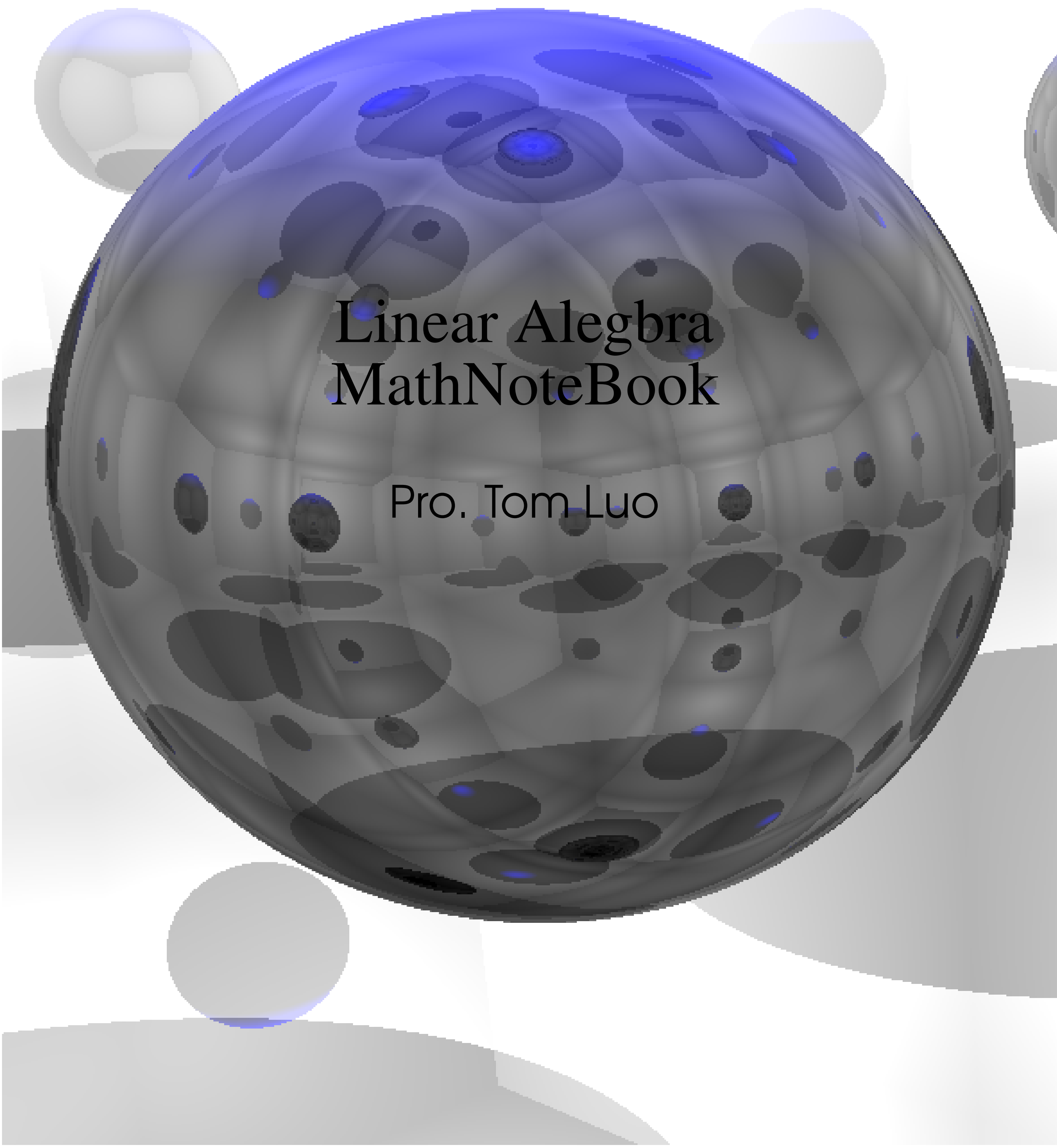
