Mathematics for Machine Learning

— Linear Algebra: Singular Value Decomposition

&

Matrix Approximation

Joseph Chuang-Chieh Lin

Department of Computer Science & Engineering,
National Taiwan Ocean University

Fall 2025

Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213

Outline

1. Singular Value Decomposition (SVD)
   - Construction of the SVD
   - Example
2. Matrix Approximation

Outline

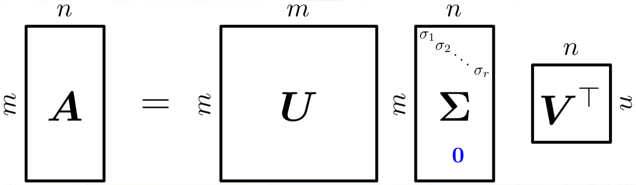
1. Singular Value Decomposition (SVD)
   - Construction of the SVD
   - Example
2. Matrix Approximation

Why Singular Value Decomposition?

- It can be applied to all matrices (not only to square matrices).
- It always exists.

Illustration

$\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\text{rank}(\boldsymbol{A}) = r \leq \min(m, n)$:

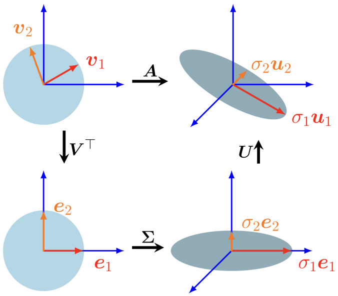


- $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ with orthogonal columns vectors $\boldsymbol{u}_i$, $i = 1, \ldots, m$.
- $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ with orthogonal columns vectors $\boldsymbol{v}_j$, $j = 1, \ldots, n$.
- $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0$ for $i \neq j$.
  - σ_i: singular values; $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$.
  - $\boldsymbol{u}_i$: left-singular vectors;
    $\boldsymbol{v}_j$: right-singular vectors;

Illustration & Example

$$\boldsymbol{A} = \begin{bmatrix} 1 & -0.8 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

$$= \begin{bmatrix} -0.79 & 0 & -0.62 \\ 0.38 & -0.78 & -0.49 \\ -0.48 & -0.62 & 0.62 \end{bmatrix} \begin{bmatrix} 1.62 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.78 & 0.62 \\ -0.62 & -0.78 \end{bmatrix}$$



Exercise

Exercise

Prove that for an $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same nonzero eigenvalues.

Outline

1. Singular Value Decomposition (SVD)
   - Construction of the SVD
   - Example
2. Matrix Approximation

SVD & Eigendecomposition

- Recall the eigendecomposition of a symmetric positive definite matrix

$$\boldsymbol{S} = \boldsymbol{S}^\top = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top.$$

SVD & Eigendecomposition

- Recall the eigendecomposition of a symmetric positive definite matrix

$$\boldsymbol{S} = \boldsymbol{S}^\top = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top.$$

with the corresponding SVD

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

so $\boldsymbol{U} = \boldsymbol{P} = \boldsymbol{V}$, $\boldsymbol{D} = \boldsymbol{\Sigma}$.

The Overall Idea

- Computing the SVD of $\boldsymbol{A} \in \mathbb{R}^{m \times n} \Leftrightarrow$ Finding two sets of orthonormal bases $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ respectively.
- Images of $\boldsymbol{A}\mathbf{v}_i$'s form a set of orthogonal vectors.

The first step: Constructing the right-singular vectors

- **Recall:** Eigenvectors of a *symmetric* matrix form an orthonormal basis (the Spectral theorem).
- Also, we can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{n \times n}$ from any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.
- Thus,

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top = \boldsymbol{P} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \boldsymbol{P}^\top,$$

where $\boldsymbol{P}$ is orthogonal and composed of orthonormal eigenbasis.

⋆ $\lambda_i \geq 0$ are the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$.

The first step (2/2)

- Assume the SVD of $\boldsymbol{A}$ exists.

