IFT 6760A - Lecture 3
Linear Algebra Refresher

Scribe(s): Tayssir Doghri, Bogdan Mazoure **Instructor:** Guillaume Rabusseau

1 Summary

In the previous lecture, we have introduced one of the matrix decompositions called the singular value decomposition. Then, we have introduced some definitions related to orthogonality and projections such that orthonormal basis and orthogonal matrix.
In this lecture, we will continue to introduce some notions related to orthogonality and projections which are orthogonal complement and orthogonal projection. Then we will present another matrix decomposition called the QR decomposition along with an application in linear regression. In addition, we will present some basics about matrix inverse. Finally, we will introduce eigenvalues and eigenvectors.

2 Orthogonality and Projections (continued)

Definition 1 (Orthogonal complement)**.** *If $\mathcal{U}$ is a subspace of $\mathbb{R}^n$, the* orthogonal complement *of $\mathcal{U}$ is defined as*

$$\mathcal{U}^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \langle \mathbf{u}, \mathbf{v} \rangle = 0 \quad \forall \mathbf{u} \in \mathcal{U}\}$$

A graphical illustration of the orthogonal complement can be found in figure 1.

The orthogonal complement $\mathcal{U}^\perp$ is a subspace of $\mathbb{R}^n$ and we can define $\mathbb{R}^n$ as a direct sum between $\mathcal{U}$ and $\mathcal{U}^\perp$, i.e, $\mathbb{R}^n = \mathcal{U} \oplus \mathcal{U}^\perp$.
If $U \in \mathbb{R}^{n \times k}$ is orthogonal then for all $\mathbf{x} \in \mathcal{R}(\mathbf{U})$ we have the following property: $\mathbf{U}\mathbf{U}^T\mathbf{x} = \mathbf{x}$.

Proof. We want to show that for all $\mathbf{x} \in \mathcal{R}(\mathbf{U})$, we have $\mathbf{U}\mathbf{U}^T\mathbf{x} = \mathbf{x}$. Let $\mathbf{x} \in \mathcal{R}(\mathbf{U})$ then there exists $\mathbf{a} \in \mathbb{R}^k$ such that $\mathbf{x} = \mathbf{U}\mathbf{a}$. Thus, $\mathbf{U}\mathbf{U}^T\mathbf{x} = \mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{a} = \mathbf{U}\mathbf{a} = \mathbf{x}$ since we have $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. □

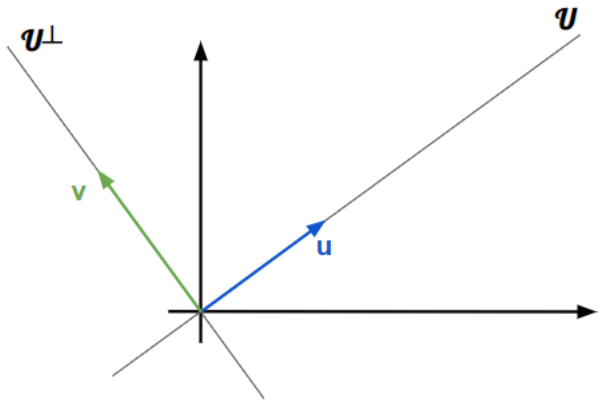

Figure 1: Illustration of the orthogonal complement of U

Definition 2 (Orthogonal projection)**.** *Let* $\mathbf{u}_1, \cdots \mathbf{u}_k \in \mathbb{R}^n$ *an orthonormal basis of a subspace* $\mathcal{U}$ *and* $\mathbf{U} \in \mathbb{R}^{n \times k}$ *such that*

$$\mathbf{U} = \begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_k \\ | & & | \end{bmatrix}$$

The orthogonal projection *onto* $\mathcal{U}$ *is defined as:*

$$\begin{aligned} \Pi_u : \mathbb{R}^n &\to \mathbb{R}^n \\ \mathbf{x} &\mapsto \mathbf{U}\mathbf{U}^T\mathbf{x} \end{aligned}$$

A graphical illustration of the orthogonal projection can be found in figure 2.

