CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin

Spring 2024

Lecture 4: Matrix factorizations I - QR, SVD

Outline

Orthogonality

- Two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$.
- A set of vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$ is orthogonal if $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0$ for $i \neq j$; and orthonormal if $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \delta_{ij}$ for $i = j$.
- $\boldsymbol{U} \in \mathbb{R}^{n \times d}$ is orthonormal if $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$. If $\boldsymbol{U}$ is square, then it is orthogonal (or **unitary** if complex), and $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}$.
- *Orthonormal matrices preserve norms:* $\|\boldsymbol{U}\boldsymbol{y}\|_2 = \|\boldsymbol{y}\|_2$.

Projectors

Projection matrix: A symmetric matrix $\boldsymbol{P}$ of the form $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^\top$ is an orthogonal projection matrix, with:

- $\boldsymbol{P}^2 = \boldsymbol{P}$.
- If $\boldsymbol{P}$ is a (orthogonal) projection matrix, then:

$$\bar{\boldsymbol{P}} = \boldsymbol{I} - \boldsymbol{P}$$

  is also a projection matrix.
- If $\boldsymbol{U}$ is an orthonormal basis of $\mathbb{X} \subseteq \mathbb{R}^n$, then:

$$Ran(\boldsymbol{P}) = \mathbb{X}, \text{ and } Ran(\boldsymbol{I} - \boldsymbol{P}) = Null(\boldsymbol{P}) = \mathbb{X}^\perp$$

Question: $\boldsymbol{P}\bar{\boldsymbol{P}} = ?$

Subspaces of a matrix

Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and consider $Ran(\boldsymbol{A})^{\perp}$, then :

$$Ran(\boldsymbol{A})^{\perp} = Null(\boldsymbol{A}^{\top})$$

Proof: Any $\boldsymbol{x} \in Ran(\boldsymbol{A})^{\perp}$ iff $\langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{x} \rangle = 0$ for all $\boldsymbol{y}$.
This is same as $\langle \boldsymbol{y}, \boldsymbol{A}^{\top}\boldsymbol{x} \rangle = 0$ for all $\boldsymbol{y}$.

Similarly, we also have:

$$Ran(\boldsymbol{A}^{\top}) = Null(\boldsymbol{A})^{\perp}$$

Thus:

$$\begin{aligned} \mathbb{R}^{n} &= Ran(\boldsymbol{A}) \oplus Null(\boldsymbol{A}^{\top}) \\ \mathbb{R}^{d} &= Ran(\boldsymbol{A}^{\top}) \oplus Null(\boldsymbol{A}) \end{aligned}$$

Finding an orthonormal basis of a subspace

- **Goal:** Find vector in $span(\boldsymbol{A})$ closest to some vector $\boldsymbol{b}$.
- Much easier with an orthonormal basis for $span(\boldsymbol{A})$.

Given $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d]$, compute $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d]$ which has orthonormal columns and s.t. $span(\boldsymbol{Q}) = span(\boldsymbol{A})$.

Each column of $\boldsymbol{A}$ must be a linear combination of certain columns of $\boldsymbol{Q}$.

Gram-Schmidt process: Compute $\boldsymbol{Q}$ so that $\boldsymbol{a}_j$ (j column of $\boldsymbol{A}$) is a linear combination of the first j columns of $\boldsymbol{Q}$.

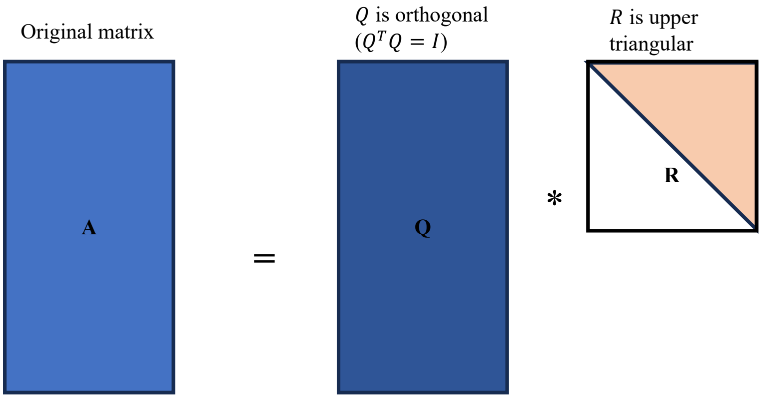
The QR Decomposition

Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with $n \geq d$, and $\text{rank}(\boldsymbol{A}) = d$, there is a $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{R} \in \mathbb{R}^{d \times d}$, s.t.

- $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$
- $\boldsymbol{Q}$ has orthonormal columns, $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$.
- $\boldsymbol{R}$ is upper triangular, $r_{ij} = 0$ for $i > j$.

We have $span(\boldsymbol{Q}) = span(\boldsymbol{A})$, the columns of $\boldsymbol{Q}$ are an orthonormal basis of $span(\boldsymbol{A})$.