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

where $\boldsymbol{U}$, $\boldsymbol{V}$ are orthonormal matrices

The first step (2/2)

- Assume the SVD of $\boldsymbol{A}$ exists.

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

where $\boldsymbol{U}$, $\boldsymbol{V}$ are orthonormal matrices $(\because \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I})$.

The first step (2/2)

- Assume the SVD of $\boldsymbol{A}$ exists.

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U\Sigma V}^\top)^\top (\boldsymbol{U\Sigma V}^\top) = \boldsymbol{V\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U\Sigma V}^\top$$

where $\boldsymbol{U}, \boldsymbol{V}$ are orthonormal matrices $(\because \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I})$. So,

The first step (2/2)

- Assume the SVD of $\boldsymbol{A}$ exists.

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

where $\boldsymbol{U}, \boldsymbol{V}$ are orthonormal matrices $(\because \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I})$. So,

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix}\boldsymbol{V}^\top$$

- Hence, we identify $\boldsymbol{V}^\top = \boldsymbol{P}^\top$ (right-singular vectors) and $\sigma_i^2 = \lambda_i$.

The second step: Constructing the left-singular vectors

- Similarly, we can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{m\times m}$ from any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.
- Thus, by assuming the SVD of $\boldsymbol{A}$ exists, we have

$$\begin{aligned}\boldsymbol{A}\boldsymbol{A}^\top &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \\ &= \boldsymbol{U}\begin{bmatrix}\sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m^2\end{bmatrix}\boldsymbol{U}^\top\end{aligned}$$

Note: $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same nonzero eigenvalues.

The second step: Constructing the left-singular vectors

- Similarly, we can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{m\times m}$ from any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.
- Thus, by assuming the SVD of $\boldsymbol{A}$ exists, we have

$$
\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^\top &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \\
&= \boldsymbol{U}\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m^2 \end{bmatrix}\boldsymbol{U}^\top
\end{aligned}
$$

Note: $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same nonzero eigenvalues.

$\Rightarrow$ The nonzero entries of $\boldsymbol{\Sigma}$ in the SVD for both steps must be the same.

The last step: Link up all parts (1/2)

Images of the $\boldsymbol{v}_i$ under $\boldsymbol{A}$ must be orthogonal.

$$(\boldsymbol{A}\mathbf{v}_i)^\top(\boldsymbol{A}\mathbf{v}_j) = \mathbf{v}_i^\top(\boldsymbol{A}^\top\boldsymbol{A})\mathbf{v}_j = \mathbf{v}_i^\top(\lambda_j\mathbf{v}_j) = \lambda_j\mathbf{v}_i^\top\mathbf{v}_j = 0.$$

(For $m \geq r$) We observe that $\{\boldsymbol{A}\mathbf{v}_1, \ldots, \boldsymbol{A}\mathbf{v}_r\}$ is a basis of an r-dimensional subspace of $\mathbb{R}^m$.

The last step: Link up all parts (1/2)

Images of the $\boldsymbol{v}_i$ under $\boldsymbol{A}$ must be orthogonal.

$$(\boldsymbol{A}\mathbf{v}_i)^\top(\boldsymbol{A}\mathbf{v}_j) = \mathbf{v}_i^\top(\boldsymbol{A}^\top\boldsymbol{A})\mathbf{v}_j = \mathbf{v}_i^\top(\lambda_j\mathbf{v}_j) = \lambda_j\mathbf{v}_i^\top\mathbf{v}_j = 0.$$

(For $m \geq r$) We observe that $\{\boldsymbol{A}\mathbf{v}_1, \ldots, \boldsymbol{A}\mathbf{v}_r\}$ is a basis of an r-dimensional subspace of $\mathbb{R}^m$.