Proof. We need to show that Π_u is well-defined, i.e, if $\mathbf{V} \in \mathbb{R}^{n \times k}$ is orthogonal and $\mathcal{R}(\mathbf{V}) = \mathcal{R}(\mathbf{U})$ then $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T$. Let $\mathbf{v}_1, \cdots \mathbf{v}_k$ another orthogonal basis of $\mathcal{U}$ and $\mathbf{V} \in \mathbb{R}^{n \times k}$ such that

$$\mathbf{V} = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_k \\ | & & | \end{bmatrix}$$

We have $\mathcal{R}(\mathbf{V}) = \mathcal{R}(\mathbf{U})$ then there exists $\mathbf{P} \in \mathbb{R}^{k \times k}$ such that $\mathbf{V} = \mathbf{U}\mathbf{P}$.
Since:

- $\mathbf{P}$ is square
- $\mathbf{I} = \mathbf{V}^T\mathbf{V} = \mathbf{P}^T \underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}} \mathbf{P} = \mathbf{P}^T\mathbf{P}$

then $\mathbf{P}\mathbf{P}^T = \mathbf{I}$. Thus, we have $\mathbf{V}\mathbf{V}^T = \mathbf{U}\mathbf{P}\mathbf{P}^T\mathbf{U}^T = \mathbf{U}\mathbf{U}^T$ □

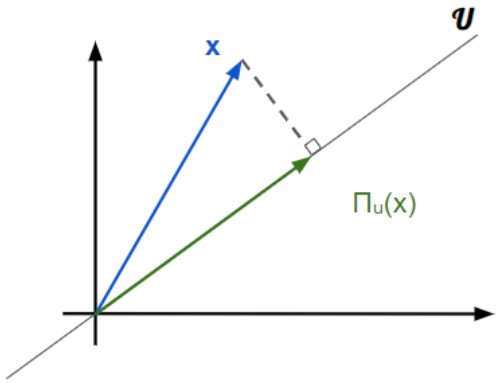

Figure 2: Illustration of the orthogonal projection of $\mathbf{x}$ onto $\mathbf{U}$

The orthogonal projection has the following properties:

- $\Pi_u^2 = \Pi_u$
- For all $\mathbf{x} \in \mathbb{R}^n$, $\langle \Pi_u(\mathbf{x}), \Pi_u(\mathbf{x}) - \mathbf{x} \rangle = 0$
- $\mathrm{Im}(\Pi_u) = \mathcal{U}$
- $\mathrm{Ker}(\Pi_u) = \mathcal{U}^\perp$

- For all $\mathbf{x} \in \mathbb{R}^n$, $\|\Pi_u(\mathbf{x})\| \leq \|\mathbf{x}\|$
- For all $\mathbf{x} \in \mathbb{R}^n$, $\arg\min_{u \in \mathbf{U}} \|\mathbf{x} - \mathbf{u}\| = \Pi_u(\mathbf{x})$

Proof. We want to show that for all $\mathbf{x} \in \mathbb{R}^n$, we have $\langle \Pi_u(\mathbf{x}), \Pi_u(\mathbf{x}) - \mathbf{x} \rangle = 0$.
Let $\mathbf{x} \in \mathbb{R}^n$ and $\Pi_u(\mathbf{x})$ its orthogonal projection onto $\mathcal{U}$, i.e, $\Pi_u(\mathbf{x}) = \mathbf{U}\mathbf{U}^T\mathbf{x}$. Then,

$$\begin{aligned} \langle \Pi_u(\mathbf{x}), \Pi_u(\mathbf{x}) - \mathbf{x} \rangle &= \langle \mathbf{U}\mathbf{U}^T\mathbf{x}, \mathbf{U}\mathbf{U}^T\mathbf{x} - \mathbf{x} \rangle \\ &= \mathbf{x}^T\mathbf{U}\underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}}\mathbf{U}^T\mathbf{x} - \mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x} \\ &= \mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x} - \mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x} \\ &= 0 \end{aligned} \tag{1}$$

□

Proof. We want to show that for all $\mathbf{x} \in \mathbb{R}^n$, we have $\arg\min_{u \in \mathbf{U}} \|\mathbf{x} - \mathbf{u}\| = \Pi_u(\mathbf{x})$.
Let $\mathbf{v} \in \mathcal{U}$, we have $\|\mathbf{x} - \mathbf{v}\|^2 = \|\mathbf{x} - \Pi_u(\mathbf{x})\|^2 + \|\mathbf{v} - \Pi_u(\mathbf{x})\|^2$. Therefore, the minimum distance between $\mathbf{x}$ and $\mathbf{v}$ is when $\mathbf{v}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathcal{U}$ which gives us $\|\mathbf{v} - \Pi_u(\mathbf{x})\|^2 = 0$ as shown in figure 3. □

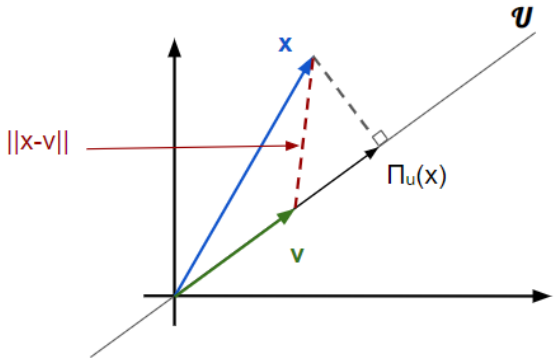

Figure 3: Illustration of the distance between $\mathbf{x}$ and $\mathbf{v}$

3 The QR decomposition

In order to solve some matrix problems, we use matrix decompositions (factorizations). In this section, we present the QR decomposition which can be used to solve the linear least squares problem for example.

