SI231b: Matrix Computations

Lecture 19: Singular Value Decomposition

Yue Qiu

qiuyue@shanghaitech.edu.cn

School of Information Science and Technology

ShanghaiTech University

Nov. 13, 2022

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ admits a singular value decomposition (SVD)

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

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal, and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ has $[\mathbf{\Sigma}]_{ij} = 0$ for all $i \neq j$ and $[\mathbf{\Sigma}]_{ii} = \sigma_i$ for all i, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$.

- matrix 2-norm: $\|\mathbf{A}\|_2 = \sigma_1$
- let r be the number of nonzero σ_i's, partition $\mathbf{U} = [\ \mathbf{U}_1\ \mathbf{U}_2\ ]$, $\mathbf{V} = [\ \mathbf{V}_1\ \mathbf{V}_2\ ]$ with $\mathbf{U}_1 \in \mathbb{R}^{m \times r}$ and $\mathbf{V}_1 \in \mathbb{R}^{n \times r}$, and let $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$
  - $\mathrm{rank}(\mathbf{A}) = r$
  - pseudo-inverse: $\mathbf{A}^\dagger = \mathbf{V}_1 \tilde{\mathbf{\Sigma}}^{-1} \mathbf{U}_1^T$
  - LS solution: $\mathbf{x}_{\mathrm{LS}} = \mathbf{A}^\dagger \mathbf{y} + \boldsymbol{\eta}$ for any $\boldsymbol{\eta} \in \mathcal{R}(\mathbf{V}_2)$
  - orthogonal projection: $\mathbf{P}_{\mathbf{A}} = \mathbf{U}_1 \mathbf{U}_1^T$

Main Results

Low-rank Approximation

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $k \in \{1, \ldots, \min\{m, n\}\}$, the problem

$$\min_{\mathbf{B} \in \mathbb{R}^{m \times n},\ \ \mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2^2$$

has an optimal solution given by $\mathbf{B}^\star = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Or equivalently, $\mathbf{B}^\star$ gives the best rank k approximation of $\mathbf{A}$ while using the matrix 2-norm to optimize $\|\mathbf{A} - \mathbf{B}\|^2$.

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

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

$\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and $\boldsymbol{\Sigma}$ takes the form

$$[\boldsymbol{\Sigma}]_{ij} = \begin{cases} \sigma_i, & i = j \\ 0, & i \neq j \end{cases}, \qquad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0, \ \ p = \min\{m, n\}.$$

- the above decomposition is called the singular value decomposition (SVD)
- σ_i is called the ith singular value
- $\mathbf{u}_i$ and $\mathbf{v}_i$ are called the ith left and right singular vectors, resp.
- the following notations may be used to denote singular values of a given $\mathbf{A}$

$$\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \ldots \geq \sigma_p(\mathbf{A}) = \sigma_{\min}(\mathbf{A})$$

- partitioned form: let r be the number of nonzero singular values, and note $\sigma_1 \geq \ldots \sigma_r > 0$, $\sigma_{r+1} = \ldots = \sigma_p = 0$. Then,

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

  where

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

- economic SVD: $\mathbf{A} = \mathbf{U}_1 \tilde{\boldsymbol{\Sigma}} \mathbf{V}_1^T$

- outer-product form: $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$

From the SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, we see that

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{D}_1\mathbf{U}^T, \qquad \mathbf{D}_1 = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2, \underbrace{0, \ldots, 0}_{m-p \text{ zeros}}) \qquad (*)$$

$$\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{D}_2\mathbf{V}^T, \qquad \mathbf{D}_2 = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2, \underbrace{0, \ldots, 0}_{n-p \text{ zeros}}) \qquad (**)$$

Observations:

- $(*)$ and $(**)$ are the eigenvalue decompositions of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$, resp.
- the left singular vector matrix $\mathbf{U}$ of $\mathbf{A}$ is the eigenvector matrix of $\mathbf{A}\mathbf{A}^T$
- the right singular vector matrix $\mathbf{V}$ of $\mathbf{A}$ is the eigenvector matrix of $\mathbf{A}^T\mathbf{A}$
- the squares of nonzero singular values of $\mathbf{A}$, $\sigma_1^2, \ldots, \sigma_r^2$, are the nonzero eigenvalues of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.

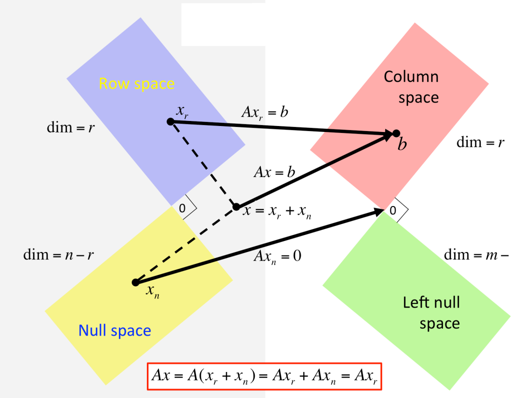

Figure 1: Four fundamental subspaces

In lecture 3, we have learnt that for $\mathbf{A} \in \mathbb{R}^{m \times n}$

- $\mathcal{R}(\mathbf{A}) \perp \mathcal{N}(\mathbf{A}^T)$, and $\mathcal{R}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A}^T) = \mathbb{R}^m$
- $\mathcal{R}(\mathbf{A}^T) \perp \mathcal{N}(\mathbf{A})$, and $\mathcal{R}(\mathbf{A}^T) \oplus \mathcal{N}(\mathbf{A}) = \mathbb{R}^n$

SVD and Four Fundamental Subspaces

Property: The following properties hold:

(a) $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{U}_1)$, $\mathcal{R}(\mathbf{A})^\perp = \mathcal{N}(\mathbf{A}^T) = \mathcal{R}(\mathbf{U}_2)$;

(b) $\mathcal{R}(\mathbf{A}^T) = \mathcal{R}(\mathbf{V}_1)$, $\mathcal{R}(\mathbf{A}^T)^\perp = \mathcal{N}(\mathbf{A}) = \mathcal{R}(\mathbf{V}_2)$;

(c) $\text{rank}(\mathbf{A}) = r$ (the number of nonzero singular values).