- Normalize the images of these right-singular vectors:

$$\mathbf{u}_i := \frac{\boldsymbol{A}\mathbf{v}_i}{\|\boldsymbol{A}\mathbf{v}_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\mathbf{v}_i = \frac{1}{\sigma_i}\boldsymbol{A}\mathbf{v}_i.$$

- That is, $\boldsymbol{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i$, for $i = 1, \ldots, r$.

The last step: Link up all parts (2/2)

- Concatenate the $\mathbf{v}_i$'s as the columns of $\boldsymbol{V}$;
- Concatenate the $\mathbf{u}_i$'s as the columns of $\boldsymbol{U}$;

The last step: Link up all parts (2/2)

- Concatenate the $\mathbf{v}_i$'s as the columns of $\boldsymbol{V}$;
- Concatenate the $\mathbf{u}_i$'s as the columns of $\boldsymbol{U}$;
- Then,

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}.$$

The last step: Link up all parts (2/2)

- Concatenate the $\mathbf{v}_i$'s as the columns of $\boldsymbol{V}$;
- Concatenate the $\mathbf{u}_i$'s as the columns of $\boldsymbol{U}$;
- Then,

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}.$$

Thus,

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

The last step: Link up all parts (2/2)

- Concatenate the $\mathbf{v}_i$'s as the columns of $\boldsymbol{V}$;
- Concatenate the $\mathbf{u}_i$'s as the columns of $\boldsymbol{U}$;
- Then,

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}.$$

Thus,

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

Exercise

Why do we have $\boldsymbol{A} = \boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^\top$?

Outline

1. Singular Value Decomposition (SVD)
   - Construction of the SVD
   - Example
2. Matrix Approximation

Example

Find the singular value decomposition of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}$$

Example

Find the singular value decomposition of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}$$

Note:

$$\boldsymbol{A}^\top \boldsymbol{A} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

SVD Example (step 1/2)

Goal: Find $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

Perform eigendecomposition of $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top$:

SVD Example (step 1/2)

Goal: Find $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

Perform eigendecomposition of $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top$:

$$\boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

SVD Example (step 1/2)

Goal: Find $\boldsymbol{A} = \boldsymbol{U\Sigma V}^\top$.

Perform eigendecomposition of $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{PDP}^\top$:

$$\boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

So,

$$\boldsymbol{V} = \boldsymbol{P} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

where $\sigma^2 = 6, \ \sigma_2^2 = 1 \ \Rightarrow \ \sigma_1 = \sqrt{6}, \sigma_2 = 1.$

SVD Example (step 2/2)

Left-singular vectors:

SVD Example (step 2/2)

Left-singular vectors:

$$\mathbf{u}_1 = \frac{1}{\sigma_1}\boldsymbol{A}\mathbf{v}_1 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} 5/\sqrt{30} \\ -2/\sqrt{30} \\ 1/\sqrt{30} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix},$$

$$\mathbf{u}_2 = \frac{1}{\sigma_2}\boldsymbol{A}\mathbf{v}_2 = \frac{1}{1}\begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}.$$

Then, we derive $\boldsymbol{U} = [\mathbf{u}_1, \mathbf{u}_2] = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$.

Outline

1 Singular Value Decomposition (SVD)
- Construction of the SVD
- Example

2 Matrix Approximation

Motivation

- Represent a matrix $\boldsymbol{A}$ as a sum of simpler low-rank matrices $\boldsymbol{A}_i$.
- Cheaper than computing the full SVD.
- Rank-1 matrix $\boldsymbol{A}_i \in \mathbb{R}^{m \times n}$:

$$\boldsymbol{A}_i = \mathbf{u}_i \mathbf{v}_i^\top . \quad \text{(outer product)}$$