Theorem 3. *Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be written as $\mathbf{A} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is orthogonal and $\mathbf{R} \in \mathbb{R}^{m \times n}$ is upper triangular. This decomposition of $\mathbf{A}$ is called the QR decomposition.*
If $m > n$ then the reduced (thin) QR decomposition *of $\mathbf{A}$ is defined as:*

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ 0 \end{bmatrix} = \mathbf{Q}_1\mathbf{R}_1$$

where $\mathbf{Q}_1 \in \mathbb{R}^{m \times n}$ is orthogonal and $\mathbf{R}_1 \in \mathbb{R}^{n \times n}$ is upper triangular.

Remark 4. *If* $\mathbf{U} \in \mathbb{R}^{n\times k}$ *and* $\text{rank}(\mathbf{U}) = k$ *then its thin QR decomposition* $\mathbf{U} = \mathbf{QR}$ *is such that:*

- $\mathcal{R}(\mathbf{Q}) = \mathcal{R}(\mathbf{U})$
- $\mathbf{R}$ *is invertible*

where $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *and* $\mathbf{R} \in \mathbb{R}^{k\times k}$

If $\mathbf{u}_1, \cdots \mathbf{u}_k \in \mathbb{R}^n$ is a basis of $\mathcal{U}$, which is **not necessarily orthonormal**, and $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that

$$\mathbf{U} = \begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_k \\ | & & | \end{bmatrix}$$

then we have the following property: $\Pi_u(\mathbf{x}) = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{x}$

Proof. In order to show the previous property, let's consider the thin QR decomposition of $\mathbf{U}$, i.e, $\mathbf{U} = \mathbf{QR}$ where $\mathbf{Q} \in \mathbb{R}^{n\times k}$ is orthogonal and $\mathbf{R} \in \mathbb{R}^{k\times k}$ is upper triangular and invertible. We have

$$\begin{aligned} \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{x} &= \mathbf{Q}\underbrace{\mathbf{R}(\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T}_{\mathbf{I}}\mathbf{Q}^T\mathbf{x} \\ &= \mathbf{Q}\mathbf{Q}^T\mathbf{x} \\ &= \Pi_u(\mathbf{x}) \end{aligned} \tag{2}$$

□

4 Linear regression

In the context of statistical learning theory, we are often interested in fitting the best model to a training set (i.e. perform regression). Formally, we aim to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ from a set of examples which has the following form: $\mathcal{D} = \{(\mathbf{x}_1, y_1), .., (\mathbf{x}_N, y_N)\}$ with $y_i \approx f(\mathbf{x}_i), \forall i = 1, 2, .., N$.
Suppose the function f takes a linear form. That is, $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ for some weight vector $\mathbf{w} \in \mathbb{R}^n$. One plausible approach to learning this function is by minimizing the Squared Error (SE) loss on an observed dataset $\mathcal{D}$:

$$\mathbf{w}^* = \underset{w\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{N} (\mathbf{w}^T\mathbf{x}_i - y_i)^2 \tag{3}$$

If we take

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \in \mathbb{R}^{N\times d}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \tag{4}$$

then (3) can be written in matrix form as

$$\mathbf{w}^* = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2, \tag{5}$$

in which case $\mathbf{X}\mathbf{w} \in \mathcal{R}(\mathbf{X})$.
In fact, finding a solution to the linear regression problem can be seen as projecting the dataset onto the hyperplane spanned by $\mathbf{X}$. For instance, assuming the rank of $\mathbf{X}$ is d

$$\begin{aligned} \hat{\mathbf{y}} &= \underset{\mathbf{v}\in\mathcal{R}(\mathbf{X})}{\arg\min} ||\mathbf{v} - \mathbf{y}||^2 \\ &= \Pi_{\mathbf{X}}(y) \\ &= \mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}_{\mathbf{w}^*}, \end{aligned} \tag{6}$$

5 Matrix inverses and pseudo-inverses

Definition 5 (Matrix inversion). *A matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is invertible if $\exists \mathbf{A}^{-1} \in \mathbb{R}^{m \times m}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.*

The matrix inverse has the following properties:

- $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$
- $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

The following statements regarding the matrix inverse are equivalent:

- $\det(\mathbf{A}) \neq 0$
- $\mathbf{A}^{-1}$ exists
- $\mathbf{A}$ has full rank
- $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$

Definition 6 (Moore-Penrose pseudo-inverse). *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and let $\mathbf{A} = \underbrace{\mathbf{U}}_{m \times R} \mathbf{D} \underbrace{\mathbf{V}^T}_{R \times m}$ be a truncated SVD where $R = rank(\mathbf{A})$.*
Then, the Moore-Penrose pseudo-inverse of $\mathbf{A}$ is defined as $\mathbf{A}^\dagger = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T$.

The Moore-Penrose pseudo-inverse has the following properties:

- $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$
- $\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger$
- $(\mathbf{A}\mathbf{A}^\dagger)^T = \mathbf{A}\mathbf{A}^\dagger$
- $(\mathbf{A}^\dagger\mathbf{A})^T = \mathbf{A}^\dagger\mathbf{A}$

Suppose (as a special case) that $\text{rank}(\mathbf{A}) = m$. Then,

$$\begin{aligned} \mathbf{A}\mathbf{A}^\dagger &= (\mathbf{U}\mathbf{D}\mathbf{V}^T)(\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T) \\ &= \mathbf{U}\mathbf{U}^T \\ &= \mathbf{I}, \end{aligned} \tag{7}$$

where the last equality holds since $\text{rank}(\mathbf{A}) = m$ hence $\mathbf{U}$ is square and $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U}$.
However, this simplification does not hold for $\mathbf{A}^\dagger\mathbf{A}$. For instance, if $m < n$, then

$$\begin{aligned} \mathbf{A}^\dagger\mathbf{A} &= (\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T)(\mathbf{U}\mathbf{D}\mathbf{V}^T) \\ &= \mathbf{V}\mathbf{V}^T \neq \mathbf{I} \end{aligned} \tag{8}$$

6 Eigenvalues

Definition 7 (Eigenvalue, eigenvector and eigenspace). *Let $\mathbf{A} \in \mathbb{R}^{m \times m}$. Any $\mathbf{v} \in \mathbb{R}^m$ such that $\mathbf{v} \neq 0$ and satisfying*

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

for $\lambda \in \mathbb{C}$ is called an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ. The space $E_\lambda = \{\mathbf{v} \in \mathbb{R}^m | \mathbf{A}\mathbf{v} = \lambda\mathbf{v}\}$ is called the eigenspace of $\mathbf{A}$ corresponding to λ.

For example, if $\mathbf{A} = \mathbf{I}$ then $\mathbf{Av} = \mathbf{Iv} = \mathbf{v}$ for all $\mathbf{v}$ and 1 is an eigenvalue with corresponding eigenspace $E_1 = \mathbb{R}^m$. Eigenvalues can be found by finding the roots of the characteristic polynomial:

$$\begin{aligned} \mathbf{Av} = \lambda\mathbf{v} &\iff (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0 \\ &\iff \mathbf{v} \in \mathcal{N}(\mathbf{A} - \lambda\mathbf{I}) \\ &\iff \det(\mathbf{A} - \lambda\mathbf{I}) = 0 \end{aligned} \tag{9}$$

As an example, let's find the eigenvalues for a given matrix $\mathbf{A}$.

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$$

It's characteristic polynomial is

$$\det\left(\begin{pmatrix} 1-\lambda & 1 \\ 0 & 2-\lambda \end{pmatrix}\right) = (1-\lambda)(2-\lambda)$$

which implies that the eigenvalues are $\lambda \in \{1, 2\}$.
As a special case, if $\mathbf{A}$ is triangular, then its determinant is the product of its eigenvalues and the eigenvalues are the diagonal entries of $\mathbf{A}$.

Definition 8 (Diagonalization)**.** *A matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is diagonalizable iff there exists a basis $\mathbf{v}_1, .., \mathbf{v}_n$ of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.*
In this case, $\mathbf{V}^{-1}\mathbf{AV} = \mathbf{D}$ is diagonal.

An example of a matrix diagonalizable over $\mathbb{C}$ but not over $\mathbb{R}$ is

$$\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

For instance, consider the problem of finding the eigenvectors of $\mathbf{A}$. Simplifying the characteristic polynomial equation implies that we have to solve $\lambda^2 + 1 = 0$. The equation has no real roots but has two complex roots, $\lambda \in \{i, -i\}$, which allows for $\mathbf{A}$ to be diagonalizable over $\mathbb{C}$.
An example of a non-diagonalizable matrix is

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$