CSE 392: Matrix and Tensor Algorithms for Data Spring 2024

Lecture 4 — 01/29/2024

Instructor: Shashanka Ubaru *Scribe: Sung Jung*

1 Orthogonality and Projections

1.1 Orthogonality

The key concept in this lecture is **orthogonality**. Two vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. For a set of vectors $\{\mathbf{u}_1, ..., \mathbf{u}_d\}$, we say that it is orthogonal of $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for $i \neq j$. If the set also has the property that $\langle \mathbf{u}_i, \mathbf{u}_i \rangle = 1$, then we say that it is **orthonormal**.

If these orthonormal vectors were to be made columns of a matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$, then you can see that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$; we call such a matrix as an *orthonormal matrix*. If $d = n$ (square matrix), then $\mathbf{U}^T = \mathbf{U}^{-1}$ (since multiplying it by $\mathbf{U}$ results in identity), so $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ as well.

The key property of an orthonormal matrix is that it preserves norms of vectors:

$$||\mathbf{U}\mathbf{y}||_2^2 = \mathbf{y}^T\mathbf{U}^T\mathbf{U}\mathbf{y} = \mathbf{y}^T\mathbf{y} = ||\mathbf{y}||_2^2 \tag{1}$$

1.2 Subspaces of a Matrix

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and consider its column space, $\mathcal{C}(\mathbf{A})$. The null space of $\mathbf{A}^T$ is the orthogonal complement of $\mathcal{C}(\mathbf{A})$ in $\mathbb{R}^n$:

$$\mathcal{C}(\mathbf{A})^\perp = Null(\mathbf{A}^T) \tag{2}$$

This is because any $\mathbf{x} \in \mathcal{C}(\mathbf{A})^\perp$ if and only if $\langle \mathbf{A}\mathbf{y}, \mathbf{x} \rangle = 0$ for all $\mathbf{y}$. This is the same as saying $\langle \mathbf{y}, \mathbf{A}^T\mathbf{x} \rangle = 0$ for all $\mathbf{y}$, in which case it must be true that $\mathbf{A}^T\mathbf{x} = 0$.

Similarly, we also have:

$$\mathcal{C}(\mathbf{A}^T) = Null(\mathbf{A})^\perp \tag{3}$$

Thus:

$$\mathbb{R}^n = \mathcal{C}(\mathbf{A}) \bigoplus Null(\mathbf{A}^T) \tag{4}$$

$$\mathbb{R}^d = \mathcal{C}(\mathbf{A}^T) \bigoplus Null(\mathbf{A}) \tag{5}$$

1.3 Projection

An important operator that makes use of orthogonality is the projector. The definition of the **projection matrix** of some matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ is a matrix $\mathbf{P}$ that keeps any vector in the column space of $\mathbf{X}$ unchanged and eliminates any vector that is orthogonal to the column space. In other words:

Definition. $\mathbf{P}$ is a projection matrix of $\mathbf{X}$ if:

- $\mathbf{v} \in \mathcal{C}(\mathbf{X}) \Rightarrow \mathbf{P}\mathbf{v} = \mathbf{v}$
- $\mathbf{w} \in \mathcal{C}(\mathbf{X})^{\perp} \Rightarrow \mathbf{P}\mathbf{w} = 0$

For the matrix multiplication to work out, we see that $\mathbf{P} \in \mathbb{R}^{n \times n}$.

With the definition above, we can prove that the column spaces of $\mathbf{X}$ and $\mathbf{P}$ are equivalent.

Theorem 1. $\mathcal{C}(\mathbf{P}) = \mathcal{C}(\mathbf{X})$

Proof. $\Rightarrow$ Take a vector $\mathbf{v}_1 \in \mathcal{C}(\mathbf{X}) \subseteq \mathbb{R}^n$. Then, by definition, $\mathbf{P}\mathbf{v}_1 = \mathbf{v}_1 \in \mathcal{C}(\mathbf{P})$. Therefore, $\mathcal{C}(\mathbf{X}) \subseteq \mathcal{C}(\mathbf{P})$.

$\Leftarrow$ Take a vector $\mathbf{v}_2 \in \mathcal{C}(\mathbf{P}) \subseteq \mathbb{R}^n$. This means that $\mathbf{v}_2 = \mathbf{P}\mathbf{y}$ for some $\mathbf{y} \in \mathbb{R}^n$. Since $\mathcal{C}(\mathbf{X}) \subseteq \mathbb{R}^n$, we can write $\mathbf{y} = \alpha\mathbf{v} + \beta\mathbf{w}$, where $\mathbf{v} \in \mathcal{C}(\mathbf{X})$, $\mathbf{w} \in \mathcal{C}(\mathbf{X})^{\perp}$, and $\alpha, \beta \in \mathbb{R}$.

Then, $\mathbf{v}_2 = \mathbf{P}\mathbf{y} = \mathbf{P}(\alpha\mathbf{v} + \beta\mathbf{w}) = \alpha\mathbf{P}\mathbf{v} + \beta\mathbf{P}\mathbf{w} = \alpha\mathbf{P}\mathbf{v} = \alpha\mathbf{v} \in \mathcal{C}(\mathbf{X})$. So, $\mathcal{C}(\mathbf{P}) \subseteq \mathcal{C}(\mathbf{X})$.

Since $\mathcal{C}(\mathbf{X}) \subseteq \mathcal{C}(\mathbf{P})$ and $\mathcal{C}(\mathbf{P}) \subseteq \mathcal{C}(\mathbf{X})$, $\mathcal{C}(\mathbf{P}) = \mathcal{C}(\mathbf{X})$ ■

From the definition of a projection matrix above, notice that $(\mathbf{I} - \mathbf{P})\mathbf{w} = \mathbf{w}$ and $(\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{0}$. Therefore, the matrix $(\mathbf{I} - \mathbf{P})$ is a projection matrix onto $\mathcal{C}(\mathbf{X})^{\perp}$, and we can show that $\mathcal{C}(\mathbf{X})^{\perp} = \mathcal{C}(\mathbf{I} - \mathbf{P})$ following the proof above with $(\mathbf{I} - \mathbf{P})$.

1.3.1 Projection Matrices are Symmetric and Idempotent

Some key properties of a projection matrix are given by the following theorem:

Theorem 2. $\mathbf{P}$ *is a projection matrix onto* $\mathcal{C}(\mathbf{P})$ *if and only if* $\mathbf{P} = \mathbf{P}^2 = \mathbf{P}^T$

Proof. $\Rightarrow$ Suppose $\mathbf{P}$ is a projection matrix. For any $\mathbf{v}$ and $\mathbf{w}$, we can write it as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ and $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$, where $\mathbf{v}_1, \mathbf{w}_1 \in \mathcal{C}(\mathbf{P})$ and $\mathbf{v}_2, \mathbf{w}_2 \in \mathcal{C}(\mathbf{P})^{\perp}$.