Motivation

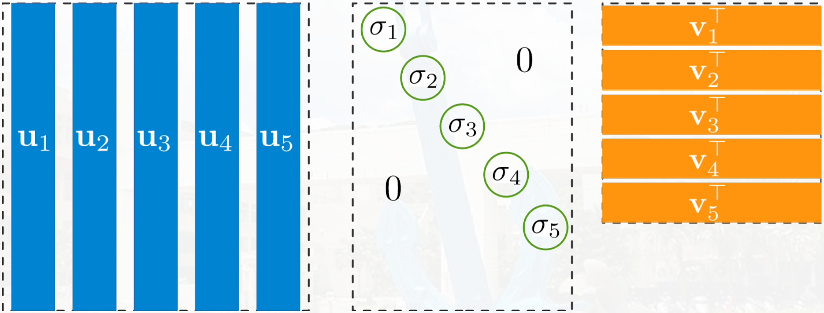
- Represent a matrix $\boldsymbol{A}$ as a sum of simpler low-rank matrices $\boldsymbol{A}_i$.
- Cheaper than computing the full SVD.
- Rank-1 matrix $\boldsymbol{A}_i \in \mathbb{R}^{m \times n}$:

$$\boldsymbol{A}_i = \mathbf{u}_i \mathbf{v}_i^\top . \quad \text{(outer product)}$$

- In fact, we can derive

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i .$$

  - Outer-product matrices $\boldsymbol{A}_i$ weighted by the ith singular value σ_i.



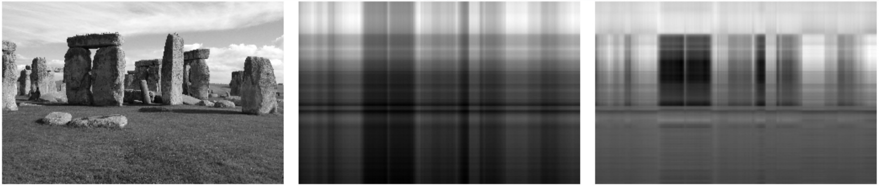
- σ_i is multiplied with $\mathbf{u}_i$ and $\mathbf{v}_i^\top$.

Rank-k Approximation

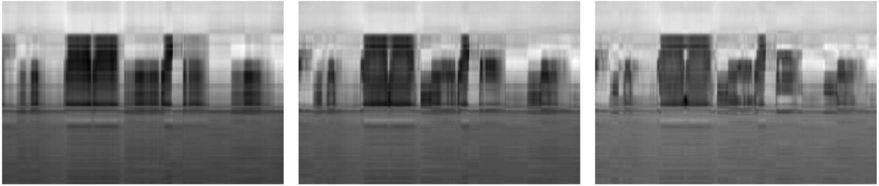
Up to an intermediate value $k < r$ (assume that σ_i's are sorted in decreasing order),

$$\hat{\boldsymbol{A}}(k) := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \sum_{i=1}^{k} \sigma_i \boldsymbol{A}_i$$

Illustrating example



(a) Original image $\boldsymbol{A}$. (b) Rank-1 approximation $\widehat{\boldsymbol{A}}(1)$. (c) Rank-2 approximation $\widehat{\boldsymbol{A}}(2)$.



(d) Rank-3 approximation $\widehat{\boldsymbol{A}}(3)$. (e) Rank-4 approximation $\widehat{\boldsymbol{A}}(4)$. (f) Rank-5 approximation $\widehat{\boldsymbol{A}}(5)$.

Measure the difference b/w $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$

Spectral Norm

For $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, the spectral norm of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is

$$\|\boldsymbol{A}\|_2 = \max_{\mathbf{x}} \frac{\|\boldsymbol{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Measure the difference b/w $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$

Spectral Norm

For $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, the spectral norm of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is

$$\|\boldsymbol{A}\|_2 = \max_{\mathbf{x}} \frac{\|\boldsymbol{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

- Think about why we need to divide the norm $\|\mathbf{x}\|_2$.

Theorem & Exercise

Theorem (4.24)

The spectral norm of $\boldsymbol{A}$ is its largest singular value σ_1.

Eckart-Young Theorem

Theorem [Eckart & Young 1936]

Consider $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r and let $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ be a matrix of rank k.