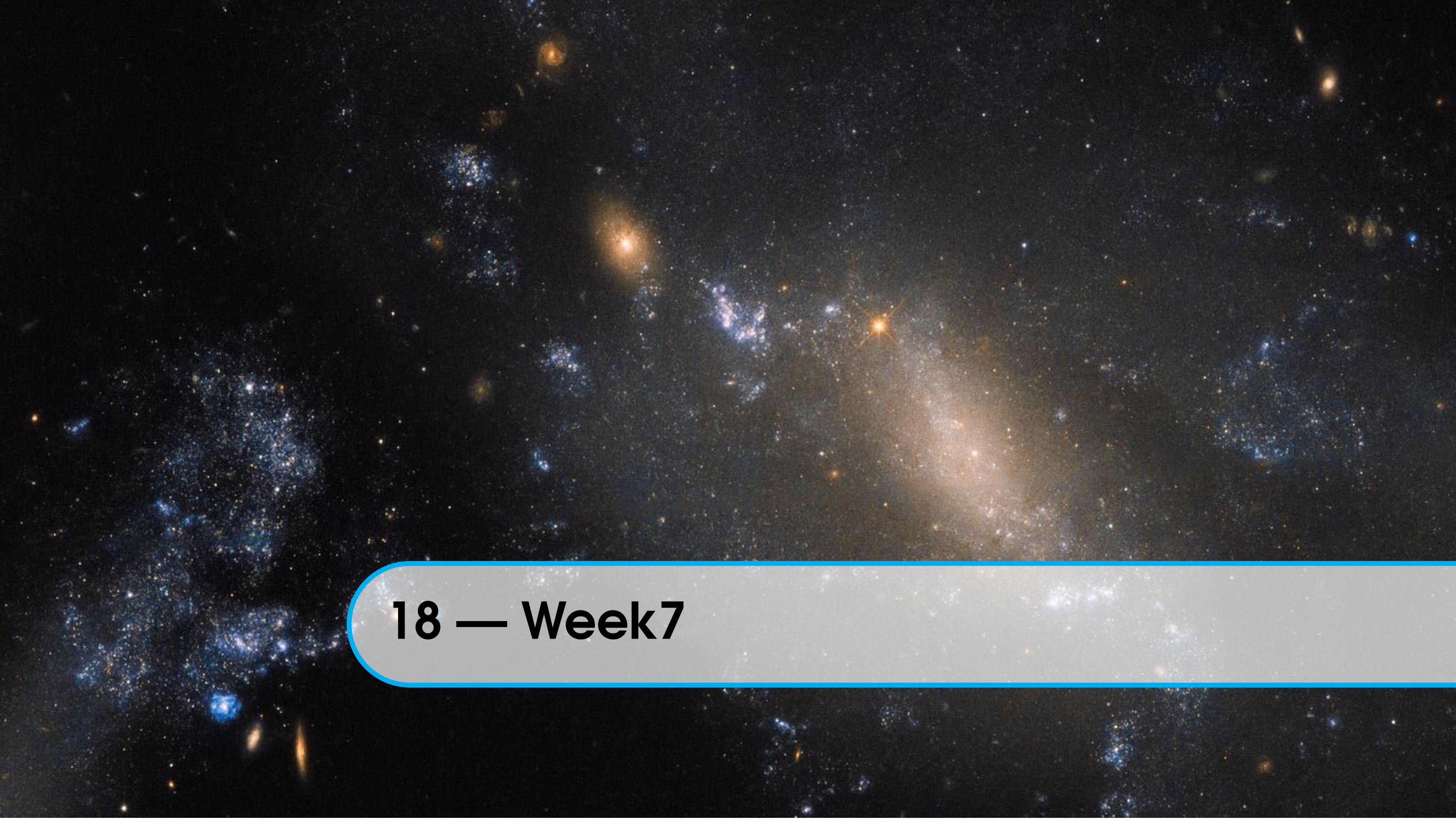