Note that $(\mathbf{I} - \mathbf{P})\mathbf{v} = (\mathbf{I} - \mathbf{P})(\mathbf{v}_1 + \mathbf{v}_2) = (\mathbf{I} - \mathbf{P})\mathbf{v}_1 + (\mathbf{I} - \mathbf{P})\mathbf{v}_2 = \mathbf{v}_2$. Similarly, $\mathbf{P}\mathbf{w} = \mathbf{P}\mathbf{w}_1 = \mathbf{w}_1$.

So,

$$(\mathbf{P}\mathbf{w})^T(\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{w}_1^T\mathbf{v}_2 = 0 \text{ for any } \mathbf{v} \text{ and } \mathbf{w}$$

$$\mathbf{P}^T(\mathbf{I} - \mathbf{P}) = 0$$

$$\mathbf{P}^T = \mathbf{P}^T\mathbf{P}$$

Since $\mathbf{P}^T\mathbf{P}$ is symmetric, $\mathbf{P}^T$ must also be symmetric. So, $\mathbf{P} = \mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{P}^2$.

$\Leftarrow$ If $\mathbf{P} = \mathbf{P}^T = \mathbf{P}^2$, we show that the definitions of a projection matrix hold.

- For $\mathbf{v} \in \mathcal{C}(\mathbf{P})$, $\mathbf{v} = \mathbf{P}\mathbf{b}$ for some $\mathbf{b}$. So, $\mathbf{P}\mathbf{v} = \mathbf{P}(\mathbf{P}\mathbf{b}) = (\mathbf{P}\mathbf{P})\mathbf{b} = \mathbf{P}\mathbf{b} = \mathbf{v}$.
- For $\mathbf{w} \in \mathcal{C}(\mathbf{P})^\perp$, it is orthogonal to all the columns of $\mathbf{P}$. Therefore, $\mathbf{P}^T\mathbf{w} = \mathbf{P}\mathbf{w} = 0$.

■

Note that $\mathcal{C}(\mathbf{P})^\perp = \mathcal{C}(\mathbf{P}^T)^\perp = Null(\mathbf{P})$ by Equation 3. So, $\bar{\mathbf{P}} \equiv (\mathbf{I} - \mathbf{P})$ is a projection matrix onto the null space of $\mathbf{P}$.

1.3.2 The Projection Matrix onto a Column Space is Unique

Suppose we have two matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ that are both projection matrices onto $\mathcal{C}(\mathbf{X})$. Take an arbitrary vector $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1 \in \mathcal{C}(\mathbf{X})$ and $\mathbf{v}_2 \in \mathcal{C}(\mathbf{X})^\perp$. Then, $\mathbf{P}_1\mathbf{v} = \mathbf{v}_1 = \mathbf{P}_2\mathbf{v}$. Rearranging, $(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{v} = 0$.

Since $\mathbf{v}$ is arbitrary, it must be that $\mathbf{P}_1 - \mathbf{P}_2 = 0$. Therefore, $\mathbf{P}_1 = \mathbf{P}_2$.

1.3.3 Construction of a Projection Matrix with an Orthonormal Matrix

Suppose we want to construct a projection matrix $\mathbf{P}$ onto $\mathcal{C}(\mathbf{X})$, and we have a matrix $\mathbf{U} \in \mathbb{R}^{n\times d}$ with orthonormal columns that span $\mathcal{C}(\mathbf{X})$. Then, the (unique) projection matrix onto $\mathcal{C}(\mathbf{X})$ is $\mathbf{P} = \mathbf{U}\mathbf{U}^T$.

To prove this, we just need to show that Theorem 2 holds:

- $\mathbf{P}^T = (\mathbf{U}\mathbf{U}^T)^T = (\mathbf{U}^T)^T\mathbf{U}^T = \mathbf{U}\mathbf{U}^T = \mathbf{P}$
- $\mathbf{P}^2 = (\mathbf{U}\mathbf{U}^T)(\mathbf{U}\mathbf{U}^T) = \mathbf{U}(\mathbf{U}^T\mathbf{U})\mathbf{U}^T = \mathbf{U}\mathbf{I}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T = \mathbf{P}$

Since we showed in the previous section that the projection matrix onto a given column space is unique, $\mathbf{U}\mathbf{U}^T$ is the one and only projection matrix onto $\mathcal{C}(\mathbf{X})$.

With $\mathbf{u}_i$ as the i-th column of $\mathbf{U}$ and a vector $v \in \mathbb{R}^n$, notice that

$$\mathbf{P}\mathbf{v} = \mathbf{U}\mathbf{U}^T\mathbf{v} = (\sum_{i=1}^{d} \mathbf{u}_i\mathbf{u}_i^T)\mathbf{v} = \sum_{i=1}^{d}(\mathbf{u}_i\mathbf{u}_i^T\mathbf{v}) \tag{6}$$

With a set of orthonormal vectors $\{\mathbf{u}_1, ..., \mathbf{u}_d\}$, projecting a vector $\mathbf{v}$ onto their span is equivalent to projecting $\mathbf{v}$ onto each individual $\mathbf{u}_i$ and summing all the projections.

2 Gram-Schmidt and the QR Decomposition

2.1 Gram-Schmidt Process

One such process to find an orthonormal basis of a subspace is the **Gram-Schmidt process**. Given a matrix $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_d]$ (assume full rank for illustration), we would like to compute $\mathbf{Q} = [\mathbf{q}_1, ..., \mathbf{q}_d]$ which has orthonormal columns and such that $\mathcal{C}(\mathbf{A}) = \mathcal{C}(\mathbf{Q})$.