Requires a proof.

Note:

- SVD can be used as a numerical tool to compute basis of $\mathcal{R}(\mathbf{A})$, $\mathcal{R}(\mathbf{A})^\perp$, $\mathcal{R}(\mathbf{A}^T)$, $\mathcal{N}(\mathbf{A})$
- we have previously learnt the following properties
  - $\text{rank}(\mathbf{A}^T) = \text{rank}(\mathbf{A})$
  - $\dim \mathcal{N}(\mathbf{A}) = n - \text{rank}(\mathbf{A})$

  By SVD, the above properties are easily seen to be true.
- SVD is also used as a numerical tool to compute the rank of a matrix.

Matrix 2-Norm

Induced matrix p-norm from the vector p-norm

$$\|\mathbf{A}\|_p = \max_{\mathbf{x}\neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_p$$

$p = 2$: matrix 2-norm or spectral norm

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}).$$

Proof:

- for any $\mathbf{x}$ with $\|\mathbf{x}\|_2 \leq 1$,
$$\begin{aligned}\|\mathbf{A}\mathbf{x}\|_2^2 &= \|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}\|_2^2 = \|\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}\|_2^2 \\ &\leq \sigma_1^2\|\mathbf{V}^T\mathbf{x}\|_2^2 = \sigma_1^2\|\mathbf{x}\|_2^2 \leq \sigma_1^2\end{aligned}$$
- $\|\mathbf{A}\mathbf{x}\|_2 = \sigma_1$ if we choose $\mathbf{x} = \mathbf{v}_1$

Implication to linear transformation: let $\mathbf{y} = \mathbf{A}\mathbf{x}$ be a linear transformation maps $\mathbf{x}$ to $\mathbf{y}$. Under the constraint $\|\mathbf{x}\|_2 = 1$, the system output $\|\mathbf{y}\|_2^2$ is maximized when $\mathbf{x}$ is chosen as the 1st right singular vector.

Illustration of Matrix 2-Norm

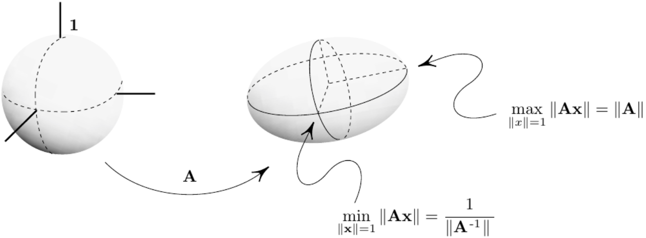

Figure 2: Linear transformation by nonsingular matrix $\mathbf{A}$

When $\mathbf{A} \in \mathbb{R}^{m \times n}$ is of full rank and $m \geq n$,

- $\|\mathbf{Ax}\|_2 \geq \sigma_{\min}(\mathbf{A})\|\mathbf{x}\|_2$ (hands-on exercise)
- can you use Figure 1 to help to understand?

Properties of Matrix 2-Norm

- $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$
  - in fact, $\|\mathbf{AB}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_p$ for any $p \geq 1$
- $\|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2$
  - a special case of the 1st property
- $\|\mathbf{QAW}\|_2 = \|\mathbf{A}\|_2$ for any orthogonal $\mathbf{Q}, \mathbf{W}$
  - we also have $\|\mathbf{QAW}\|_F = \|\mathbf{A}\|_F$ for any orthogonal $\mathbf{Q}, \mathbf{W}$
- $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{p}\|\mathbf{A}\|_2$ (here $p = \min\{m, n\}$)
  - proof: $\|\mathbf{A}\|_F = \|\mathbf{\Sigma}\|_F = \sqrt{\sum_{i=1}^p \sigma_i^2}$, and $\sigma_1^2 \leq \sum_{i=1}^p \sigma_i^2 \leq p\sigma_1^2$
- let $\mathbf{A}$ be square and nonsingular. Then, $\|\mathbf{A}^{-1}\|_2 = 1/\sigma_{\min}(\mathbf{A})$

Schatten p-Norm

The function

$$f(\mathbf{A}) = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})^p \right)^{1/p}, \qquad p \geq 1,$$

defines a matrix norm and is called the Schatten p-norm. Here $\sigma_i(\mathbf{A})$ $(i = 1,\ 2,\ \cdots,\ p)$ are the singular values of $\mathbf{A}$.

Nuclear norm:

$$\|\mathbf{A}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})$$

- a special case of the Schatten p-norm
- finds applications in rank approximation, e.g., for compressive sensing and matrix completion [Recht-Fazel-Parrilo'10]

1. B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, *S*IAM Review, vol. 52, no. 3, pp. 471–501, 2010.

Readings

You are supposed to read

- Gene H. Golub and Charles F. Van Loan. *Matrix Computations*, Johns Hopkins University Press, 2013.

  Chapter 2.4.