Linear Alegbra MathNoteBook

Pro. Tom Luo



18 — Week7

18.1 Thursday

Three ways for matrix decomposition are significant in linear alegbra:

$$\begin{cases} \text{LU (from } \textbf{elimination}\text{)} \\ \text{QR (from } \textbf{orthogonalization}\text{)} \\ \text{SVD (from } \textbf{eigenvectors}\text{)} \end{cases}$$

We have learnt the first two decomposition. And the third way is increasingly significant in the information age.
In the last lecture we talk about *eigendecomposition* for **real symmetric** matrices and *diagonalization*. However, can we get some **universal** decomposition? Is there any decomposition that can be applied to all matrices?
The anwer is yes. The key idea is to do *symmetrization*, we have to consider $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

18.1.1 SVD: Singular Value Decomposition

Any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ could be factorized into

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

where $\boldsymbol{U}$ is a $m\times m$ **orthogonal** matrix, $\boldsymbol{\Sigma}$ is a $m\times n$ "*diagonal*" (we will define it later) matrix, $\boldsymbol{V}$ is a $n\times n$ **orthogonal** matrix.
If $\boldsymbol{V}$=$\boldsymbol{U}$ (then consequently $m = n$), then this is exactly *eigendecomposition*.
Specifically speaking,
$\boldsymbol{U}$ is $m\times m$ matrix s.t. *columns are eigenvectors of* $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.
$\boldsymbol{V}$ is $n\times n$ matrix s.t. *columns are eigenvectors of* $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

$\boldsymbol{\Sigma}$ is $m \times n$ matrix which has the form:

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

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

Theorem 18.1 SVD always exists for any **real** matrix.

Proof. For any $m \times n$ matrix $\boldsymbol{A}$, WLOG, we set $m \geq n$.

- Firstly, we consider the case that all $\lambda_j \neq 0$ for $j = 1, \dots, n$. (λ_j's are eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.) Since $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is *real symmetric*, we do the eigendecomposition:

  $$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$$

  where $\boldsymbol{V}$ is *orthonormal* matrix and $\boldsymbol{D}$ is *diagonal* matrix.
  Also, the eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ are orthogonal (note that in proposition (16.3) we claim that the eigenvectors of diagonalizable matrix are orthogonal.):

  $$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j \text{ for } j = 1, \dots, n$$

  where $\boldsymbol{v}_j$'s are eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ s.t. they form orthonormal basis of $\mathbb{R}^n$.
  Note that given any matrix $\boldsymbol{A}$, $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix}$ is immediately defined. (This is because $\boldsymbol{v}_j$'s are eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.)
  If we want to show $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, since $\boldsymbol{A}$ and $\boldsymbol{V}$ is defined, we only need to show there exists *special* $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ such that

  $$\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{A}(\boldsymbol{V}^{\mathrm{T}})^{-1} = \boldsymbol{A}\boldsymbol{V}.$$

  - First step, we construct such $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$:
    Since λ_j's are eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ associated with eigenvectors $\boldsymbol{v}_j$, we obtain:

    $$\|\boldsymbol{A}\boldsymbol{v}_j\|^2 = \boldsymbol{v}_j^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_j) = \boldsymbol{v}_j^{\mathrm{T}}(\lambda_j \boldsymbol{v}_j) = \lambda_j(\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{v}_j) = \lambda_j\|\boldsymbol{v}_j\|^2.$$

    Hence $\lambda_j = \frac{\|\boldsymbol{A}\boldsymbol{v}_j\|^2}{\|\boldsymbol{v}_j\|^2} > 0$. (As we assume $\lambda_j \neq 0$, this is strictly inequality.)
    Hence we define $\boldsymbol{u}_j := \boldsymbol{A}\boldsymbol{v}_j \frac{1}{\sqrt{\lambda_j}} \in \mathbb{R}^{m \times 1}$. for $j = 1, \dots, n$.
    And then we construct $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$:

    $$\boldsymbol{U} := \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_n \end{bmatrix} \in \mathbb{R}^{m \times n}.$$
    $$\boldsymbol{\Sigma} := \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}) \in \mathbb{R}^{n \times n}.$$

    It's easy to verify that $\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{V}$.

