogy with least square problem

For least squares problem, the key is to do approximation for $\boldsymbol{b} \in \mathbb{R}^m$. In other words, we just do a projection from $\boldsymbol{b}$ to the plane $\{\boldsymbol{A}\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^n\}$:

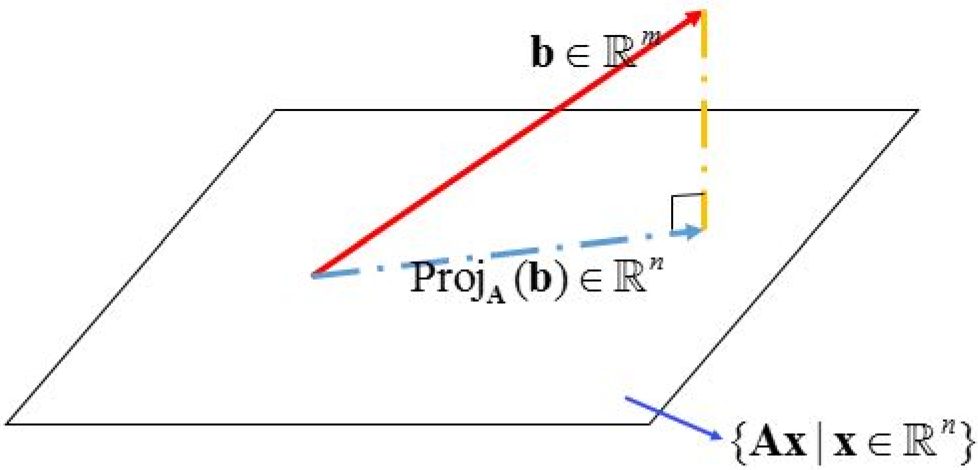


Figure 8.3: Least square problem: find $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \mathrm{Proj}_{\mathcal{C}(\boldsymbol{A})}(\boldsymbol{b})$.

R For the least squares problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, the key is to do the projection of $\boldsymbol{b}$ onto $\mathcal{C}(\boldsymbol{A})$, thus it suffices to solve the **equality**

$$\boldsymbol{A}\boldsymbol{x} = \mathrm{Proj}_{\mathcal{C}(\boldsymbol{A})}(\boldsymbol{b}).$$

Similarly, the beast rank k approximation could be viewed as a projction from $\boldsymbol{A}$ with rank r to the "plane" that contains all rank k matrices:

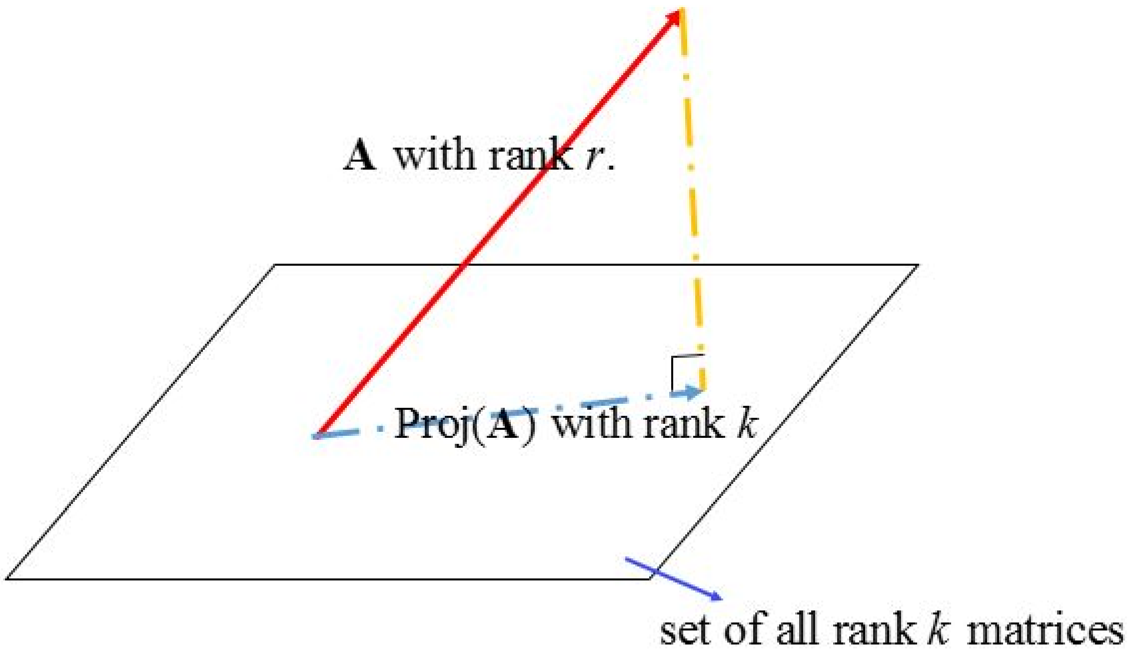


Figure 8.4: Best rank k approximation: find the projection from matrix $\boldsymbol{A}$ with rank r onto the plane that contains all rank k matrices

R Similarly, for the best rank k approximation problem

$$\begin{aligned} \min_{\boldsymbol{Z}} \quad & \|\boldsymbol{A}-\boldsymbol{Z}\|_F^2 \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{Z}) = k \\ & \boldsymbol{Z} \in \mathbb{R}^{m\times n} \end{aligned}$$

with $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the key is to do the projection of $\boldsymbol{A}$ onto the set $\mathcal{M} = \{\boldsymbol{M} \in \mathbb{R}^{m\times n} \mid \operatorname{rank}(\boldsymbol{M}) = k\}$, thus it suffices to solve the **equality**

$$\boldsymbol{Z} = \operatorname{Proj}_{\mathcal{M}}(\boldsymbol{A}).$$

For some non-convex optimization problems, this idea is very useful. The further reading is recommended:

Jain, Prateek, and P. Kar. "Non-convex Optimization for Machine Learning." Foundations & Trends® in Machine Learning 10.3-4(2017):142-336.