In the Gram-Schmidt process, we compute $\mathbf{Q}$ such that $\mathbf{a}_j$ (the j-th column of $\mathbf{A}$) is a linear combination of the first j columns of $\mathbf{Q}$. The key idea is that for $\mathbf{a}_i$, we subtract from it the projection of it to the span of the already-calculated $\{\mathbf{q}_1, ..., \mathbf{q}_{i-1}\}$; since $\{\mathbf{q}_1, ..., \mathbf{q}_{i-1}\}$ is a set of orthonormal vectors, we can subtract the projections one by one ($\mathbf{q}_k \mathbf{q}_k^T \mathbf{a}_i$), as seen in Equation 6. It goes as follows:

$\tilde{\mathbf{q}}_1 = \mathbf{a}_1$	$\mathbf{q}_1 = \tilde{\mathbf{q}}_1 / \|\|\tilde{\mathbf{q}}_1\|\|_2$
$\tilde{\mathbf{q}}_2 = \mathbf{a}_2 - \mathbf{q}_1 \mathbf{q}_1^T \mathbf{a}_2$	$\mathbf{q}_2 = \tilde{\mathbf{q}}_2 / \|\|\tilde{\mathbf{q}}_2\|\|_2$
$\tilde{\mathbf{q}}_3 = \mathbf{a}_3 - \mathbf{q}_1 \mathbf{q}_1^T \mathbf{a}_3 - \mathbf{q}_2 \mathbf{q}_2^T \mathbf{a}_3$	$\mathbf{q}_3 = \tilde{\mathbf{q}}_3 / \|\|\tilde{\mathbf{q}}_3\|\|_2$
$\tilde{\mathbf{q}}_4 = \mathbf{a}_4 - \mathbf{q}_1 \mathbf{q}_1^T \mathbf{a}_4 - \mathbf{q}_2 \mathbf{q}_2^T \mathbf{a}_4 - \mathbf{q}_2 \mathbf{q}_2^T \mathbf{a}_4$	$\mathbf{q}_3 = \tilde{\mathbf{q}}_3 / \|\|\tilde{\mathbf{q}}_3\|\|_2$
...	...
$\tilde{\mathbf{q}}_d = \mathbf{a}_d - \sum_{i=1}^{d} \mathbf{q}_i \mathbf{q}_i^T \mathbf{a}_d$	$\mathbf{q}_d = \tilde{\mathbf{q}}_d / \|\|\tilde{\mathbf{q}}_d\|\|_2$

The computation of $\alpha_k = \mathbf{q}_k^T \mathbf{a}_i$ costs about $2n$ operations, and the multiplication $\alpha_k \mathbf{q}_k$ costs n operations. In each step, we sum $\alpha_k \mathbf{q}_k$ roughly d times, and there are d steps. So, the left column of the table above has a total cost of $O(nd^2)$ operations. In the right column, the cost of calculating the norm takes $O(n)$ operations, and the division takes $O(1)$ operations. With d steps, the total cost of the right column is $O(nd)$, which is negligible compared to the cost of the left column. Therefore, the total cost of the Gram-Schmidt process is $O(nd^2)$.

2.2 QR Decomposition

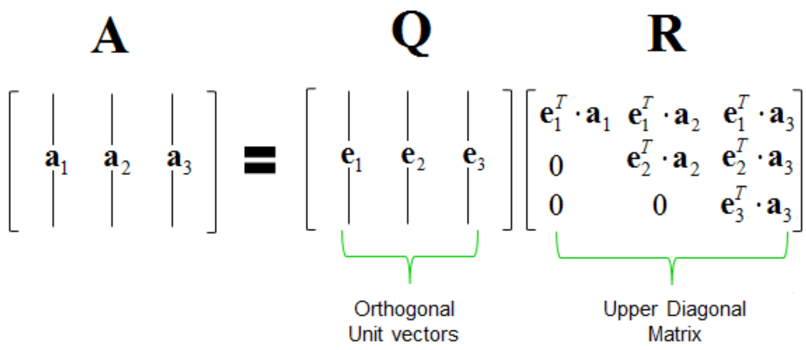
The Gram-Schmidt process is one method to calculate the **QR decomposition**. It states that given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $n \geq d$ and $\text{rank}(\mathbf{A}) = d$ (these conditions are just for the purposes of this class; QR can still be done without them), there is a $\mathbf{Q} \in \mathbb{R}^{n \times d}$ and $\mathbf{R} \in \mathbb{R}^{d \times d}$ such that

- $\mathbf{A} = \mathbf{QR}$
- $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ (orthonormal columns)
- $R_{ij} = 0$ for $i > j$ (upper triangular)

In this setup, the columns of $\mathbf{Q}$ are an orthonormal basis of $\mathcal{C}(\mathbf{A})$.

Using the $\{\mathbf{q}_1, ..., \mathbf{q}_d\}$ from Gram-Schmidt, we can rearrange the Table above to find that we can recover $\mathbf{A}$ if $\mathbf{R}$ collects the inner products between $\mathbf{q}_i$ and $\mathbf{a}_j$:

- $R_{ij} = \mathbf{q}_i^T \mathbf{a}_j$ for $i < j$
- $R_{ii} = ||\tilde{\mathbf{q}}_i||_2 = \mathbf{q}_i^T \mathbf{a}_i$



More numerically accurate methods exist for computing the QR decomposition, including the Givens and Householder's methods.

2.2.1 Least Squares using QR

Recall from lecture 3 that in the least-squares regression problem, we solve

$$\mathbf{x}^* = \arg\min_{x \in \mathbb{R}^d} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 \tag{7}$$

for $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$.

The solution satisfied the normal equation:

$$\mathbf{A}^T\mathbf{A}\mathbf{x}^* = \mathbf{A}^T\mathbf{b} \tag{8}$$

However, we saw that working with $\mathbf{A}^T\mathbf{A}$ is not ideal because we need to compute its inverse, and it may be highly ill-conditioned.

Instead, if we write the normal equation with the QR decomposition of $\mathbf{A}$,

$$\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\mathbf{x}^* = \mathbf{R}^T\mathbf{Q}^T\mathbf{b} \tag{9}$$

$$\mathbf{R}^T\mathbf{R}\mathbf{x}^* = \mathbf{R}^T\mathbf{Q}^T\mathbf{b} \tag{10}$$

$$\mathbf{R}\mathbf{x}^* = \mathbf{Q}^T\mathbf{b} \tag{11}$$

Alternatively, using the column picture of least squares from lecture 3, we know that least squares seeks to minimize the length (ℓ_2 norm) of the error vector $\mathbf{e} = \mathbf{b} - \mathbf{A}\mathbf{x}$. This happens when it is orthogonal to the column space of $\mathbf{A}$, which is equivalent to that of $\mathbf{Q}$. Thus,

$$\mathbf{Q}^T\mathbf{e}^* = 0 \tag{12}$$
$$\mathbf{Q}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^*) = 0 \tag{13}$$
$$\mathbf{Q}^T\mathbf{A}\mathbf{x}^* = \mathbf{Q}^T\mathbf{b} \tag{14}$$
$$\mathbf{Q}^T\mathbf{Q}\mathbf{R}\mathbf{x}^* = \mathbf{Q}^T\mathbf{b} \tag{15}$$
$$\mathbf{R}\mathbf{x}^* = \mathbf{Q}^T\mathbf{b} \tag{16}$$

Since $\mathbf{R}$ is upper triangular, this system of equations can be solved using back substitution.