Question: What is the computational cost of QR?


Original matrix
A
=
Q is orthogonal
(Q^T Q = I)
Q
*
R is upper
triangular
R

Least squares using QR

- *Recall:* In the *least-squares* regression problem, assuming $n \geq d$, we solve:

$$\boldsymbol{x}^* = \min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2.$$

- If $\boldsymbol{A}$ is full rank then we compute $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$.
- The normal equation can be written as:

$$\begin{aligned} \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{b} \quad &\rightarrow \quad \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{b} \\ &\rightarrow \quad \boldsymbol{R}^\top \boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{b} \\ &\rightarrow \quad \boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^\top \boldsymbol{b}. \end{aligned}$$

- Therefore,

$$\boldsymbol{x}^* = \boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{b}.$$

Note that $\boldsymbol{R}$ is non-singular.

- Alternatively, recall that $span(\boldsymbol{Q}) = span(\boldsymbol{A})$.
- We know that $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2$ is minimum when $\boldsymbol{Ax} - \boldsymbol{b} \perp span(\boldsymbol{Q})$.
- This implies what?

As a rule it is not a good idea to form $\boldsymbol{A}^\top \boldsymbol{A}$ and solve the normal equations.
Methods using the QR factorization are better.
Why?

QR factorization is also used in direct solvers of linear system $\boldsymbol{Ax} = \boldsymbol{b}$.

The Singular Value Decomposition

SVD

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ there exist unitary matrices $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ such that

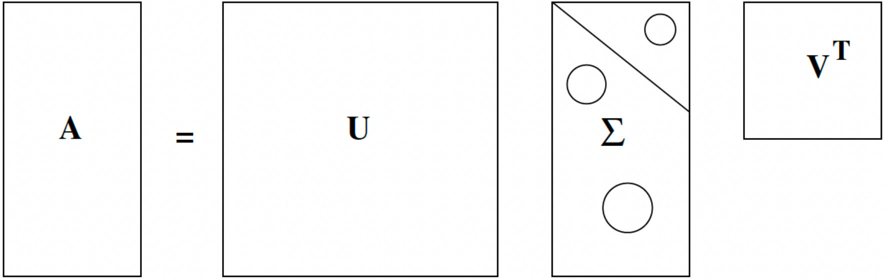
$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$$

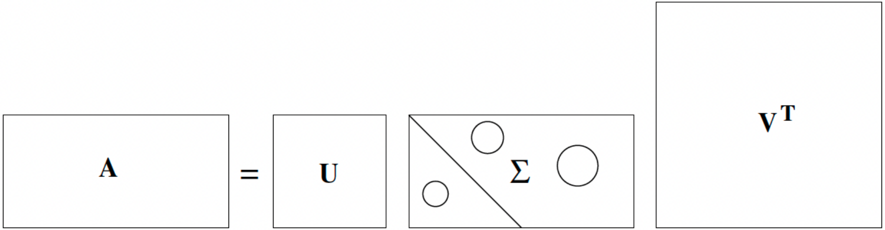
where Σ is a diagonal matrix with entries $\sigma_i \geq 0$.

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \text{ with } p = \min(n, d)$$

Let $\sigma_1 = \|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}, \|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$. There exists a pair of unit vectors such that

$$\boldsymbol{A}\boldsymbol{v}_1 = \sigma_1 \boldsymbol{u}_1.$$


A
=
U
Σ
V T


A
=
U
Σ
V T

Thin SVD

- In the first case, suppose , we can write

$$\boldsymbol{A} = [\boldsymbol{U}_1 \ \boldsymbol{U}_2] \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \boldsymbol{V}^\top,$$

  where $\boldsymbol{U}_1 \in \mathbb{R}^{n \times d}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{n \times n-d}$. Then,

$$\boldsymbol{A} = \boldsymbol{U}_1 \Sigma_1 \boldsymbol{V}^\top,$$

  where $\Sigma_1, \boldsymbol{V} \in \mathbb{R}^{d \times d}$.
- Referred to as *thin or economical* SVD.

Question: How to compute the thin SVD of $\boldsymbol{A}$ from its QR factorization?

SVD Properties

Suppose

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \text{ and } \sigma_{r+1} = \cdots = \sigma_p = 0$$

Then:

- $\text{rank}(\boldsymbol{A}) = r =$ number of nonzero singular values.
- $Ran(\boldsymbol{A}) = span\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\}$
- $Null(\boldsymbol{A}^\top) = span\{\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \ldots, \boldsymbol{u}_n\}$
- $Ran(\boldsymbol{A}^\top) = span\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$
- $Null(\boldsymbol{A}) = span\{\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_d\}$

SVD Properties II

- A matrix $\boldsymbol{A}$ admits the SVD expansion

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

- $\|\boldsymbol{A}\|_2 = \sigma_1 =$ largest singular value.
- $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$.