- Next step, we show that $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n\}$ is orthonormal set:
  For any $\boldsymbol{u}_i,\boldsymbol{u}_j$, we have

$$\begin{aligned}\langle \boldsymbol{u}_i,\boldsymbol{u}_j\rangle &= \frac{1}{\sqrt{\lambda_i}\sqrt{\lambda_j}}\langle \boldsymbol{A}\boldsymbol{v}_i,\boldsymbol{A}\boldsymbol{v}_j\rangle \\ &= \frac{1}{\sqrt{\lambda_i}\sqrt{\lambda_j}}\langle \boldsymbol{v}_i,\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_j\rangle \qquad \text{Due to the useful formula } \langle \boldsymbol{A}\boldsymbol{x},\boldsymbol{y}\rangle = \langle \boldsymbol{x},\boldsymbol{A}^{\mathrm{H}}\boldsymbol{y}\rangle. \\ &= \frac{1}{\sqrt{\lambda_i}\sqrt{\lambda_j}}\langle \boldsymbol{v}_i,\lambda_j\boldsymbol{v}_j\rangle = \sqrt{\frac{\lambda_j}{\lambda_i}}\langle \boldsymbol{v}_i,\boldsymbol{v}_j\rangle\end{aligned}$$

  Since $\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_n\}$ are orthonormal, we obtain

$$\langle \boldsymbol{v}_i,\boldsymbol{v}_j\rangle = \begin{cases}1, & i=j\\ 0, & i\ne j\end{cases} \implies \langle \boldsymbol{u}_i,\boldsymbol{u}_j\rangle = \begin{cases}1, & i=j\\ 0, & i\ne j\end{cases}$$

  Hence $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n\}$ are orthonormal.
- Then we show that $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n\}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$:
  For $j=1,\dots,n$, we obtain:

$$\begin{aligned}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}_j &= \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_j\frac{1}{\sqrt{\lambda_j}} \qquad \text{by definition of } \boldsymbol{u}_j. \\ &= \boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_j)\frac{1}{\sqrt{\lambda_j}} = \boldsymbol{A}\lambda_j\boldsymbol{v}_j\frac{1}{\sqrt{\lambda_j}} \\ &= \sqrt{\lambda_j}\boldsymbol{A}\boldsymbol{v}_j = \lambda_j\times(\frac{1}{\sqrt{\lambda_j}}\boldsymbol{A}\boldsymbol{v}_j) \\ &= \lambda_j\boldsymbol{u}_j.\end{aligned}$$

- We notice that in SVD $\boldsymbol{U}$ is m by m matrix, $\boldsymbol{\Sigma}$ is m by n matrix. Hence we need to *reconstruct* our $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ in step 1:
  Since $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n\}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$, and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ has m orthogonal eigenvectors, so we pick $\boldsymbol{u}_{n+1},\dots,\boldsymbol{u}_m$ s.t. $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n,\boldsymbol{u}_{n+1},\dots,\boldsymbol{u}_m\}$ are m orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.
  Then we let

$$\boldsymbol{U} := \begin{bmatrix}\boldsymbol{u}_1 & \dots & \boldsymbol{u}_m\end{bmatrix} \in \mathbb{R}^{m\times m}.$$

$$\boldsymbol{\Sigma} := \begin{pmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0\end{pmatrix} \in \mathbb{R}^{m\times n}.$$

  It's easy to verify that

$$\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{V}$$

  Hence finally we obtain

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

- For there exists some $\lambda_j = 0$ case, we discuss it in next section.

■

18.1.2 Remark on SVD decomposition

Remark 1

The eigenvalues for $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ are not always nonzero.