Then for any $k \leq r$ with $\hat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_1 \mathbf{u}_i \mathbf{v}_i^\top$, it holds that

$$\hat{\boldsymbol{A}}(k) = \underset{\text{rank}(\boldsymbol{B})=k}{\arg\min} \|\boldsymbol{A} - \boldsymbol{B}\|_2,$$

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 = \sigma_{k+1}.$$

Physical meaning:

- We can view the rank-k approximation as a projection of the matrix $\boldsymbol{A}$ onto a lower-dimensional space of rank-at-most-k matrices.
- The approximation error: the next singular value (i.e., σ_{k+1})!

Sketch of the Proof (1/2)

Note that

$$\boldsymbol{A} - \hat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

By Theorem 4.24, we have $\|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 = \sigma_{k+1}$ (spectral norm).

Sketch of the Proof (1/2)

Note that

$$\boldsymbol{A} - \hat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

By Theorem 4.24, we have $\|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 = \sigma_{k+1}$ (spectral norm).

But why $\hat{\boldsymbol{A}}$ is the *best* approximation in some sense?

Sketch of the Proof (1/2)

Note that

$$\boldsymbol{A} - \hat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\top}.$$

By Theorem 4.24, we have $\|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 = \sigma_{k+1}$ (spectral norm).

But why $\hat{\boldsymbol{A}}$ is the *best* approximation in some sense?

Assume that $r > k$ and there is another $\boldsymbol{B}$ with $\operatorname{rank}(\boldsymbol{B}) \leq k$, such that

$$\|\boldsymbol{A} - \boldsymbol{B}\|_2 < \|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2.$$

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z, \ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z, \ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

$\Rightarrow \|\boldsymbol{A}\mathbf{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\mathbf{x}\|_2$

Sketch of the Proof (2/2)

Note that $\dim(\mathrm{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z,\ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

$\Rightarrow \|\boldsymbol{A}\mathbf{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\mathbf{x}\|_2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\mathbf{x}\|_2$

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z,\ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

$$\Rightarrow \|\boldsymbol{A}\mathbf{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\mathbf{x}\|_2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\mathbf{x}\|_2 < \sigma_{k+1} \|\mathbf{x}\|_2. \quad \cdots (\dagger)$$

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z, \ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

$\Rightarrow \|\boldsymbol{A}\mathbf{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\mathbf{x}\|_2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\mathbf{x}\|_2 < \sigma_{k+1} \|\mathbf{x}\|_2. \quad \cdots (\dagger)$

However, there exists a $(k+1)$-dimensional subspace Y spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$.

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z,\ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

$\Rightarrow \|\boldsymbol{A}\mathbf{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\mathbf{x}\|_2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\mathbf{x}\|_2 < \sigma_{k+1} \|\mathbf{x}\|_2. \quad \cdots (\dagger)$

However, there exists a $(k+1)$-dimensional subspace Y spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$.

Note that for $\mathbf{x} \in Y$, $\|\boldsymbol{A}\mathbf{x}\|_2 \geq \sigma_{k+1} \|\mathbf{x}\|_2. \quad \cdots (\ddagger)$ (Exercise!)

Sketch of the Proof (2/2)

Note that $\dim(\text{null}(\boldsymbol{B})) \geq n - k$.

There exists an $\geq (n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$ such that

$$\mathbf{x} \in Z,\ \boldsymbol{B}\mathbf{x} = \mathbf{0}.$$

$\Rightarrow \|\boldsymbol{A}\mathbf{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\mathbf{x}\|_2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\mathbf{x}\|_2 < \sigma_{k+1}\|\mathbf{x}\|_2. \quad \cdots (\dagger)$

However, there exists a $(k+1)$-dimensional subspace Y spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$.

Note that for $\mathbf{x} \in Y$, $\|\boldsymbol{A}\mathbf{x}\|_2 \geq \sigma_{k+1}\|\mathbf{x}\|_2. \quad \cdots (\ddagger)$ (Exercise!)

