Preliminaries

Statistical Machine Learning (ENGG*6600*02)

School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh
Summer 2023

Dataset, Learning Model, and Learning Tasks

Dataset

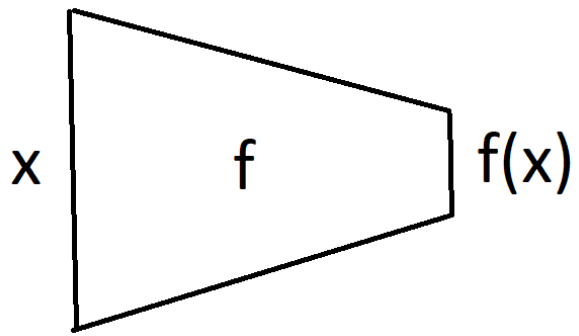
- Consider the measurement of a quantity. This quantity can be:
  - personal health data, including blood pressure, blood sugar, and blood fat,
  - images from a specific scene but taken from different perspectives,
  - images from several categories of animals, such as cat, dog, frog, etc.,
  - medical images, such as digital pathology image patches, including both healthy and tumorous tissues,
  - or any other measured signal.
- The quantity can be multidimensional, i.e., a set of values, and therefore, every quantity can be considered a multidimensional data point in a Euclidean space.
- Let the dimensionality of this space be d, meaning that every quantity is a d-dimensional vector, or data point, in $\mathbb{R}^d$. The set of d values for the quantity can be called **features** of the quantity.
- Multiple measurements of a quantity can exist, each of which is a d-dimensional data point. Therefore, there will be a set of d-dimensional data points, called a **dataset**.
- For example, the quantity can be an image, whose features are its pixels. The dataset can be a set of images from a specific scene but with different perspectives and angles.

Learning Model

- Consider a dataset of n data points $\{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, each of which is a d-dimensional vector in the d-dimensional Euclidean space. We can put these vectors column-wise in a matrix $\boldsymbol{X} \in \mathbb{R}^{d \times n}$.
- Consider a learning model f which is a map from data space to some output space:

$$\begin{aligned} f &: \mathbb{R}^d \to \mathbb{R}^p, \\ f &: \boldsymbol{x} \mapsto f(\boldsymbol{x}). \end{aligned} \tag{1}$$

- Usually, $p \leq d$ but not necessarily.



Learning Tasks

- Learning model is like a new-born baby (it first knows nothing and we should teach it)
- **Supervised**:
  - **Regression**: example of learning EMG signals for artificial leg, example of weather prediction

$$f(\boldsymbol{x}) \in [0,1]^p \text{ or } f(\boldsymbol{x}) \in \mathbb{R}.$$

  - **Classification**: example of teaching apples and cucumbers to a baby

$$f(\boldsymbol{x}) \in \{\ell_1, \ell_2, \ldots, \ell_m\}.$$

- **Unsupervised**:
  - **Clustering**: example of clustering apples and cucumbers by a baby

$$f(\boldsymbol{x}) \in \{\ell_1, \ell_2, \ldots, \ell_m\}.$$

- **Environment (world)**:
  - **Reinforcement learning**: example of teaching a dog

$$f(\boldsymbol{x}) = a \in \mathcal{A},$$

where $\mathcal{A}$ is the set of possible actions.

Learning Tasks

- **Dimensionality reduction (manifold learning)**: learning an embedding space

$$f : \mathbb{R}^d \rightarrow \mathbb{R}^p.$$

  where $p \leq d$, and usually $p \ll d$.
  - **Unsupervised** dimensionality reduction: embedding similar patterns close to each other
  - **Supervised** dimensionality reduction: Decreasing the intra-class variances and increase the inter-class variances
  - For more information on dimensionality reduction , you can see our textbook [1]: `https://link.springer.com/book/10.1007/978-3-031-10602-6`
- **Numerosity processing**:
  - **Outlier (anomaly) detection**: detecting outliers in data
  - **Prototype selection** [2]: selecting important instances
  - **Prototype generation** [3]: selecting and generating important instances
  - For more information on dimensionality reduction and numerosity processing, you can see my PhD thesis: [4]: `https://uwspace.uwaterloo.ca/handle/10012/16813`

Other Fields of AI

Some other fields of Artificial Intelligence (AI):

- Soft computing:
  - Fuzzy logic and fuzzy control
  - Metaheuristic optimization and intelligent search
- Biological-inspired (third generation) neural networks - relation to neuroscience and cognitive science - Example: spiking neural network
- Feature engineering (pre-processing):
  - Feature selection
  - Feature extraction (dimensionality reduction)
- Application of AI in various fields of science and technology

Expectation, Variance, Covariance, Bias, MSE

Random Variable, PDF, PMF, CDF

- **Random variable** X is a mathematical formalization of a quantity which depends on random events.
  - **Discrete** random variable: a countable set of possible values, e.g., $\{x_1, \ldots, x_m\}$
  - **Continuous** random variable: an uncountable set of possible values, e.g., $(0, 1]$
- **Probability mass function (PMF)** shows the distribution of a discrete random variable:

$$f(x) = \mathbb{P}(X = x), \quad x \in \{x_1, \ldots, x_m\}. \tag{2}$$

- **Probability density function (PDF)** shows the distribution of a continuous random variable:

$$f(x) = \lim_{\Delta x \to 0} \frac{\mathbb{P}(x < X < x + \Delta x)}{\Delta x}. \tag{3}$$

- **Cumulative distribution function (CDF)** shows the cumulative probabilities until that value:

$$F(x) = \mathbb{P}(X \leq x). \tag{4}$$

Expectation

- Let X be a random variable with probability mass/density function $f(x)$.
- **Expectation** or **expected value** is the mean of distribution.
- Expectation for discrete data, $x \in \{x_1, x_2, \ldots, x_m\}$:

$$\mathbb{E}(X) = \sum_{i=1}^{m} x_i \, f(x_i) = x_1 \, f(x_1) + \cdots + x_m \, f(x_m). \tag{5}$$

- Expectation for continuous data, $x \in \mathcal{D}$:

$$\mathbb{E}(X) = \int_{\mathcal{D}} x \, f(x) \, dx. \tag{6}$$

- If $h(x)$ is a function on X, then:

$$\mathbb{E}(h(x)) = \int_{\mathcal{D}} h(x) \, f(x) \, dx \tag{7}$$

- Expectation is a linear operator:

$$\mathbb{E}\Big(\sum_{i=1}^{k} a_i X_i \Big) = \sum_{i=1}^{k} a_i \, \mathbb{E}(X_i). \tag{8}$$

Variance

- Assume we have variable X and we estimate it. Let the random variable $\widehat{X}$ denote the estimate of X. Let $\mathbb{E}(\cdot)$ and $\mathbb{P}(\cdot)$ denote expectation and probability, respectively.
- The **variance** of estimating this random variable is defined as:

$$\mathbb{V}\text{ar}(\widehat{X}) := \mathbb{E}\big((\widehat{X} - \mathbb{E}(\widehat{X}))^2\big), \tag{9}$$

  which means average deviation of $\widehat{X}$ from the mean of our estimate, $\mathbb{E}(\widehat{X})$, where the deviation is squared for symmetry of difference.
- This variance can be restated as:

$$\mathbb{V}\text{ar}(\widehat{X}) = \mathbb{E}(\widehat{X}^2) - (\mathbb{E}(\widehat{X}))^2. \tag{10}$$

- Proof:

$$\begin{aligned}
\mathbb{V}\text{ar}(\widehat{X}) &= \mathbb{E}\big(\widehat{X}^2 + (\mathbb{E}(\widehat{X}))^2 - 2\widehat{X}\mathbb{E}(\widehat{X})\big) \\
&\overset{(a)}{=} \mathbb{E}(\widehat{X}^2) + (\mathbb{E}(\widehat{X}))^2 - 2\mathbb{E}(\widehat{X})\mathbb{E}(\widehat{X}) \\
&= \mathbb{E}(\widehat{X}^2) - (\mathbb{E}(\widehat{X}))^2,
\end{aligned}$$