Proposition 18.1 For $m \times n$ matrix $\boldsymbol{A}$, suppose $\operatorname{rank}(\boldsymbol{A}) = r$, then all eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ are given by

$$\begin{aligned}\operatorname{eig}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) &= \{\lambda_1, \dots, \lambda_r, \overbrace{0, \dots\dots, 0}^{\text{Totally } n-r \text{ terms}}\} \\ \operatorname{eig}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}) &= \{\lambda_1, \dots, \lambda_r, \underbrace{0, \dots\dots, 0}_{\text{Totally } m-r \text{ terms}}\}.\end{aligned}$$

Proof.
- Firstly we prove that the nonzero eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ are exactly the same (counted with multiplicity):
We only need to show $\frac{\det(\lambda \boldsymbol{I} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}{\lambda^{n-r}} = \frac{\det(\lambda \boldsymbol{I} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}{\lambda^{m-r}}$.
And we find that

$$\begin{aligned}\det(\lambda \boldsymbol{I} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) &= \lambda^n \det(\boldsymbol{I} - \lambda^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) \\ &= \lambda^n \det(\boldsymbol{I} - \lambda^{-1}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}) \qquad \text{Due to Sylvester's determinant identity} \\ &\qquad\qquad \det(\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{I}_n + \boldsymbol{B}\boldsymbol{A}) \\ &\qquad\qquad \text{for } \boldsymbol{A} \in \mathbb{R}^{m\times n} \text{ and } \boldsymbol{B} \in \mathbb{R}^{n\times m}.\end{aligned}$$

Hence we obtain

$$\begin{aligned}\frac{\det(\lambda \boldsymbol{I} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}{\lambda^{n-r}} &= \frac{\lambda^n \det(\boldsymbol{I} - \lambda^{-1}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})}{\lambda^{n-r}} \\ &= \frac{\lambda^m \det(\boldsymbol{I} - \lambda^{-1}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})}{\lambda^{m-r}} \\ &= \frac{\det(\lambda \boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})}{\lambda^{m-r}}\end{aligned}$$

- Secondly we show the eigenvalues for $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ have exactly $(n-r)$ zeros; the eigenvalues for $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ have exactly $(m-r)$ zeros.
Assume there are n ind. eigenvectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ for $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ corresponding to their eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. Hence we have

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i \text{ for } i = 1, \dots, n.$$

Since $\operatorname{rank}(\boldsymbol{A}) = r = \operatorname{rank}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$, the dimension of the eigenspace for $\lambda = 0$ is $n-r$. Hence among $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ there are $n-r$ ind. eigenvectors belong to the eigenspace for $\lambda = 0$.
Thus there are exactly $(n-r)$ zeros for eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.
- How to prove there are exactly $(m-r)$ zeros for eigenvalues of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$? We just need to obtain $\operatorname{rank}(\boldsymbol{A}^{\mathrm{T}}) = r = \operatorname{rank}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})$ and proceed similarly.

■

For SVD decomposition

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

we can convert it into the following two forms:

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

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

If we write it into vector forms, we obtain:

$$\begin{cases} \boldsymbol{A}\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}_j = \sigma_j \boldsymbol{v}_j \end{cases}$$

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

Remark 2: Four fundamental subspaces

The general SVD decomposition for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is given by

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_r & \dots & \boldsymbol{u}_m \end{bmatrix} \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_r^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_n^{\mathrm{T}} \end{pmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$$

For such $\boldsymbol{A}$, the matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ contain orthonormal basis for all four fundamental subspaces:

First r columns of $\boldsymbol{V}$: row space of $\boldsymbol{A}$.
last $n-r$ columns of $\boldsymbol{V}$: null space of $\boldsymbol{A}$.
First r columns of $\boldsymbol{U}$: column space of $\boldsymbol{A}$.
last $m-r$ columns of $\boldsymbol{U}$: null space of $\boldsymbol{A}^{\mathrm{T}}$.