3 Singular Value Decomposition

Another important matrix decomposition involving orthonormal matrices is the **singular value decomposition** (SVD). It states that for any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, there exist orthonormal matrices $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{17}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times d}$ is a matrix whose top left $(p \times p)$ principal submatrix is diagonal with entries $\sigma_i \geq 0$, with $p = \min(n, d)$.

The columns of $\mathbf{U}$ are called the *left singular vectors* of $\mathbf{A}$, and they are the same as the eigenvectors of $\mathbf{A}\mathbf{A}^T$, which span $\mathbb{R}^n$:

$$\mathbf{A}\mathbf{A}^T = (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)(\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T) = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T = \mathbf{U}\Sigma_n^2\mathbf{U}^T \tag{18}$$

Since $\mathbf{A}\mathbf{A}^T$ is positive semidefinite, $\mathbf{U}$ is orthonormal; it can also be made square by including the bases of the null space of $\mathbf{A}\mathbf{A}^T$ as eigenvectors with their corresponding eigenvalues equal to 0. Because the eigenvalues in Σ_n^2 and the corresponding eigenvectors in the columns of $\mathbf{U}$ can be ordered in any way to give the product $\mathbf{A}\mathbf{A}^T$, assume that the eigenvalues are ordered in a descending manner:

$$\sigma_1^2 \geq \sigma_2^2 \geq ... \geq \sigma_p^2 \geq 0 \tag{19}$$

The columns of $\mathbf{V}$ are called the *right singular vectors* of $\mathbf{A}$, and they are the same as the eigenvectors of $\mathbf{A}^T\mathbf{A}$, which span $\mathbb{R}^d$:

$$\mathbf{A}^T\mathbf{A} = (\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T)(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T) = \mathbf{V}\Sigma^T\Sigma\mathbf{V}^T = \mathbf{V}\Sigma_d^2\mathbf{V}^T \tag{20}$$

Again, $\mathbf{V}$ can be made square and orthonormal because $\mathbf{A}^T\mathbf{A}$ is positive semidefinite, and we can order the columns of $\mathbf{V}$ such that their corresponding σ_i^2 are in descending order.

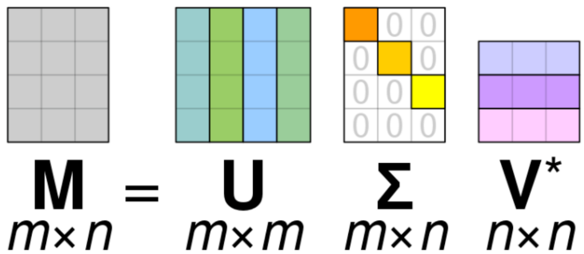
σ_i is called the i-th *singular value* of $\mathbf{A}$, and σ_i^2 is equal to the i-th eigenvalue of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$. It can be proven that $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ have the same nonzero eigenvalues.

Proof. Let $\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\lambda \neq 0$ and $||\mathbf{x}||_2 \neq 0$; λ is a nonzero eigenvalue of $\mathbf{A}^T\mathbf{A}$ with its corresponding eigenvector being $\mathbf{x}$.

Then, multiplying it by $\mathbf{A}$ gives $\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}) = \lambda(\mathbf{A}\mathbf{x})$; λ is also a (nonzero) eigenvalue of $\mathbf{A}\mathbf{A}^T$ with its corresponding eigenvector being $\mathbf{A}\mathbf{x}$.

Note that $\mathbf{A}\mathbf{x} \neq \mathbf{0}$ without violating our conditions of λ and $\mathbf{x}$ being nonzero. If it were zero, then it would imply that $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{0} = \mathbf{0}$, but since $\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, it would mean that either $\lambda = 0$ or $\mathbf{x} = \mathbf{0}$, both of which cannot happen given our starting conditions.

Conversely, if $\mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\mathbf{x}$ with $\lambda \neq 0$, then multiplying by $\mathbf{A}^T$ gives $\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{x}) = \lambda(\mathbf{A}^T\mathbf{x})$. ■



3.1 SVD Properties

Let $\text{rank}(\mathbf{A}) = r \leq p$. Then:

1. $r =$ number of nonzero singular values

Proof. The rank of a matrix does not change when multiplied by non-singular matrices. Since $\mathbf{U}$ and $\mathbf{V}$ are square and orthonormal, they are invertible. Therefore,

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \text{rank}(\mathbf{\Sigma})$$

which is equal to the number of nonzero singular values of $\mathbf{A}$. ■

2. $\mathcal{C}(\mathbf{A}^T) = span\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r\}$
3. $Null(\mathbf{A}) = span\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_d\}$

Proof. We can rewrite Equation 17 as $\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$, or equivalently:

$$\mathbf{A}\mathbf{v}_1 = \sigma_1\mathbf{u}_1 \ , \ \mathbf{A}\mathbf{v}_2 = \sigma_2\mathbf{u}_2, \ , \ ... \ , \ \mathbf{A}\mathbf{v}_r = \sigma_r\mathbf{u}_r \ , \ ... \ , \ \mathbf{A}\mathbf{v}_p = \sigma_p\mathbf{u}_p$$

where $\mathbf{v}_i$ and $\mathbf{u}_i$ are the i-th columns of $\mathbf{V}$ and $\mathbf{U}$, respectively.

If $d > n$, then the rewrite can continue outside the principal submatrix of Σ as $\mathbf{A}\mathbf{v}_k = \mathbf{0}$ for $p < k \leq d$.

Since, σ_{r+1} through σ_p are 0, it means that

$$\mathbf{A}\mathbf{v}_k = \mathbf{0}$$

for $(r+1) \leq k \leq d$. Thus, the set $\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_d\}$ is an orthonormal set for at least some subspace of $Null(\mathbf{A})$.

Any vector $\mathbf{x} \in \mathbb{R}^d$ not in this subspace must be in the orthogonal complement of this subspace in $\mathbb{R}^d$. Since $\mathbf{V} \in \mathbb{R}^{d \times d}$ is square and orthonormal, its columns create a orthonormal spanning set of $\mathbb{R}^d$. Therefore, the orthonormal complement of $span\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_d\}$ is $span\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r\}$. We can represent $\mathbf{x}$ in terms of these vectors:

$$\mathbf{x} = \sum_{k=1}^{r} \alpha_k \mathbf{v}_k$$

for $\alpha_1, ..., \alpha_r \in \mathbb{R}$.