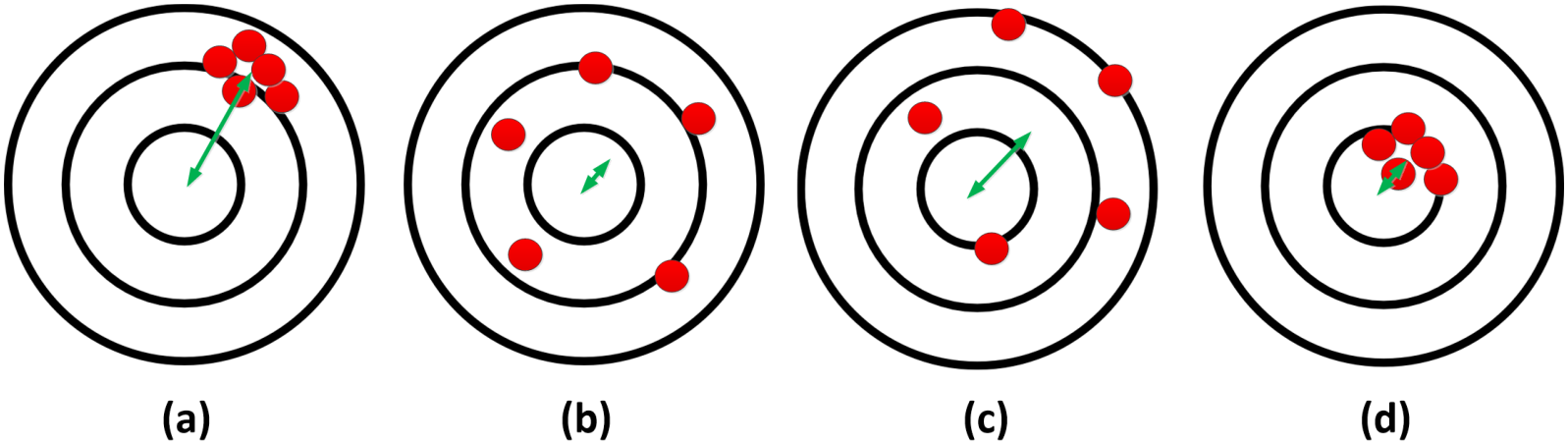
  where (a) is because expectation is a linear operator and $\mathbb{E}(\widehat{X})$ is not a random variable.

Bias

- Our estimation can have a bias. The **bias** of our estimate is defined as:

$$\mathbb{B}\text{ias}(\widehat{X}) := \mathbb{E}(\widehat{X}) - X, \tag{11}$$

  which means how much the mean of our estimate deviates from the original X.
- If the bias of an estimator is zero, i.e., $\mathbb{E}(\widehat{X}) = X$, the estimator is **unbiased**. Otherwise, it is **biased**.
- The trade-off of bias and variance (dart example):



Mean Squared Error (MSE)

- The **Mean Squared Error (MSE)** of our estimate, $\widehat{X}$, is defined as:

$$\mathrm{MSE}(\widehat{X}) := \mathbb{E}\big((\widehat{X} - X)^2\big), \tag{12}$$

which means how much our estimate deviates from the original X.

- The relation of MSE, variance, and bias is as follows:

$$\mathrm{MSE}(\widehat{X}) = \mathbb{V}\mathrm{ar}(\widehat{X}) + (\mathbb{B}\mathrm{ias}(\widehat{X}))^2. \tag{13}$$

- Proof:

$$\begin{aligned}
\mathrm{MSE}(\widehat{X}) &= \mathbb{E}\big((\widehat{X} - X)^2\big) = \mathbb{E}\big((\widehat{X} - \mathbb{E}(\widehat{X}) + \mathbb{E}(\widehat{X}) - X)^2\big) \\
&= \mathbb{E}\big((\widehat{X} - \mathbb{E}(\widehat{X}))^2 + (\mathbb{E}(\widehat{X}) - X)^2 + 2(\widehat{X} - \mathbb{E}(\widehat{X}))(\mathbb{E}(\widehat{X}) - X)\big) \\
&\overset{(a)}{=} \mathbb{E}\big((\widehat{X} - \mathbb{E}(\widehat{X}))^2\big) + (\mathbb{E}(\widehat{X}) - X)^2 + 2\underbrace{(\mathbb{E}(\widehat{X}) - \mathbb{E}(\widehat{X}))}_{0}(\mathbb{E}(\widehat{X}) - X) \\
&\overset{(b)}{=} \mathbb{V}\mathrm{ar}(\widehat{X}) + (\mathbb{B}\mathrm{ias}(\widehat{X}))^2,
\end{aligned}$$

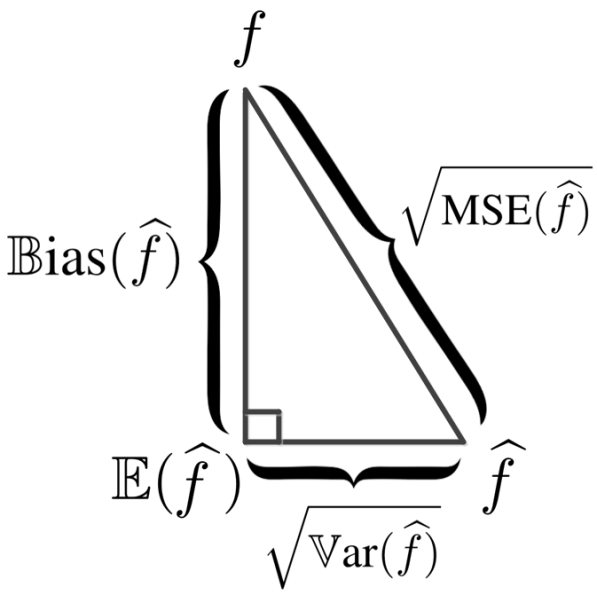
where (a) is because expectation is a linear operator and X and $\mathbb{E}(\widehat{X})$ are not random, and (b) is because of Eqs. (9) and (11).

Relation of Variance, Bias, and MSE

- The relation of MSE, variance, and bias is as follows:

$$\mathrm{MSE}(\widehat{X}) = \mathbb{V}\mathrm{ar}(\widehat{X}) + (\mathbb{B}\mathrm{ias}(\widehat{X}))^2.$$

- According to Pythagorean theorem:



Covariance

- **Covariance** is defined as:

$$\mathbb{C}\mathrm{ov}(\widehat{X}, \widehat{Y}) := \mathbb{E}\Big(\big(\widehat{X} - \mathbb{E}(\widehat{X})\big) \big(\widehat{Y} - \mathbb{E}(\widehat{Y})\big) \Big). \tag{14}$$

- We can restate it as:

$$\mathbb{C}\mathrm{ov}(\widehat{X}, \widehat{Y}) = \mathbb{E}(\widehat{X}\,\widehat{Y}) - \mathbb{E}(\widehat{Y})\,\mathbb{E}(\widehat{Y}). \tag{15}$$

- Proof:

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\widehat{X}, \widehat{Y}) &= \mathbb{E}\Big(\big(\widehat{X} - \mathbb{E}(\widehat{X})\big) \big(\widehat{Y} - \mathbb{E}(\widehat{Y})\big) \Big) \\
&= \mathbb{E}\Big(\widehat{X}\widehat{Y} - \widehat{X}\mathbb{E}(\widehat{Y}) - \mathbb{E}(\widehat{X})\widehat{Y} + \mathbb{E}(\widehat{X})\mathbb{E}(\widehat{Y}) \Big) \\
&= \mathbb{E}(\widehat{X}\widehat{Y}) - \mathbb{E}\big(\widehat{X}\mathbb{E}(\widehat{Y})\big) - \mathbb{E}\big(\mathbb{E}(\widehat{X})\widehat{Y}\big) + \mathbb{E}\big(\mathbb{E}(\widehat{X})\mathbb{E}(\widehat{Y})\big) \\
&= \mathbb{E}(\widehat{X}\widehat{Y}) - \mathbb{E}(\widehat{X})\mathbb{E}(\widehat{Y}) - \mathbb{E}(\widehat{X})\mathbb{E}(\widehat{Y}) + \mathbb{E}(\widehat{X})\mathbb{E}(\widehat{Y}) \\
&= \mathbb{E}(\widehat{X}\widehat{Y}) - \mathbb{E}(\widehat{X})\mathbb{E}(\widehat{Y}).
\end{aligned}$$

Variance and Covariance

- If we have two random variables $\widehat{X}$ and $\widehat{Y}$, we can say:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(a\widehat{X}+b\widehat{Y}) &\overset{(10)}{=} \mathbb{E}\big((a\widehat{X}+b\widehat{Y})^2\big) - \big(\mathbb{E}(a\widehat{X}+b\widehat{Y})\big)^2 \\
&\overset{(a)}{=} a^2\,\mathbb{E}(\widehat{X}^2) + b^2\,\mathbb{E}(\widehat{Y}^2) + 2ab\,\mathbb{E}(\widehat{X}\widehat{Y}) - a^2\,(\mathbb{E}(\widehat{X}))^2 - b^2\,(\mathbb{E}(\widehat{Y}))^2 - 2ab\,\mathbb{E}(\widehat{Y})\mathbb{E}(\widehat{Y}) \\
&\overset{(10)}{=} a^2\,\mathbb{V}\mathrm{ar}(\widehat{X}) + b^2\,\mathbb{V}\mathrm{ar}(\widehat{X}) + 2ab\,\mathbb{C}\mathrm{ov}(\widehat{X},\widehat{Y}),
\end{aligned} \tag{16}$$

  where (a) is for linearity of expectation.

- If the two random variables are independent, i.e., $X \perp\!\!\!\perp Y$, we have:

$$\begin{aligned}
\mathbb{E}(\widehat{X}\widehat{Y}) &\overset{(a)}{=} \iint \widehat{x}\widehat{y} f(\widehat{x},\widehat{y})\,d\widehat{x}\,d\widehat{y} \overset{\perp\!\!\!\perp}{=} \iint \widehat{x}\widehat{y} f(\widehat{x}) f(\widehat{y})\,d\widehat{x}\,d\widehat{y} \\
&= \int \widehat{y} f(\widehat{y}) \underbrace{\int \widehat{x} f(\widehat{x})\,d\widehat{x}}_{\mathbb{E}(\widehat{X})}\,d\widehat{y} = \mathbb{E}(\widehat{X}) \underbrace{\int \widehat{y} f(\widehat{y})\,d\widehat{y}}_{\mathbb{E}(\widehat{Y})} = \mathbb{E}(\widehat{X})\,\mathbb{E}(\widehat{Y}) \implies \mathbb{C}\mathrm{ov}(\widehat{X},\widehat{Y}) = 0,
\end{aligned} \tag{17}$$

  where (a) is according to definition of expectation. Note that Eq. (17) is not true for the reverse implication (we can prove by counterexample).

- So, if two random variables $\widehat{X}$ and $\widehat{Y}$ are independent, then:

$$X \perp\!\!\!\perp Y \implies \mathbb{E}(\widehat{X}\widehat{Y}) = \mathbb{E}(\widehat{X})\,\mathbb{E}(\widehat{Y}) \implies \mathbb{C}\mathrm{ov}(\widehat{X},\widehat{Y}) = 0. \tag{18}$$

Variance and Covariance

- We can extend Eqs. (16) and (15) to multiple random variables:

$$\mathbb{V}\mathrm{ar}\Big(\sum_{i=1}^{k} a_i X_i\Big) = \sum_{i=1}^{k} a_i^2\, \mathbb{V}\mathrm{ar}(X_i) + \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} a_i a_j \mathbb{C}\mathrm{ov}(X_i, X_j), \tag{19}$$

$$\mathbb{C}\mathrm{ov}\Big(\sum_{i=1}^{k_1} a_i X_i, \sum_{j=1}^{k_2} b_j Y_j\Big) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} a_i\, b_j\, \mathbb{C}\mathrm{ov}(X_i, Y_j), \tag{20}$$

  where a_i's and b_j's are not random.
- If X_i's are independent, we have:

$$\mathbb{V}\mathrm{ar}\Big(\sum_{i=1}^{k} a_i X_i\Big) = \sum_{i=1}^{k} a_i^2\, \mathbb{V}\mathrm{ar}(X_i). \tag{21}$$

Dependence and Correlation

- **Dependence** is any dependence of two random variables.
- **Correlation** is linear dependence.
- **Pearson correlation coefficient** is defined as:

$$\mathbb{C}\mathrm{orr}(X, Y) := \frac{\mathbb{C}\mathrm{ov}(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}}.$$

$$-1 \leq \mathbb{C}\mathrm{orr}(X, Y) \leq 1.$$

Conditional Expectation and Variance

- Conditional probability:

$$\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\,\mathbb{P}(Y) = \mathbb{P}(Y|X)\,\mathbb{P}(X). \tag{22}$$

- Bayes' rule:

$$\mathbb{P}(X|Y)\,\mathbb{P}(Y) = \mathbb{P}(Y|X)\,\mathbb{P}(X) \implies \mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\,\mathbb{P}(X)}{\mathbb{P}(Y)}. \tag{23}$$

- We can have:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y, X)}{\mathbb{P}(Y)}. \tag{24}$$

- **Law of total probability** or **marginalization**:

$$f(y) = \sum_x f(y|x)\,f(x), \tag{25}$$

$$f(y) = \int_x f(y|x)\,f(x)\,dx. \tag{26}$$

Conditional Expectation and Variance

- Conditional expectation:

$$\mathbb{E}(Y|X) := \sum_y y\, f(y|x) \tag{27}$$

- **Law of total expectation** or **Adam's law**:

$$\mathbb{E}(Y) = \mathbb{E}\big(\mathbb{E}(Y|X)\big) \tag{28}$$

- Proof:

$$\begin{aligned}\mathbb{E}(Y) &= \sum_y y\, f(y) = \sum_y y \sum_x f(y|x) f(x) = \sum_x \Big[\sum_y y\, f(y|x)\Big] f(x) \\ &= \sum_x \mathbb{E}(Y|X) f(x) = \mathbb{E}\big(\mathbb{E}(Y|X)\big).\end{aligned}$$

Conditional Expectation and Variance

- **Law of total variance** or **Eve's law**:

$$\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}\big(\mathbb{V}\mathrm{ar}(Y|X)\big) + \mathbb{V}\mathrm{ar}\big(\mathbb{E}(Y|X)\big). \tag{29}$$

- Proof:

$$\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \implies \mathbb{E}(Y^2) = \mathbb{V}\mathrm{ar}(Y) + \mathbb{E}(Y)^2.$$

By law of total expectation:

$$\begin{aligned}
&\mathbb{E}(Y^2) = \mathbb{E}\Big(\mathbb{V}\mathrm{ar}(Y|X) + \mathbb{E}(Y|X)^2\Big) \\
&\implies \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \mathbb{E}\Big(\mathbb{V}\mathrm{ar}(Y|X) + \mathbb{E}(Y|X)^2\Big) - \mathbb{E}(Y)^2 \\
&= \mathbb{E}\Big(\mathbb{V}\mathrm{ar}(Y|X) + \mathbb{E}(Y|X)^2\Big) - \mathbb{E}\big(\mathbb{E}(Y|X)\big)^2 \\
&= \mathbb{E}\big(\mathbb{V}\mathrm{ar}(Y|X)\big) + \Big(\mathbb{E}\big(\mathbb{E}(Y|X)^2\big) - \mathbb{E}\big(\mathbb{E}(Y|X)\big)^2\Big) = \mathbb{E}\big(\mathbb{V}\mathrm{ar}(Y|X)\big) + \mathbb{V}\mathrm{ar}\big(\mathbb{E}(Y|X)\big).
\end{aligned}$$