Maybe it's easy to understand it in graph:

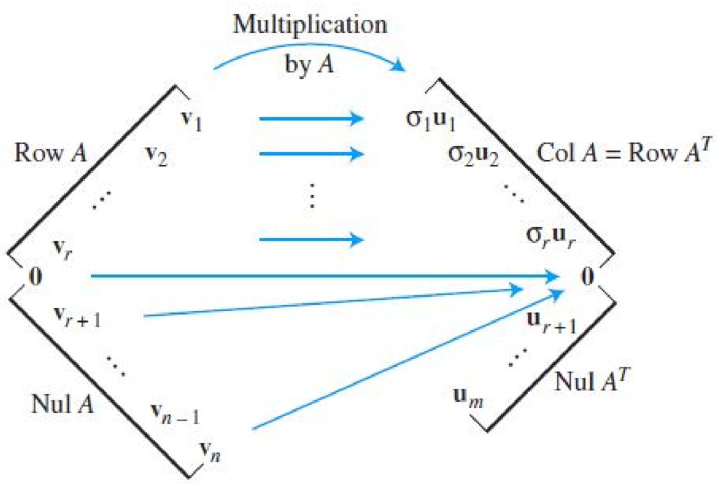


Figure 18.1: The fundamental spaces and the action of $\boldsymbol{A}$.

Remark 3: vector form

Recall we can write eigendecomposition in *vector form:*

$$A = \lambda_1 v_1 v_1^{\mathrm{T}} + \cdots + \lambda_n v_n v_n^{\mathrm{T}}$$

Also, we could write the **general** SVD decomposition in remark 2 into *vector form*:

$$A = \sigma u_1 v_1^{\mathrm{T}} + \cdots + \sigma u_r v_r^{\mathrm{T}}$$

where $r = \operatorname{rank}(A) =$ number of nonzero singular values. Here leads to the third meaning for the rank:

Proposition 18.2 The rank of $m \times n$ matrix A is the number of nonzero singular values.

Proof. We assume there are exactly s zero singular values of A, which means there are s zero eigenvalues of $A^{\mathrm{T}}A$ assocaiated with their s ind. eigenvectors. (Independence is due to the diagonalizable of $A^{\mathrm{T}}A$.) In other words, the eigenspace of $A^{\mathrm{T}}A$ for $\lambda = 0$ has dimension s.
The eigenspace of $A^{\mathrm{T}}A$ for $\lambda = 0$ is given by

$$\{x : A^{\mathrm{T}}Ax = 0\}.$$

Hence its dimension is given by

$$\operatorname{rank}(A^{\mathrm{T}}A) = \operatorname{rank}(A) := n - r.$$

Hence $s = n - r$. And obviously, the number of **nonzero** singular values is $n - s = r$. ■

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

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

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

Compact SVD

Hence any matrix with rank r can be factorized into

$$\begin{aligned} A &= U\Sigma V^{\mathrm{T}} \\ &= \begin{bmatrix} u_1 & \dots & u_r \end{bmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \begin{bmatrix} v_1^{\mathrm{T}} \\ \vdots \\ v_r^{\mathrm{T}} \end{bmatrix} \end{aligned}$$

where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{r\times n}$ are both orthogonal matrix. And $\Sigma = (\sigma_1, \dots, \sigma_r)$, where $\sigma_i > 0$ for $i = 1, 2, \dots, r$.

Corollary 18.1 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 18.1 — perpendicular for matrix. For two real $n \times n$ matrix A and B, if they are **perpendicular** (**orthogonal**), then the inner product between A and B is zero:

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

■

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:

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

So what is rank? 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 matrix:

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

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

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

- $\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{row}(\boldsymbol{A}))$
- $\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{col}(\boldsymbol{A}))$
- $\operatorname{rank}(\boldsymbol{A}) =$ smallest number of rank 1 matrices with sum $\boldsymbol{A}$.

18.1.3 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 model:

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

Firstly let's introduce the definition for Frobenius norm:

Definition 18.2 — 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 18.2 Suppose the SVD if $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is given by

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

and suppose $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.
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.