The product with $\mathbf{A}$ gives

$$\mathbf{A}\mathbf{x} = \mathbf{A} \sum_{k=1}^{r} \alpha_k \mathbf{v}_k = \sum_{k=1}^{r} \alpha_k \mathbf{A}\mathbf{v}_k = \sum_{k=1}^{r} \alpha_k \sigma_k \mathbf{u}_k$$

Since the columns of $\mathbf{U}$ are linearly independent (moreover, orthogonal) and σ_1 through σ_r are nonzero, there are no $\{\alpha_1, ..., \alpha_r\}$ that gives the zero vector other than $\alpha_k = 0 \; \forall \; 1 \leq k \leq r$ (this is by the definition of linear independence). Therefore, there are no other nonzero vectors in $span\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r\}$ that is part of the null space of $\mathbf{A}$. Therefore, the null space of $\mathbf{A}$ must be $span\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_d\}$, and the row space must be its orthogonal complement in $\mathbb{R}^d$, which is $span\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r\}$. ■

4. $\mathcal{C}(\mathbf{A}) = span\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_r\}$
5. $Null(\mathbf{A}^T) = span\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, ..., \mathbf{v}_n\}$

The proof follows the same steps with $\mathbf{A}^T$.

3.2 Thin SVD

If $\mathbf{A}$ is not square or is less than full rank, Σ contains rows or columns of zeros that don't contribute anything to $\mathbf{A}$. For instance, consider a tall and skinny $\mathbf{A}$; in other words, $n > d$. Then, we can write Equation 17 as

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T = \mathbf{U}_1 \Sigma_1 \mathbf{V}^T \quad (21)$$

where $\mathbf{U}_1 \in \mathbb{R}^{n \times d}$, $\mathbf{U}_2 \in \mathbb{R}^{n \times n-d}$, and $\Sigma_1, \mathbf{V} \in \mathbb{R}^{d \times d}$.

If $n < d$, then

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \Sigma_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} = \mathbf{U}\Sigma_1\mathbf{V}_1^T \tag{22}$$

where $\mathbf{V}_1 \in \mathbb{R}^{d \times n}$, $\mathbf{V}_2 \in \mathbb{R}^{d \times d-n}$, and $\Sigma_1, \mathbf{U} \in \mathbb{R}^{n \times n}$.

More generally, with $\text{rank}(\mathbf{A}) = r \leq p$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^T \tag{23}$$

where $\mathbf{U}_1 \in \mathbb{R}^{n \times r}$, $\mathbf{U}_2 \in \mathbb{R}^{n \times n-r}$, $\mathbf{V}_1 \in \mathbb{R}^{d \times r}$, $\mathbf{V}_2 \in \mathbb{R}^{d \times d-r}$, and $\Sigma_1 \in \mathbb{R}^{r \times r}$.

When the full square $\mathbf{U}$ and $\mathbf{V}$ are both not used, we refer to the resulting SVD as a *thin SVD* or an *economical SVD*. In a thin SVD, we only include the left and right singular vectors that span the column and row spaces of $\mathbf{A}$.

We can also write the thin SVD (Equation 23) as a sum of rank-1 matrices that are scaled outer products between the left and right singular vectors:

$$\mathbf{A} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{24}$$

3.3 Matrix Norms in terms of Singular Values

Certain matrix norms can be written in terms of the singular values.

In particular, the matrix 2-norm is equal to the largest singular value:

$$||\mathbf{A}||_2 = \sigma_1 \tag{25}$$

Proof. The definition of a matrix 2-norm is

$$||\mathbf{A}||_2 = \max_{\mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0}} \frac{||\mathbf{A}\mathbf{x}||_2}{||\mathbf{x}||_2}$$

for $\mathbf{A} \in \mathbb{R}^{n \times d}$.

Writing $\mathbf{A}$ as its full SVD and expressing $\mathbf{x}$ in terms of the full right singular vectors of $\mathbf{A}$ (which span $\mathbb{R}^d$):

$$
\begin{aligned}
||\mathbf{A}||_2^2 &= \max_{\mathbf{x}\in\mathbb{R}^d,\mathbf{x}\neq\mathbf{0}} \frac{||\mathbf{A}\mathbf{x}||_2^2}{||\mathbf{x}||_2^2} && (26)\\
&= \max_{\mathbf{x}\in\mathbb{R}^d,\mathbf{x}\neq\mathbf{0}} \frac{||\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}||_2^2}{||\mathbf{x}||_2^2} && (27)\\
&= \max_{\mathbf{x}\in\mathbb{R}^d,\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}}{\mathbf{x}^T\mathbf{x}} && (28)\\
&= \max_{\mathbf{x}\in\mathbb{R}^d,\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{V}\mathbf{\Sigma}_d^2\mathbf{V}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} & \mathbf{U}^T\mathbf{U}=\mathbf{I} \quad & (29)\\
&= \max_{\alpha_1,\dots,\alpha_d\in\mathbb{R},\sum\alpha_i\neq 0} \frac{(\sum_{i=1}^d \alpha_i\mathbf{v}_i^T\mathbf{V})\mathbf{\Sigma}_d^2(\sum_{j=1}^d \alpha_j\mathbf{V}^T\mathbf{v}_j)}{\sum_{k=1}^d \alpha_k\mathbf{v}_k^T\sum_{l=1}^d \alpha_l\mathbf{v}_l} & \mathbf{x}=\sum_{i=1}^d \alpha_i\mathbf{v}_i \quad & (30)\\
&= \max_{\alpha_1,\dots,\alpha_d\in\mathbb{R},\sum\alpha_i\neq 0} \frac{(\sum_{i=1}^d \alpha_i\mathbf{v}_i^T\mathbf{V})\mathbf{\Sigma}_d^2(\sum_{j=1}^d \alpha_j\mathbf{V}^T\mathbf{v}_j)}{\sum_{k=1}^d \alpha_k^2} & \mathbf{v}_i^T\mathbf{v}_j=\delta_{ij} \quad & (31)\\
&= \max_{\alpha_1,\dots,\alpha_d\in\mathbb{R},\sum\alpha_i\neq 0} \frac{\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_d\end{bmatrix}\mathbf{\Sigma}_d^2\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_d\end{bmatrix}}{\sum_{k=1}^d \alpha_k^2} && (32)\\
&= \max_{\alpha_1,\dots,\alpha_d\in\mathbb{R},\sum\alpha_i\neq 0} \frac{\begin{bmatrix}(\sigma_1\alpha_1) & \dots & (\sigma_r\alpha_r) & 0 & \dots & 0\end{bmatrix}\begin{bmatrix}\sigma_1\alpha_1\\ \vdots\\ \sigma_r\alpha_r\\ 0\\ \vdots\\ 0\end{bmatrix}}{\sum_{k=1}^d \alpha_k^2} && (33)\\
&= \max_{\alpha_1,\dots,\alpha_d\in\mathbb{R},\sum\alpha_i\neq 0} \frac{\sum_{i=1}^r \sigma_i^2\alpha_i^2}{\sum_{k=1}^d \alpha_k^2} && (34)\\
&= \sigma_1^2 && (35)
\end{aligned}
$$

The maximum is obtained when $\alpha_i = 0$ for $i \neq 1$.