But by the Dimension Theorem (rank-nullity theorem), there must be $\mathbf{x} \in Y \cap Z$. ($\Rightarrow\!\Leftarrow$)

Sketch of the Proof of ($\ddagger$))

- For any $\mathbf{x} \in Y = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$, write $\mathbf{x} = \sum_{i=1}^{k+1} \alpha_i \mathbf{v}_i$.
- $\mathbf{v}_i$'s and $\mathbf{u}_i$'s are orthonormal $\Rightarrow$

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{k+1} \alpha_i^2, \qquad \boldsymbol{A}\mathbf{x} = \sum_{i=1}^{k+1} \alpha_i \boldsymbol{A}\mathbf{v}_i = \sum_{i=1}^{k+1} \alpha_i \sigma_i \mathbf{u}_i.$$

- Hence,

$$\|\boldsymbol{A}\mathbf{x}\|_2^2 = \left\| \sum_{i=1}^{k+1} \alpha_i \sigma_i \mathbf{u}_i \right\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 \alpha_i^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} \alpha_i^2 = \sigma_{k+1}^2 \|\mathbf{x}\|_2^2.$$

Proof of Theorem 4.24

$$\|\boldsymbol{A}\|_2 = \sup_{\mathbf{x}\neq 0} \frac{\|\boldsymbol{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x}\neq 0} \left\| \boldsymbol{A}\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|\boldsymbol{A}\mathbf{x}\|_2.$$

- Let $\boldsymbol{B} := \boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{n\times n}$.
- Then $\boldsymbol{B}$ is symmetric positive semidefinite and admits the eigen-decomposition $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{PDP}^\top$, for $\boldsymbol{P} = \left[\mathbf{v}_1, \ldots, \mathbf{v}_n\right]$ containing orthonormal column vectors, and $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),\ \lambda_i \geq 0$.

Proof of Theorem 4.24 (Upper Bound)

Write $\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{v}_i$. Then

$$\|\boldsymbol{A}\mathbf{x}\|_2^2 = \mathbf{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} = \left(\sum_{i=1}^{n} a_i \mathbf{v}_i\right)^\top \left(\sum_{i=1}^{n} a_i \lambda_i \mathbf{v}_i\right) = \sum_{i=1}^{n} a_i^2 \lambda_i$$

$$\leq \left(\max_{1 \leq i \leq n} \lambda_i\right) \sum_{i=1}^{n} a_i^2 = \left(\max_i \lambda_i\right) \|\mathbf{x}\|_2^2.$$

Hence,

$$\|\boldsymbol{A}\mathbf{x}\|_2 \leq \sqrt{\max_i \lambda_i}\, \|\mathbf{x}\|_2 \quad \Rightarrow \quad \|\boldsymbol{A}\|_2 \leq \sqrt{\max_i \lambda_i}\,.$$

Proof of Theorem 4.24 (Lower Bound and Final)

- Let $\mathbf{v}_k$ be an eigenvector of $\boldsymbol{A}^\top \boldsymbol{A}$ with $\lambda_k = \max_i \lambda_i$ and $\|\mathbf{v}_k\|_2 = 1$. Then
$$\|\boldsymbol{A}\mathbf{v}_k\|_2^2 = \mathbf{v}_k^\top \boldsymbol{A}^\top \boldsymbol{A}\mathbf{v}_k = \lambda_k.$$
- Therefore $\|\boldsymbol{A}\|_2 \geq \|\boldsymbol{A}\mathbf{v}_k\|_2 \geq \sqrt{\lambda_k} = \sqrt{\max_i \lambda_i}$. Combining with the upper bound gives
$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})} = \sigma_{\max}(\boldsymbol{A}) = \max_i \sigma_i,$$
where σ_i are the singular values of $\boldsymbol{A}$ and $\lambda_i = \sigma_i^2$.

Discussions