Monte Carlo Approximation

- Suppose we are considering some d-dimensional data $x \in \mathbb{R}^d$. Let $f(x)$ be the Probability Density Function (PDF) of data. Consider $h(x)$ is a function over the data x. According to definition, the expectation of function $h(x)$ over the distribution $f(x)$ and the probability of function $h(x)$ belonging to a set $\mathcal{A}$ are:

$$\mathbb{E}(h(x)) = \int h(x)\, f(x)\, dx, \tag{30}$$

$$\mathbb{P}(h(z) \in \mathcal{A}) = \int_{h(x)\in\mathcal{A}} f(x)\, dx, \tag{31}$$

  respectively.
- Using a sample of size n from distribution $f(x)$ (i.e., $\{x_1, \ldots, x_n\} \sim f(x)$), we can approximate Eqs. (30) and (31) by (**Monte Carlo approximation**):

$$\mathbb{E}(h(x)) \approx \frac{1}{n} \sum_{i=1}^{n} h(x_i), \tag{32}$$

$$\mathbb{P}(h(z) \in \mathcal{A}) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\big(h(x_i) \in \mathcal{A}\big), \tag{33}$$

  where $\mathbb{I}(\cdot)$ denotes the indicator function which is one and zero when its condition is and is not satisfied, respectively.
- As the above definition states, the MC approximation generates many samples from the distribution in order to approximate the expectation by mean (or average) of the samples. Obviously, the more the n is, the better the approximation becomes.

Sample Mean and Covariance

- Assume we have n data points $\{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$.
- Sample mean:

$$\mathbb{R}^d \ni \boldsymbol{\mu} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i. \tag{34}$$

- Sample covariance matrix:

$$\mathbb{R}^{d\times d} \ni \boldsymbol{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top. \tag{35}$$

- For this to be an unbiased sample covariance, n should be $n-1$ in the denominator:

$$\mathbb{R}^{d\times d} \ni \boldsymbol{\Sigma} := \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top. \tag{36}$$

Linear Projection

Column Space

- Consider p basis vectors. We can define a p-dimensional Euclidean space by these p basis vectors. For example, two vectors define a plane and three vectors define a 3D space.
- In terminology: The p basis vectors **span** the p-dimensional Euclidean space. Or the p-dimensional Euclidean space is **spanned by** the p basis vectors.
- Consider the p vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$. These vectors can be stacked columnwise in matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p] \in \mathbb{R}^{d \times p}$.
- The space spanned by the columns of matrix $\boldsymbol{U}$ is called the column space of matrix $\boldsymbol{U}$, denoted by $\mathbb{C}\text{ol}(\boldsymbol{U})$:

$$\mathbb{C}\text{ol}(\boldsymbol{U}) := \textbf{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}. \tag{37}$$

- In other words, the space whose bases are the columns of matrix $\boldsymbol{U}$ is called the column space of matrix $\boldsymbol{U}$.

Linear Projection

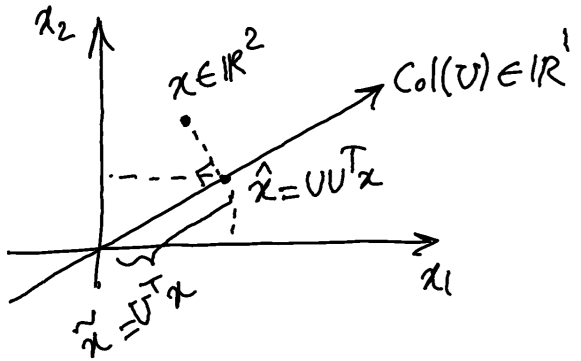
- Assume there is a data point $\boldsymbol{x} \in \mathbb{R}^d$. The aim is to project this data point onto the vector space spanned by p vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$, where each vector is d-dimensional and usually $p \ll d$.
- These vectors can be stacked columnwise in matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p] \in \mathbb{R}^{d \times p}$. In other words, the goal is to project $\boldsymbol{x}$ onto the column space of $\boldsymbol{U}$, denoted by $\mathbb{C}\text{ol}(\boldsymbol{U})$.
- As $p < d$, this projection is projection onto a **subspace** because we are projecting from d-dimensional space onto a lower dimensional space.
- The **projection** of $\boldsymbol{x} \in \mathbb{R}^d$ onto $\mathbb{C}\text{ol}(\boldsymbol{U}) \in \mathbb{R}^p$ is:

$$\mathbb{R}^p \ni \widetilde{\boldsymbol{x}} := \boldsymbol{U}^\top \boldsymbol{x}. \tag{38}$$

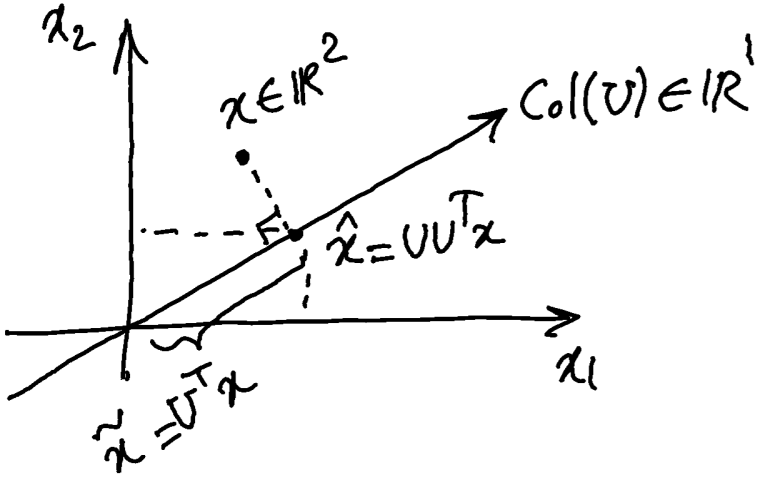
- The **reconstruction** of $\widetilde{\boldsymbol{x}} \in \mathbb{R}^p$ in the d-dimensional space is:

$$\mathbb{R}^d \ni \widehat{\boldsymbol{x}} := \boldsymbol{U}\widetilde{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}. \tag{39}$$

- Reconstruction is its representation in $\mathbb{R}^d$ again, but after projection.

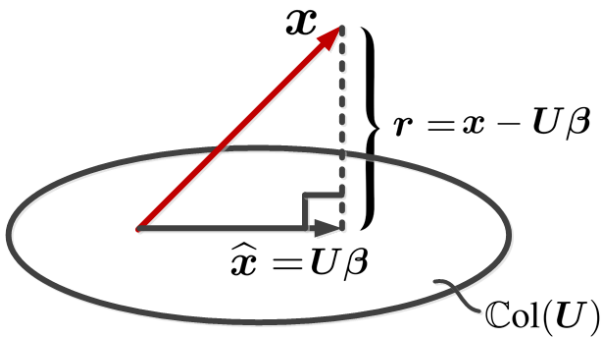


Linear Projection



- **Reconstruction error**: There is a residual/error between the original data $\boldsymbol{x}$ and its reconstruction (if the data point is already in the column space, this residual is zero):

$$\boldsymbol{r} = \boldsymbol{x} - \widehat{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}. \tag{40}$$



Centering Matrix

Centering Matrix

- Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{\alpha \times \beta}$, which is represented by its rows, $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_\alpha]^\top$ or by its columns, $\boldsymbol{A} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_\beta]$, where $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ denote the i-th row and j-th column of $\boldsymbol{A}$, respectively. Note that the vectors are column vectors.
- The **left centering matrix** is defined as:

$$\mathbb{R}^{\alpha \times \alpha} \ni \boldsymbol{H} := \boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top, \tag{41}$$

where $\mathbf{1} = [1, \ldots, 1]^\top \in \mathbb{R}^\alpha$ and $\boldsymbol{I} \in \mathbb{R}^{\alpha \times \alpha}$ is the identity matrix.