■

Also, the Frobenius norm is related to the sum of the square of the singular values:

$$||\mathbf{A}||_F = \sqrt{\sum_{i=1}^r \sigma_i^2} \qquad (36)$$

Proof. The definition of the Frobenius norm is

$$||\mathbf{A}||_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}^2 = \text{Tr}(\mathbf{A}^T\mathbf{A})$$

But the trace of a matrix is equal to the sum of its eigenvalues, and we saw in Equation 20 that the eigenvalues of $\mathbf{A}^T\mathbf{A}$ are the square of the singular values of $\mathbf{A}$.

■

3.4 Eckart-Young-Mirsky Theorem

A key theorem that involves SVD is the **Eckart-Young-Mirsky theorem**, which states that the best rank k approximation of a matrix is the one where its rank-1 expansion (Equation 24) is truncated at $i = k$:

Theorem 3. *For any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with rank r, let $k \leq r$ and $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i$. Then,*

$$\min_{\mathbf{B}:\text{rank}(\mathbf{B})=k} ||\mathbf{A} - \mathbf{B}||_2 = ||\mathbf{A} - \mathbf{A}_k||_2 = \sigma_{k+1} \tag{37}$$

and

$$\min_{\mathbf{B}:\text{rank}(\mathbf{B})=k} ||\mathbf{A} - \mathbf{B}||_F = ||\mathbf{A} - \mathbf{A}_k||_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i} \tag{38}$$

Proof. The proof for the 2-norm is assigned as a homework problem. Here, we will only prove the Frobenius norm version.

First, note that for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and a subspace $V \subseteq \mathbb{R}^d$ of dimension $(n-k)$ that is orthogonal to the first k singular vectors,

$$\max_{\mathbf{v}\in V, ||\mathbf{v}||_2=1} ||\mathbf{A}\mathbf{v}||_2 = \sigma_{k+1}$$

The proof is similar to the proof of the matrix 2-norm (Equation 25) but with a subspace of $\mathbb{R}^d$.

Now, we must prove the *Weyl inequality.* Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times d}$ and denote their singular values as $\sigma_i(\mathbf{X})$ and $\sigma_i(\mathbf{Y})$. Let $V_X \subseteq \mathbb{R}^d$ and $V_Y \subseteq \mathbb{R}^d$ have dimensions $(d-k)$ and $(d-l)$ and be orthogonal to the first k and l right singular vectors of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and let $W = V_X \cap V_Y$. Then,

$$\begin{aligned}
\max_{\mathbf{v}\in W, ||\mathbf{v}||_2=1} ||\mathbf{X}\mathbf{v} + \mathbf{Y}\mathbf{v}||_2 &\leq \max_{\mathbf{v}\in W, ||\mathbf{v}||_2=1} ||\mathbf{X}\mathbf{v}||_2 + ||\mathbf{Y}\mathbf{v}||_2 && \text{triangle inequality} \\
&\leq \max_{\mathbf{v}\in V_X, ||\mathbf{v}||_2=1} ||\mathbf{X}\mathbf{v}||_2 + \max_{\mathbf{v}\in V_Y, ||\mathbf{v}||_2=1} ||\mathbf{Y}\mathbf{v}||_2 \\
&\leq \sigma_{k+1}(\mathbf{X}) + \sigma_{l+1}(\mathbf{Y})
\end{aligned}$$

And note that $\dim(W) \leq (d-k) + (d-l) - d) = d - k - l$. So, by the Courant-Fischer's Min-Max theorem (proved in the next lecture):

$$\sigma_{k+l+1}(\mathbf{X}+\mathbf{Y}) = \min_{V \subseteq \mathbb{R}^d, \dim(V)=d-k-l} \max_{\mathbf{v} \in V, ||\mathbf{v}||_2=1} ||\mathbf{X}\mathbf{v} + \mathbf{Y}\mathbf{v}||_2 \tag{39}$$

$$\leq \max_{\mathbf{v} \in W, ||\mathbf{v}||_2=1} ||\mathbf{X}\mathbf{v} + \mathbf{Y}\mathbf{v}||_2 \tag{40}$$

$$\leq \sigma_{k+1}(\mathbf{X}) + \sigma_{l+1}(\mathbf{Y}) \tag{41}$$

Now, to prove the Eckart-Young-Mirsky theorem for the Frobenius norm, take $\mathbf{X} = \mathbf{B}$ and $\mathbf{Y} = \mathbf{A} - \mathbf{B}$. Applying Weyl's inequality (Equation 41):

$$\sigma_{i+k}(\mathbf{A}) \leq \sigma_{k+1}(\mathbf{B}) + \sigma_i(\mathbf{A}-\mathbf{B}) = \sigma_i(\mathbf{A}-\mathbf{B})$$

The last equality used the fact that $\text{rank}(\mathbf{B}) = k$.

Then,

$$||\mathbf{A}-\mathbf{B}||_F^2 = \sum_{i=1}^{p} \sigma_i(\mathbf{A}-\mathbf{B}) \geq \sum_{i=1}^{r-k} \sigma_{i+k}(\mathbf{A})$$

where $p = \min(n, d)$.

After showing that the lower bound is met when $\mathbf{B} = \mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i$, our proof is complete:

$$||\mathbf{A}-\mathbf{A}_k||_F^2 = ||\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i||_F^2 \tag{42}$$

$$= ||\sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i||_F^2 \tag{43}$$

$$= \sum_{i=k+1}^{r} \sigma_i^2 \tag{44}$$

$$= \sum_{i=1}^{r-k} \sigma_{i+k}^2 \tag{45}$$

■

3.5 Pseudoinverse

Recall the thin SVD for $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $\text{rank}(\mathbf{A}) = r$. (Equation 23):

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} = \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^T$$

The pseudoinverse is defined as:

$$\mathbf{A}^\dagger = \mathbf{V}_1\Sigma_1^{-1}\mathbf{U}_1^T = \sum_{i=1}^{r}\frac{1}{\sigma_i}\mathbf{v}_i\mathbf{u}_i^T \tag{46}$$

The matrices $\mathbf{A}$ and $\mathbf{A}^\dagger$ provide a bijective mapping between the row and column spaces of $\mathbf{A}$ while "zero-ing out" vectors in its null and left-null spaces. Recall that the r columns of $\mathbf{U}_1$ span $\mathcal{C}(\mathbf{A})$ and the r columns of $\mathbf{V}_1$ span $\mathcal{C}(\mathbf{A}^T)$.