Analogy with least square problem
For least square 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 $\{\boldsymbol{Ax} | \boldsymbol{x} \in \mathbb{R}^n\}$:

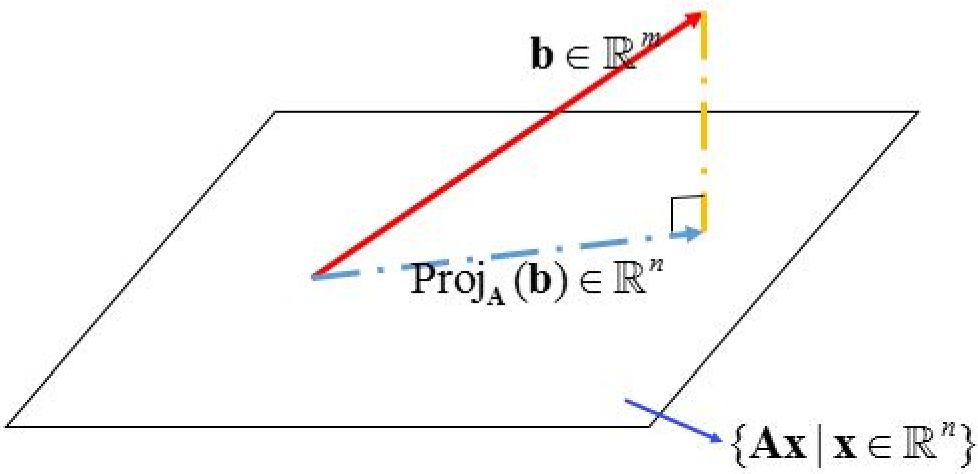


Figure 18.2: Least square problem: find $\boldsymbol{x}$ such that $\boldsymbol{Ax} = \mathrm{Proj}_{\boldsymbol{A}}(\boldsymbol{b})$.

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:

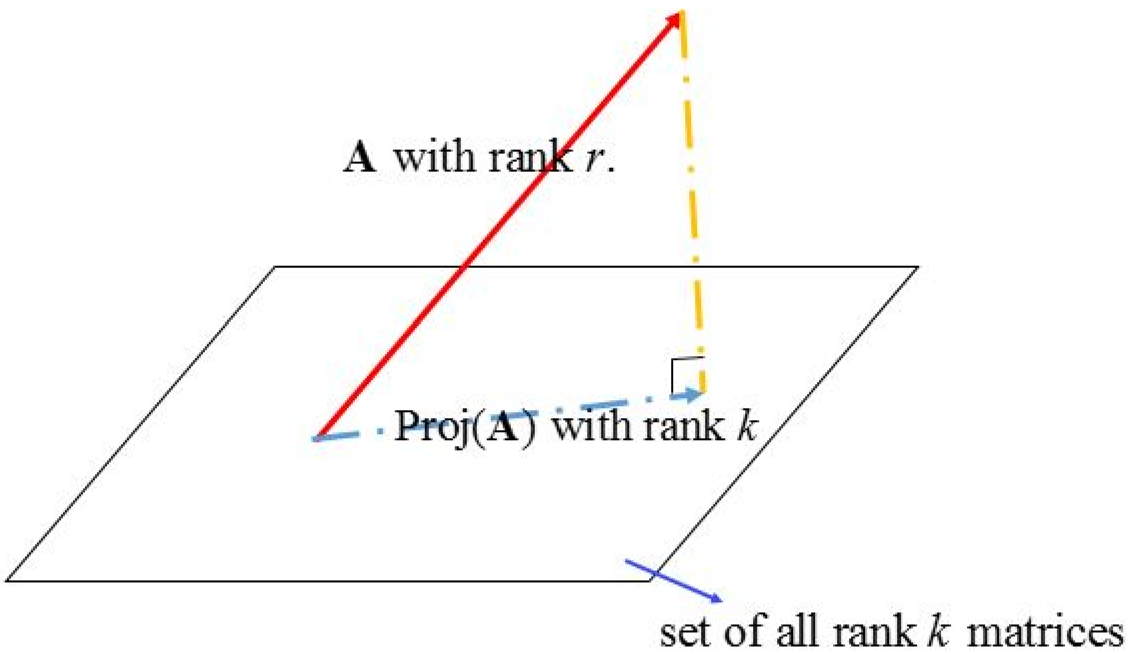


Figure 18.3: Best rank k approximation: find projection from rank r matrix to the plane that contains all rank k matrices