- Left multiplying this matrix by $\boldsymbol{A}$, i.e., $\boldsymbol{HA}$, removes the mean of rows of $\boldsymbol{A}$ from all of its rows:

$$\boldsymbol{HA} \overset{(41)}{=} \boldsymbol{A} - (1/\alpha)\mathbf{1}\mathbf{1}^\top \boldsymbol{A} = (\boldsymbol{A}^\top - \boldsymbol{\mu}_{\text{rows}})^\top, \tag{42}$$

where the column vector $\boldsymbol{\mu}_{\text{rows}} \in \mathbb{R}^\beta$ is the mean of the rows of $\boldsymbol{A}$.

- The **right centering matrix** is defined as:

$$\mathbb{R}^{\beta \times \beta} \ni \boldsymbol{H} := \boldsymbol{I} - (1/\beta)\mathbf{1}\mathbf{1}^\top, \tag{43}$$

where $\mathbf{1} = [1, \ldots, 1]^\top \in \mathbb{R}^\beta$ and $\boldsymbol{I} \in \mathbb{R}^{\beta \times \beta}$ is the identity matrix.

- Right multiplying this matrix to $\boldsymbol{A}$, i.e., $\boldsymbol{AH}$, removes the mean of the columns of $\boldsymbol{A}$ from all of its columns:

$$\boldsymbol{AH} \overset{(43)}{=} \boldsymbol{A} - (1/\beta)\boldsymbol{A}\mathbf{1}\mathbf{1}^\top = \boldsymbol{A} - \boldsymbol{\mu}_{\text{cols}}, \tag{44}$$

where the column vector $\boldsymbol{\mu}_{\text{cols}} \in \mathbb{R}^\alpha$ is the mean of the columns of $\boldsymbol{A}$.

Centering Matrix

- Both left and right centering matrices can be used at the same time to have a **double-centered** matrix $\boldsymbol{A}$:

$$\begin{aligned}\boldsymbol{HAH} &= (\boldsymbol{I}_\alpha - (1/\alpha)\mathbf{1}_\alpha\mathbf{1}_\alpha^\top)\boldsymbol{A}(\boldsymbol{I}_\beta - (1/\beta)\mathbf{1}_\beta\mathbf{1}_\beta^\top) \\ &= (\boldsymbol{A} - (1/\alpha)\mathbf{1}_\alpha\mathbf{1}_\alpha^\top\boldsymbol{A})(\boldsymbol{I}_\beta - (1/\beta)\mathbf{1}_\beta\mathbf{1}_\beta^\top) \\ &= \boldsymbol{A} - (1/\alpha)\mathbf{1}_\alpha\mathbf{1}_\alpha^\top\boldsymbol{A} - (1/\beta)\boldsymbol{A}\mathbf{1}_\beta\mathbf{1}_\beta^\top + (1/(\alpha\beta))\mathbf{1}_\alpha\mathbf{1}_\alpha^\top\boldsymbol{A}\mathbf{1}_\beta\mathbf{1}_\beta^\top. \qquad (45)\end{aligned}$$

- The second term removes the mean of the rows of $\boldsymbol{A}$ according to Eq. (42) and the third term removes the mean of the columns of $\boldsymbol{A}$ according to Eq. (44). The last term, however, adds the overall mean of $\boldsymbol{A}$ back to it.
- In summary:

$$\begin{aligned}\boldsymbol{HA} &\approx (\boldsymbol{A}^\top - \boldsymbol{\mu}_{\text{rows}})^\top, \\ \boldsymbol{AH} &\approx \boldsymbol{A} - \boldsymbol{\mu}_{\text{cols}}, \\ \boldsymbol{HAH} &\approx (\boldsymbol{A}^\top - \boldsymbol{\mu}_{\text{rows}})^\top - \boldsymbol{\mu}_{\text{cols}} + \boldsymbol{\mu}_{\text{all}}.\end{aligned}$$

Centering Matrix

- The centering matrix is symmetric because:

$$\boldsymbol{H}^\top = (\boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top)^\top = \boldsymbol{I}^\top - (1/\alpha)(\mathbf{1}\mathbf{1}^\top)^\top = \boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top \overset{(41)}{=} \boldsymbol{H}. \tag{46}$$

- The centering matrix is also idempotent:

$$\boldsymbol{H}^k = \underbrace{\boldsymbol{H}\boldsymbol{H}\cdots\boldsymbol{H}}_{k \text{ times}} = \boldsymbol{H}, \tag{47}$$

where k is a positive integer. Proof:

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{H} &= (\boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top)(\boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top) \\
&= \boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top - (1/\alpha)\mathbf{1}\mathbf{1}^\top + (1/\alpha^2)\mathbf{1}\underbrace{\mathbf{1}^\top\mathbf{1}}_{\alpha}\mathbf{1}^\top \\
&= \boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top - (1/\alpha)\mathbf{1}\mathbf{1}^\top + (1/\alpha)\mathbf{1}\mathbf{1}^\top = \boldsymbol{I} - (1/\alpha)\mathbf{1}\mathbf{1}^\top \overset{(41)}{=} \boldsymbol{H}.
\end{aligned}$$

Therefore:

$$\boldsymbol{H}^k = (\underbrace{\boldsymbol{H}\cdots(\boldsymbol{H}(\underbrace{\boldsymbol{H}\boldsymbol{H}}_{\boldsymbol{H}}))}_{\boldsymbol{H}}) = \boldsymbol{H}.$$

Norm

Inner product

Definition (Inner product of vectors)

Consider two vectors $\boldsymbol{x} = [x_1, \ldots, x_d]^\top \in \mathbb{R}^d$ and $\boldsymbol{y} = [y_1, \ldots, y_d]^\top \in \mathbb{R}^d$. Their **inner product**, also called **dot product**, is:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{d} x_i \, y_i.$$

Definition (Inner product of matrices)

We also have inner product between matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d_1 \times d_2}$. Let $\boldsymbol{X}_{ij}$ denote the (i,j)-th element of matrix $\boldsymbol{X}$. The inner product of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is:

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{Y}) = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \boldsymbol{X}_{i,j} \, \boldsymbol{Y}_{i,j},$$

where $\mathbf{tr}(.)$ denotes the trace of matrix.

Norm

Definition (Norm)

A function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$, $\|\cdot\| : \boldsymbol{x} \mapsto \|\boldsymbol{x}\|$ is a **norm** if it satisfies:

1. $\|\boldsymbol{x}\| \geq 0, \forall \boldsymbol{x}$
2. $\|a\boldsymbol{x}\| = |a| \, \|\boldsymbol{x}\|, \forall \boldsymbol{x}$ and all scalars a
3. $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = 0$
4. Triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

Important norms for vectors

Some important norms for a vector $\boldsymbol{x} = [x_1, \ldots, x_d]^\top$ are as follows.

- The ℓ_p **norm** is:

$$\|\boldsymbol{x}\|_p := \left(|x_1|^p + \cdots + |x_d|^p\right)^{1/p},$$

  where $p \geq 1$ and $|.|$ denotes the absolute value.

- Two well-known ℓ_p norms are ℓ_1 **norm** and ℓ_2 **norm** (also called the **Euclidean norm**) with $p = 1$ and $p = 2$, respectively:

$$\|\boldsymbol{x}\|_1 := |x_1| + \cdots + |x_d| = \sum_{i=1}^{d} |x_i|,$$

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_d^2} = \sqrt{\sum_{i=1}^{d} x_i^2},$$

Important norms for matrices

Some important norms for a matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ are as follows.