To illustrate, take a vector $\mathbf{x} \in \mathcal{C}(\mathbf{A})$; it can be written as $\mathbf{x} = \sum_{i=1}^{r}\alpha_i\mathbf{u}_i$. Applying the pseudoinverse:

$$\mathbf{A}^\dagger\mathbf{x} = \sum_{i=1}^{r}(\frac{1}{\sigma_i}\mathbf{v}_i\mathbf{u}_i^T\sum_{j=1}^{r}\alpha_j\mathbf{u}_j) = \sum_{i=1}^{r}\sum_{j=1}^{r}\frac{\alpha_j}{\sigma_i}\mathbf{v}_i(\mathbf{u}_i^T\mathbf{u}_j) = \sum_{i=1}^{r}\frac{\alpha_i}{\sigma_i}\mathbf{v}_i \in \mathcal{C}(\mathbf{A}^T) \tag{47}$$

Applying $\mathbf{A}$ to this result, we recover the original vector $\mathbf{x}$:

$$\mathbf{A}\sum_{i=1}^{r}\frac{\alpha_i}{\sigma_i}\mathbf{v}_i = \sum_{j=1}^{r}\sigma_j\mathbf{u}_j\mathbf{v}_j^T\sum_{i=1}^{r}\frac{\alpha_i}{\sigma_i}\mathbf{v}_i = \sum_{i=1}^{r}\sum_{j=1}^{r}\frac{\alpha_i\sigma_j}{\sigma_i}\mathbf{u}_j(\mathbf{v}_j^T\mathbf{v}_i) = \sum_{i=1}^{r}\alpha_i\mathbf{u}_i = \mathbf{x} \tag{48}$$

In other words, since $\mathbf{x} \in \mathcal{C}(\mathbf{A})$ can be expressed as $\mathbf{x} = \mathbf{A}\mathbf{b}$ for some $\mathbf{b}$, we have

$$\mathbf{A}\mathbf{A}^\dagger\mathbf{x} = \mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{b} = \mathbf{x} = \mathbf{A}\mathbf{b} \tag{49}$$

Since b is arbitrary,

$$\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A} \tag{50}$$

And for a vector $\mathbf{y} = \sum_{i=r+1}^{n}\beta_i\mathbf{u}_i \in Null(\mathbf{A}^T)$:

$$\mathbf{A}^\dagger\mathbf{y} = \sum_{i=1}^{r}(\frac{1}{\sigma_i}\mathbf{v}_i\mathbf{u}_i^T\sum_{j=r+1}^{n}\beta_j\mathbf{u}_j) = \sum_{i=1}^{r}\sum_{j=r+1}^{n}\frac{\beta_j}{\sigma_i}\mathbf{v}_i(\mathbf{u}_i^T\mathbf{u}_j) = \mathbf{0} \tag{51}$$

Thus, for any vector $\mathbf{z} = \mathbf{x} + \mathbf{y} \in \mathbb{R}^n$, we have

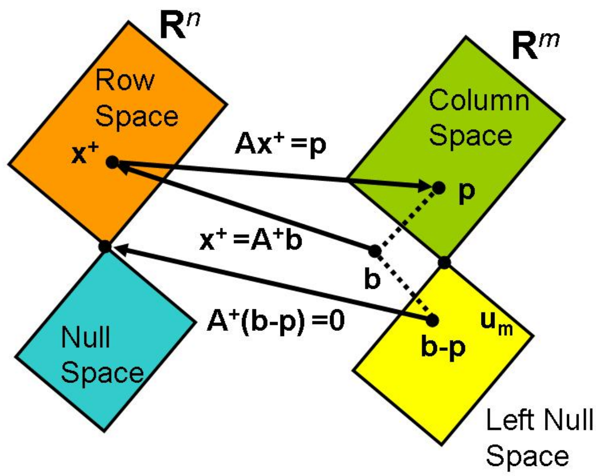
$$\begin{aligned}
\mathbf{A}\mathbf{A}^\dagger\mathbf{z} &= \mathbf{A}\mathbf{A}^\dagger(\mathbf{x}+\mathbf{y}) && (52)\\
&= \mathbf{A}\mathbf{A}^\dagger\mathbf{x} + \mathbf{A}\mathbf{A}^\dagger\mathbf{y} && (53)\\
&= \mathbf{A}\mathbf{A}^\dagger\mathbf{x} && (54)\\
&= \mathbf{x} && (55)
\end{aligned}$$

The vector $\mathbf{z}$ was projected to only the component in the column space of $\mathbf{A}$.

Similiarly, with $\mathbf{x} \in \mathcal{C}(\mathbf{A}^\dagger) = \mathcal{C}(\mathbf{A}^T)$ and $\mathbf{y} \in Null(\mathbf{A})$, we can show that

$$\mathbf{A}^\dagger \mathbf{A} \mathbf{A}^\dagger = \mathbf{A}^\dagger \tag{56}$$

and that $\mathbf{A}^\dagger \mathbf{A}$ is a projection matrix onto the row space of $\mathbf{A}$.



3.5.1 Properties of the Pseudoinverse

These properties were shown above:

1. $\mathbf{A}\mathbf{A}^\dagger \mathbf{A} = \mathbf{A}$
2. $\mathbf{A}^\dagger \mathbf{A} \mathbf{A}^\dagger = \mathbf{A}^\dagger$
3. $\mathbf{A}\mathbf{A}^\dagger$ is a projector onto $\mathcal{C}(\mathbf{A})$
4. $\mathbf{A}^\dagger \mathbf{A}$ is a projector onto $\mathcal{C}(\mathbf{A}^T)$

Some additional properties are:

5. $(\mathbf{A}^\dagger \mathbf{A})^T = \mathbf{A}^\dagger \mathbf{A}$
6. $(\mathbf{A}\mathbf{A}^\dagger)^T = \mathbf{A}\mathbf{A}^\dagger$

Proof. We proved in Theorem 2 that projection matrices are symmetric. Since $\mathbf{A}^\dagger \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\dagger$ are projection matrices, they are symmetric. ∎

And some special cases when $\mathbf{A}$ is full rank:

7. When $n \geq d$ and $\mathbf{A}$ is full rank: $\mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$

- Also called the **left inverse** because when applied to the left of $\mathbf{A}$: $\mathbf{A}^\dagger\mathbf{A} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A} = \mathbf{I}$

8. When $d \geq n$ and $\mathbf{A}$ is full rank: $\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$

   - Also called the **right inverse** because when applied to the right of $\mathbf{A}$: $\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} = \mathbf{I}$

9. When $\mathbf{A}$ is square and full rank: $\mathbf{A}^\dagger = \mathbf{A}^{-1}$

Proof. Only proving the right inverse. Others can be shown in a similar manner.