Eckart-Young-Mirsky Theorem

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with rank r, let $k \leq r$ and $\boldsymbol{A}_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ then

$$\min_{\boldsymbol{B}:\mathrm{rank}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2 = \|\boldsymbol{A} - \boldsymbol{A}_k\|_2 = \sigma_{k+1}.$$

Pseudo-inverse

- Given $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$, we rewrite it as :

$$\boldsymbol{A} = [\boldsymbol{U}_1\ \boldsymbol{U}_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^\top \\ \boldsymbol{V}_2^\top \end{bmatrix} = \boldsymbol{U}_1\Sigma_1\boldsymbol{V}_1^\top$$

- Then the pseudo inverse of $\boldsymbol{A}$ is:

$$\boldsymbol{A}^\dagger = \boldsymbol{V}_1\Sigma_1^{-1}\boldsymbol{U}_1^\top = \sum_{i=1}^{r} \frac{1}{\sigma_i}\boldsymbol{v}_i\boldsymbol{u}_i^\top$$

- The pseudo-inverse of $\boldsymbol{A}$ is the mapping from a vector $\boldsymbol{b}$ to the (unique) **Minimum Norm solution** of the LS problem: $\min_{\boldsymbol{x}\in\mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.

$$\boldsymbol{x} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b} = \boldsymbol{A}^\dagger\boldsymbol{b}.$$

- Let us express solution $\boldsymbol{x}$ in basis $\boldsymbol{V}$ as: $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y} = [\boldsymbol{V}_1,\ \boldsymbol{V}_2] \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$.
- Then left multiply by $\boldsymbol{U}^\top$ to get:

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \left\| \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}_1^\top \boldsymbol{b} \\ \boldsymbol{U}_2^\top \boldsymbol{b} \end{bmatrix} \right\|_2^2$$

- Let us find all possible solutions in terms of $\boldsymbol{y} = [\boldsymbol{y}_1; \boldsymbol{y}_2]$.
- From above, we have $\boldsymbol{y}_1 = \Sigma_1^{-1}\boldsymbol{U}_1^\top \boldsymbol{b}$ and $\boldsymbol{y}_2$ can be anything.
- Then,

$$\begin{aligned} \boldsymbol{x} &= [\boldsymbol{V}_1,\ \boldsymbol{V}_2] \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \boldsymbol{V}_1\boldsymbol{y}_1 + \boldsymbol{V}_2\boldsymbol{y}_2 \\ &= \boldsymbol{V}_1\Sigma_1^{-1}\boldsymbol{U}_1^\top \boldsymbol{b} + \boldsymbol{V}_2\boldsymbol{y}_2 \\ &= \boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{V}_2\boldsymbol{y}_2. \end{aligned}$$

- We know that $\boldsymbol{A}^{\dagger}\boldsymbol{b} \in Ran(\boldsymbol{A}^{\top})$ and $\boldsymbol{V}_2\boldsymbol{y}_2 \in Null(\boldsymbol{A})$.
- Therefore: least-squares solutions are all of the form:

$$\boldsymbol{A}^{\dagger}\boldsymbol{b} + \boldsymbol{w} \quad \text{where} \quad \boldsymbol{w} \in Null(\boldsymbol{A}).$$

- We obtain the smallest norm when $\boldsymbol{w} = 0$.
- The *Minimum Norm solution* of the LS problem: $\min_{\boldsymbol{x}\in\mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ is :

$$\boldsymbol{x}_{LS} = \boldsymbol{V}_1\Sigma_1^{-1}\boldsymbol{U}_1^{\top}\boldsymbol{b} = \boldsymbol{A}^{\dagger}\boldsymbol{b}.$$

Moore-Penrose Inverse

The pseudo-inverse of $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is given by

$$\boldsymbol{A}^\dagger = \boldsymbol{V} \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{U}^\top = \sum_{i=1}^{r} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top$$

Properties:

- $\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}$ $\quad \boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{A}^\dagger$ $\quad (\boldsymbol{A}^\dagger\boldsymbol{A})^H = \boldsymbol{A}^\dagger\boldsymbol{A}$ $\quad (\boldsymbol{A}\boldsymbol{A}^\dagger)^H = \boldsymbol{A}\boldsymbol{A}^\dagger$
- $\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{I}$ when $\operatorname{rank}(\boldsymbol{A}) = d$, and $\boldsymbol{A}^\dagger = \boldsymbol{A}^{-1}$ if $\boldsymbol{A}$ is invertible.
- **Left inverse:** $\boldsymbol{A}^\dagger = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$, when $n \geq d$, and $\boldsymbol{A}$ is full rank.
- **Right inverse:** $\boldsymbol{A}^\dagger = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}$, when $n \leq d$, and $\boldsymbol{A}$ is full rank.

Exercises

- $\boldsymbol{A}\boldsymbol{A}^{\dagger}$ is a projector onto which space?
- $\boldsymbol{A}^{\dagger}\boldsymbol{A}$ is a projector onto which space?

Questions?