- The formulation of the **Frobenius norm** for a matrix is similar to the formulation of ℓ_2 norm for a vector:

$$\|\boldsymbol{X}\|_F := \sqrt{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \boldsymbol{X}_{i,j}^2},$$

  where $\boldsymbol{X}_{ij}$ denotes the (i,j)-th element of $\boldsymbol{X}$.

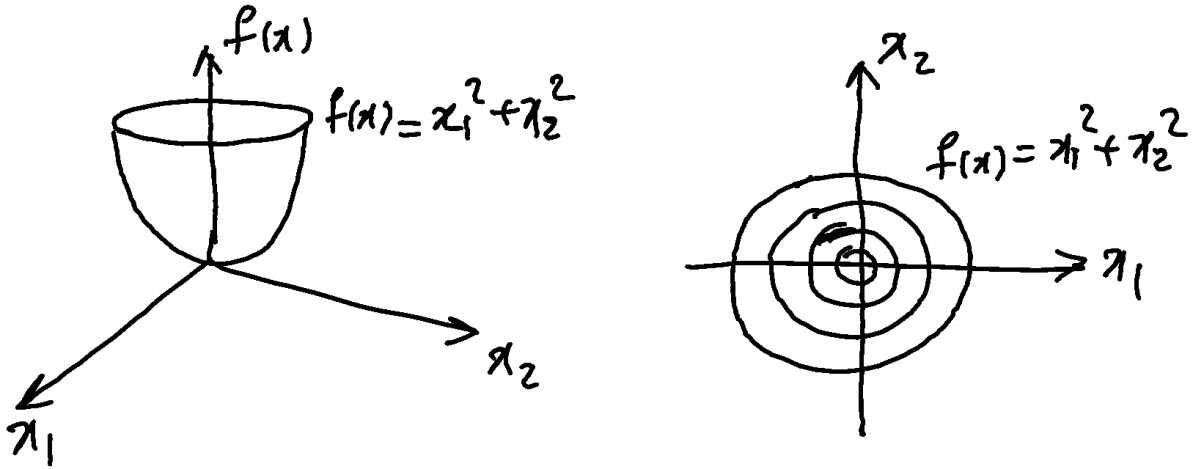
Quadratic forms using norms

For $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, we have:

$$\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \sum_{i=1}^{d} x_i^2,$$

$$\|\boldsymbol{X}\|_F^2 = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X}) = \langle \boldsymbol{X}, \boldsymbol{X} \rangle = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \boldsymbol{X}_{i,j}^2,$$

which are convex and in quadratic forms.



Preliminaries on Derivatives

Dimensionality of derivative

- Consider a function $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$.
- Derivative of function $f(\boldsymbol{x}) \in \mathbb{R}^{d_2}$ with respect to (w.r.t.) $\boldsymbol{x} \in \mathbb{R}^{d_1}$ has dimensionality $(d_1 \times d_2)$.
- This is because tweaking every element of $\boldsymbol{x} \in \mathbb{R}^{d_1}$ can change every element of $f(\boldsymbol{x}) \in \mathbb{R}^{d_2}$. The (i, j)-th element of the $(d_1 \times d_2)$-dimensional derivative states the amount of change in the j-th element of $f(\boldsymbol{x})$ resulted by changing the i-th element of $\boldsymbol{x}$.

Examples

- The derivative of a scalar w.r.t. a scalar is a scalar.
- The derivative of a scalar w.r.t. a vector is a vector.
- The derivative of a scalar w.r.t. a matrix is a matrix.
- The derivative of a vector w.r.t. a vector is a matrix.
- The derivative of a vector w.r.t. a matrix is a rank-3 tensor.
- The derivative of a matrix w.r.t. a matrix is a rank-4 tensor.

Dimensionality of derivative

In more details:

- If the function is $f : \mathbb{R} \to \mathbb{R}, f : x \mapsto f(x)$, the derivative $(\partial f(x)/\partial x) \in \mathbb{R}$ is a scalar because changing the scalar $\boldsymbol{x}$ can change the scalar $f(\boldsymbol{x})$.
- If the function is $f : \mathbb{R}^d \to \mathbb{R}, f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, the derivative $(\partial f(\boldsymbol{x})/\partial \boldsymbol{x}) \in \mathbb{R}^d$ is a vector because changing every element of the vector $\boldsymbol{x}$ can change the scalar $f(\boldsymbol{x})$.
- If the function is $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}, f : \boldsymbol{X} \mapsto f(\boldsymbol{X})$, the derivative $(\partial f(\boldsymbol{X})/\partial \boldsymbol{X}) \in \mathbb{R}^{d_1 \times d_2}$ is a matrix because changing every element of the matrix $\boldsymbol{X}$ can change the scalar $f(\boldsymbol{X})$.
- If the function is $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}, f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, the derivative $(\partial f(\boldsymbol{x})/\partial \boldsymbol{x}) \in \mathbb{R}^{d_1 \times d_2}$ is a matrix because changing every element of the vector $\boldsymbol{x}$ can change every element of the vector $f(\boldsymbol{x})$.
- If the function is $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_3}, f : \boldsymbol{X} \mapsto f(\boldsymbol{X})$, the derivative $(\partial f(\boldsymbol{X})/\partial \boldsymbol{X})$ is a $(d_1 \times d_2 \times d_3)$-dimensional tensor because changing every element of the matrix $\boldsymbol{X}$ can change every element of the vector $f(\boldsymbol{X})$.
- If the function is $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_3 \times d_4}, f : \boldsymbol{X} \mapsto f(\boldsymbol{X})$, the derivative $(\partial f(\boldsymbol{X})/\partial \boldsymbol{X})$ is a $(d_1 \times d_2 \times d_3 \times d_4)$-dimensional tensor because changing every element of the matrix $\boldsymbol{X}$ can change every element of the matrix $f(\boldsymbol{X})$.

Gradient and Hessian

Definition (Gradient)

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$. In optimizing the function f, the derivative of function w.r.t. its variable $\boldsymbol{x}$ is called the **gradient**, denoted by:

$$\nabla f(\boldsymbol{x}) := \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \in \mathbb{R}^d.$$

Definition (Hessian)

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$. The second derivative of function w.r.t. to its derivative is called the **Hessian** matrix, denoted by:

$$\boldsymbol{B} = \nabla^2 f(\boldsymbol{x}) := \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}^2} \in \mathbb{R}^{d \times d}.$$

The Hessian matrix is symmetric. If the function is convex, its Hessian matrix is positive semi-definite.

Chain rule

- When having composite functions (i.e., function of function), we use **chain rule** for derivative. Example:

$$f(x) = \sqrt{x^3 + x^2 - x + 10} = \sqrt{g(x)}, \quad g(x) = x^3 + x^2 - x + 10,$$
$$\frac{\partial f(x)}{\partial x} = \frac{\partial f(x)}{\partial g(x)} \times \frac{\partial g(x)}{\partial x} = \frac{1}{2\sqrt{g(x)}} \times (3x^2 + 2x - 1) = \frac{3x^2 + 2x - 1}{2\sqrt{x^3 + x^2 - x + 10}}$$

- The chain rule in matrix derivatives is usually stated right to left in matrix multiplications while transpose is used for matrices in multiplication.
- Let **vec**(.) denote vectorization of a $\mathbb{R}^{a \times b}$ matrix to a $\mathbb{R}^{ab}$ vector.
- Let $\textbf{vec}^{-1}_{a \times b}(.)$ be de-vectorization of a $\mathbb{R}^{ab}$ vector to a $\mathbb{R}^{a \times b}$ matrix.