For $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $d \geq n$ and $\text{rank}(A) = n$, its full SVD is:

$$\mathbf{A} = \mathbf{U}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T \\ \mathbf{V}_2^T\end{bmatrix} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^T$$

with $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\Sigma_1 \in \mathbb{R}^{n\times n}$, and $\mathbf{V}_1 \in \mathbb{R}^{d\times n}$. Furthermore, Σ_1 is fully diagonal (and invertible), and $\mathbf{U}^T = \mathbf{U}^{-1}$ as usual.

From the definition of the pseudoinverse (Equation 46):

$$\mathbf{A}^\dagger = \mathbf{V}_1\Sigma_1^{-1}\mathbf{U}^T$$

And

$$\begin{aligned}
\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} &= \mathbf{V}_1\Sigma_1^T\mathbf{U}^T(\mathbf{U}\Sigma_1\mathbf{V}_1^T\mathbf{V}_1\Sigma_1^T\mathbf{U}^T)^{-1} && (57)\\
&= \mathbf{V}_1\Sigma_1\mathbf{U}^T(\mathbf{U}\Sigma_1^2\mathbf{U}^T)^{-1} && (58)\\
&= \mathbf{V}_1\Sigma_1\mathbf{U}^T\mathbf{U}\Sigma_1^{-2}\mathbf{U}^T && (59)\\
&= \mathbf{V}_1\Sigma_1^{-1}\mathbf{U}^T && (60)\\
&= \mathbf{A}^\dagger && (61)
\end{aligned}$$

■

3.5.2 Least Squares with the Pseudoinverse

Recall the least-squares regression problem from lecture 3:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 \tag{62}$$

given a data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with n samples $\{\mathbf{a}_i\}_{i=1}^n \in \mathbb{R}$ of d-dimensional features and a column vector $\mathbf{b} \in \mathbb{R}^n$ of targets.

We showed that the solution satisfied the normal equation:

$$\mathbf{A}^T\mathbf{A}\mathbf{x}^* = \mathbf{A}^T\mathbf{b} \tag{63}$$

For a full-rank $\mathbf{A}$ with $n \geq d$, $(\mathbf{A}^T\mathbf{A})^{-1}$ exists, and the unique solution was

$$\mathbf{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{64}$$

However, if $\mathbf{A}$ is not full rank and/or $d > n$, then there can be more than one solution to the normal equation. But, out of these solutions, $\mathbf{x}^* = \mathbf{A}^\dagger b$ gives the **minimum norm solution**. The proof is shown below.

Consider the SVD of $\mathbf{A}$ from Equation 23:

$$\mathbf{A} = \mathbf{U\Sigma V}^T = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T\\ \mathbf{V}_2^T\end{bmatrix} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^T \tag{65}$$

We express any $\mathbf{x} \in \mathbb{R}^d$ in terms of the full right singular vectors of $\mathbf{A}$, which span $\mathbb{R}^d$:

$$\mathbf{x} = \mathbf{V}\mathbf{y} = [\mathbf{V}_1, \mathbf{V}_2]\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} \tag{66}$$

Using these, we evaluate the loss function

$$\begin{aligned}||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 &= ||\mathbf{U}\Sigma\mathbf{V}^T[\mathbf{V}_1, \mathbf{V}_2]\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} - \mathbf{b}||_2^2 &(67)\\ &= ||\mathbf{U}\Sigma\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} - \mathbf{b}||_2^2 &(68)\end{aligned}$$

Since multiplying by an orthonormal matrix does not change the norm, we left multiply by $\mathbf{U}^T$:

$$\begin{aligned}||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 &= ||\mathbf{U}^T(\mathbf{A}\mathbf{x} - \mathbf{b})||_2^2 &(69)\\ &= ||\Sigma\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} - \mathbf{U}^T\mathbf{b}||_2^2 &(70)\\ &= ||\begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} - \begin{bmatrix}\mathbf{U}_1^T\\ \mathbf{U}_2^T\end{bmatrix}\mathbf{b}||_2^2 &(71)\\ &= ||\mathbf{\Sigma}_1\mathbf{y}_1 - \mathbf{U}_1^T\mathbf{b}||_2^2 + ||\mathbf{0}\cdot\mathbf{y}_2 - \mathbf{U}_2^T\mathbf{b}||_2^2 &(72)\end{aligned}$$

To find the minimum of this loss function, we consider the terms one by one. The first term is 0 when

$$\mathbf{y}_1 = \mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{b} \tag{73}$$

However, the second term is always $||\mathbf{U}_2^T\mathbf{b}||_2^2$ regardless of $\mathbf{y}_2$.

Therefore, the least-squares solutions are:

$$\mathbf{x} = \mathbf{V}\mathbf{y} = [\mathbf{V}_1, \mathbf{V}_2]\begin{bmatrix}\mathbf{y}_1 \\ \mathbf{y}_2\end{bmatrix} \tag{74}$$

$$= \mathbf{V}_1\mathbf{y}_1 + \mathbf{V}_2\mathbf{y}_2 \tag{75}$$

$$= \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{b} + \mathbf{V}_2\mathbf{y}_2 \tag{76}$$

$$= \mathbf{A}^\dagger\mathbf{b} + \mathbf{V}_2\mathbf{y}_2 \tag{77}$$

We know that $\mathbf{A}^\dagger\mathbf{b} \in \mathcal{C}(\mathbf{A}^T)$ and $\mathbf{V}_2\mathbf{y}_2 \in Null(\mathbf{A})$. Therefore, the least squares solutions are of the form:

$$\mathbf{A}^\dagger\mathbf{b} + \mathbf{w} \text{ where } \mathbf{w} \in Null(\mathbf{A}) \tag{78}$$

The solution with the smallest norm is obtained when $\mathbf{w} = \mathbf{0}$, and the minimum norm solution to the least-squares regression problem is

$$\mathbf{x}_{LS} = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{b} = \mathbf{A}^\dagger\mathbf{b} \tag{79}$$