Optimization

Optimization

- Lagrangian:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & y_i(\boldsymbol{x}) \leq 0,\ i \in \{1,\dots,m_1\}, \\ & h_i(\boldsymbol{x}) = 0,\ i \in \{1,\dots,m_2\}. \end{aligned}$$

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu}) := f(\boldsymbol{x}) + \sum_{i=1}^{m_1} \lambda_i y_i(\boldsymbol{x}) + \sum_{i=1}^{m_2} \nu_i h_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{y}(\boldsymbol{x}) + \boldsymbol{\nu}^\top \boldsymbol{h}(\boldsymbol{x}).$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}} \overset{\text{set}}{=} \mathbf{0} \implies \dots$$

- Unconstrained optimization:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}),$$

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} + (\Delta \boldsymbol{x})^{(k)},$$

Gradient method: $\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \eta^{(k)} \nabla f(\boldsymbol{x}^{(k)})$,

Newton's method: $\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \eta^{(k)} \left(\nabla^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \nabla f(\boldsymbol{x}^{(k)})$.

- Constrained optimization: We can use interior-point method or proximal methods, ...

Rayleigh-Ritz Quotient

Rayleigh-Ritz Quotient

- The **Rayleigh-Ritz quotient** or **Rayleigh quotient** is defined as [5, 6]:

$$\mathbb{R} \ni R(\boldsymbol{A}, \boldsymbol{x}) := \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}, \tag{48}$$

where $\boldsymbol{A}$ is a symmetric matrix and $\boldsymbol{x}$ is a nonzero vector:

$$\boldsymbol{A} = \boldsymbol{A}^\top, \quad \boldsymbol{x} \neq \boldsymbol{0}. \tag{49}$$

- One of the properties of the Rayleigh-Ritz quotient is:

$$R(\boldsymbol{A}, c\boldsymbol{x}) = R(\boldsymbol{A}, \boldsymbol{x}), \tag{50}$$

where c is a scalar. Proof:

$$R(\boldsymbol{A}, c\boldsymbol{x}) = \frac{(c\boldsymbol{x})^\top \boldsymbol{A}\, c\boldsymbol{x}}{(c\boldsymbol{x})^\top c\boldsymbol{x}} \overset{(a)}{=} \frac{c\boldsymbol{x}^\top \boldsymbol{A}\, c\boldsymbol{x}}{c\boldsymbol{x}^\top c\boldsymbol{x}} \overset{(b)}{=} \frac{c^2}{c^2} \times \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \overset{(48)}{=} R(\boldsymbol{A}, \boldsymbol{x}),$$

where (a) and (b) are because c is a scalar.

Rayleigh-Ritz Quotient

- Consider the optimization problem of the Rayleigh-Ritz quotient:

$$\underset{\boldsymbol{x}}{\text{minimize/maximize}} \quad R(\boldsymbol{A}, \boldsymbol{x}). \tag{51}$$

- According to Eq. (50), this is equivalent to the following problem [6]:

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize/maximize}} && R(\boldsymbol{A}, \boldsymbol{x}) \\ &\text{subject to} && ||\boldsymbol{x}||_2 = 1. \end{aligned} \tag{52}$$

- Proof: Let $\boldsymbol{y} := (1/||\boldsymbol{x}||_2)\,\boldsymbol{x}$. The Rayleigh-Ritz quotient is:

$$R(\boldsymbol{A}, \boldsymbol{y}) = \frac{\boldsymbol{y}^\top \boldsymbol{A}\,\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \frac{1/||\boldsymbol{x}||_2^2}{1/||\boldsymbol{x}||_2^2} \times \frac{\boldsymbol{x}^\top \boldsymbol{A}\,\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = R(\boldsymbol{A}, \boldsymbol{x}).$$

Due to the following:

$$||\boldsymbol{y}||_2^2 = \frac{1}{||\boldsymbol{x}||_2^2} \times ||\boldsymbol{x}||_2^2 = 1 \implies ||\boldsymbol{y}||_2 = 1,$$

$R(\boldsymbol{A}, \boldsymbol{y})$ should be optimized subject to $||\boldsymbol{y}||_2 = 1$. Changing the dummy variable $\boldsymbol{y}$ to $\boldsymbol{x}$ gives Eq. (52).

Rayleigh-Ritz Quotient

- Another equivalent problem for Eq. (51) is:

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize/maximize}} && \boldsymbol{x}^\top \boldsymbol{A}\, \boldsymbol{x} \\ &\text{subject to} && ||\boldsymbol{x}||_2 = 1, \end{aligned} \tag{53}$$

  obtained by inserting the constraint $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x} = 1$ in Eqs. (48) and (52).
- The constraint in Eqs. (52) and (53) can be equal to any constant (1 is a constant and disappears in derivative of Lagrangian).
- The **generalized Rayleigh-Ritz quotient** or **generalized Rayleigh quotient** is defined as [5]:

$$\mathbb{R} \ni R(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{x}) := \frac{\boldsymbol{x}^\top \boldsymbol{A}\, \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{B}\, \boldsymbol{x}}, \tag{54}$$

  where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{x}$ is a nonzero vector:

$$\boldsymbol{A} = \boldsymbol{A}^\top, \quad \boldsymbol{B} = \boldsymbol{B}^\top, \quad \boldsymbol{x} \neq \boldsymbol{0}. \tag{55}$$

Rayleigh-Ritz Quotient

- If the symmetric $\boldsymbol{B}$ is positive definite:

$$\boldsymbol{B} \succ 0, \tag{56}$$

it has a Cholesky decomposition:

$$\boldsymbol{B} = \boldsymbol{C}\boldsymbol{C}^\top, \tag{57}$$

where $\boldsymbol{C}$ is a lower triangular matrix.

- In case $\boldsymbol{B} \succ 0$, the generalized Rayleigh-Ritz quotient can be converted to a Rayleigh-Ritz quotient:

$$R(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{x}) = R(\boldsymbol{D}, \boldsymbol{C}^\top \boldsymbol{x}), \tag{58}$$

where:

$$\boldsymbol{D} := \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C}^{-\top}. \tag{59}$$

- Proof:

$$\text{RHS} = R(\boldsymbol{D}, \boldsymbol{C}^\top \boldsymbol{x}) \overset{(48)}{=} \frac{(\boldsymbol{C}^\top \boldsymbol{x})^\top \boldsymbol{D} (\boldsymbol{C}^\top \boldsymbol{x})}{(\boldsymbol{C}^\top \boldsymbol{x})^\top (\boldsymbol{C}^\top \boldsymbol{x})}$$
$$\overset{(59)}{=} \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{C}\boldsymbol{C}^{-1})^\top \boldsymbol{x}}{\boldsymbol{x}^\top (\boldsymbol{C}\boldsymbol{C}^\top)\boldsymbol{x}} \overset{(a)}{=} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}} \overset{(54)}{=} R(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{x}) = \text{LHS},$$

where RHS and LHS are short for right and left hand sides and (a) is because of Eq. (57) and $\boldsymbol{C}\boldsymbol{C}^{-1} = \boldsymbol{I}$ because $\boldsymbol{C}$ is a square matrix.

Rayleigh-Ritz Quotient

- Similarly, one of the properties of the generalized Rayleigh-Ritz quotient is:

$$R(\boldsymbol{A}, \boldsymbol{B}; c\boldsymbol{x}) = R(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{x}), \tag{60}$$

where c is a scalar. Proof:

$$R(\boldsymbol{A}, \boldsymbol{B}; c\boldsymbol{x}) = \frac{(c\boldsymbol{x})^\top \boldsymbol{A}\, c\boldsymbol{x}}{(c\boldsymbol{x})^\top \boldsymbol{B}\, c\boldsymbol{x}} \overset{(a)}{=} \frac{c\boldsymbol{x}^\top \boldsymbol{A}\, c\boldsymbol{x}}{c\boldsymbol{x}^\top \boldsymbol{B}\, c\boldsymbol{x}} \overset{(b)}{=} \frac{c^2}{c^2} \times \frac{\boldsymbol{x}^\top \boldsymbol{A}\, \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{B}\, \boldsymbol{x}} \overset{(54)}{=} R(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{x}),$$

where (a) and (b) are because c is a scalar.

- Consider the optimization problem of the generalized Rayleigh-Ritz quotient:

$$\underset{\boldsymbol{x}}{\text{minimize/maximize}} \quad R(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{x}). \tag{61}$$

According to Eq. (60), it has an equivalent form:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize/maximize}} \quad & \boldsymbol{x}^\top \boldsymbol{A}\, \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x}^\top \boldsymbol{B}\, \boldsymbol{x} = 1, \end{aligned} \tag{62}$$

for the same reason as the Rayleigh-Ritz quotient. The constraint can be equal to any constant because in the derivative of Lagrangian, the constant will be dropped.

Eigenvalue and Singular Value Decomposition

Eigenvalue Problem

- Eigenvalue and generalized eigenvalue problems play important roles in different fields of science, including machine learning, physics, statistics, and mathematics.
- In the eigenvalue problem, the eigenvectors of a matrix represent the most important and informative directions of that matrix. For example, if the matrix is a covariance matrix of data, the eigenvectors represent the directions of the spread or variance of data and the corresponding eigenvalues are the magnitude of the spread in these directions [7].
- These directions are impacted by another matrix in the generalized eigenvalue problem. If the other matrix is the identity matrix, this impact is cancelled and the eigenvalue problem captures the directions of the maximum spread.
- The **eigenvalue problem** [8, 9] of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is defined as:

$$\boldsymbol{A}\boldsymbol{\phi}_i = \lambda_i \boldsymbol{\phi}_i, \qquad \forall i \in \{1, \ldots, d\}, \tag{63}$$

and in matrix form, it is:

$$\boldsymbol{A}\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda}, \tag{64}$$

where the columns of $\mathbb{R}^{d \times d} \ni \boldsymbol{\Phi} := [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d]$ are the eigenvectors and diagonal elements of $\mathbb{R}^{d \times d} \ni \boldsymbol{\Lambda} := \mathbf{diag}([\lambda_1, \ldots, \lambda_d]^\top)$ are the eigenvalues. Note that $\boldsymbol{\phi}_i \in \mathbb{R}^d$ and $\lambda_i \in \mathbb{R}$.

Eigenvalue Problem

- For the eigenvalue problem, the matrix $\boldsymbol{A}$ can be nonsymmetric. If the matrix is symmetric, its eigenvectors are orthogonal/orthonormal and if it is nonsymmetric, its eigenvectors are not orthogonal/orthonormal.
- Equation (64) can be restated as:

$$\boldsymbol{A\Phi} = \boldsymbol{\Phi\Lambda} \implies \boldsymbol{A}\underbrace{\boldsymbol{\Phi\Phi}^\top}_{\boldsymbol{I}} = \boldsymbol{\Phi\Lambda\Phi}^\top \implies \boldsymbol{A} = \boldsymbol{\Phi\Lambda\Phi}^\top = \boldsymbol{\Phi\Lambda\Phi}^{-1}, \tag{65}$$

  where $\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^{-1}$ because $\boldsymbol{\Phi}$ is an orthogonal matrix.
- There is always $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}$ for orthogonal $\boldsymbol{\Phi}$, but there is only $\boldsymbol{\Phi\Phi}^\top = \boldsymbol{I}$ if "all" columns of the orthogonal $\boldsymbol{\Phi}$ exist (it is not truncated, i.e., it is a square matrix). Equation (65) is referred to as "**eigenvalue decomposition**", "eigen-decomposition", or "spectral decomposition".

Generalized Eigenvalue Problem

- The **generalized eigenvalue problem** [5, 9] of two symmetric matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ is defined as:

$$\boldsymbol{A}\boldsymbol{\phi}_i = \lambda_i \boldsymbol{B}\boldsymbol{\phi}_i, \qquad \forall i \in \{1, \ldots, d\}, \tag{66}$$

  and in matrix form, it is:

$$\boldsymbol{A}\boldsymbol{\Phi} = \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{\Lambda}, \tag{67}$$

  where the columns of $\mathbb{R}^{d \times d} \ni \boldsymbol{\Phi} := [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d]$ are the eigenvectors and diagonal elements of $\mathbb{R}^{d \times d} \ni \boldsymbol{\Lambda} := \mathbf{diag}([\lambda_1, \ldots, \lambda_d]^\top)$ are the eigenvalues. Note that $\boldsymbol{\phi}_i \in \mathbb{R}^d$ and $\lambda_i \in \mathbb{R}$.
- The generalized eigenvalue problem of Eq. (66) or (67) is denoted by $(\boldsymbol{A}, \boldsymbol{B})$.
- The $(\boldsymbol{A}, \boldsymbol{B})$ is called a "pair" or "pencil" [5], and the order in the pair matters, according to Eq. (67).
- The $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$ are called the generalized eigenvectors and eigenvalues of $(\boldsymbol{A}, \boldsymbol{B})$.
- The $(\boldsymbol{\Phi}, \boldsymbol{\Lambda})$ or $(\boldsymbol{\phi}_i, \lambda_i)$ is called the "eigenpair" of the pair $(\boldsymbol{A}, \boldsymbol{B})$ in the literature [5].
- Comparing Eqs. (63) and (66) or Eqs. (64) and (67) demonstrates that the eigenvalue problem is a special case of the generalized eigenvalue problem where $\boldsymbol{B} = \boldsymbol{I}$.

Optimization Forms of (Generalized) Eigenvalue Problem

- **Optimization Form 1:**

$$\begin{aligned} \underset{\boldsymbol{\phi}}{\text{maximize}} \quad & \boldsymbol{\phi}^\top \boldsymbol{A}\, \boldsymbol{\phi}, \\ \text{subject to} \quad & \boldsymbol{\phi}^\top \boldsymbol{B}\, \boldsymbol{\phi} = 1, \end{aligned} \tag{68}$$

where $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{B} \in \mathbb{R}^{d \times d}$. The Lagrangian for Eq. (68) is:

$$\mathcal{L} = \boldsymbol{\phi}^\top \boldsymbol{A}\, \boldsymbol{\phi} - \lambda\, (\boldsymbol{\phi}^\top \boldsymbol{B}\, \boldsymbol{\phi} - 1),$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier. Equating the derivative of Lagrangian to zero gives us:

$$\mathbb{R}^d \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{\phi}} = 2\boldsymbol{A}\boldsymbol{\phi} - 2\lambda \boldsymbol{B}\boldsymbol{\phi} \overset{\text{set}}{=} 0 \implies \boldsymbol{A}\boldsymbol{\phi} = \lambda \boldsymbol{B}\boldsymbol{\phi},$$

which is a generalized eigenvalue problem $(\boldsymbol{A}, \boldsymbol{B})$ according to Eq. (66), where $\boldsymbol{\phi}$ is the eigenvector and λ is the eigenvalue.

- As Eq. (68) is a *maximization* problem, the eigenvector is the one having the largest eigenvalue. If Eq. (68) is a *minimization* problem, the eigenvector is the one having the smallest eigenvalue.

Optimization Forms of (Generalized) Eigenvalue Problem

- **Optimization Form 2:**

$$\begin{aligned} \underset{\boldsymbol{\Phi}}{\text{maximize}} \quad & \mathbf{tr}(\boldsymbol{\Phi}^\top \boldsymbol{A}\, \boldsymbol{\Phi}), \\ \text{subject to} \quad & \boldsymbol{\Phi}^\top \boldsymbol{B}\, \boldsymbol{\Phi} = \boldsymbol{I}, \end{aligned} \tag{69}$$

where $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{B} \in \mathbb{R}^{d \times d}$.
The Lagrangian for Eq. (69) is:

$$\mathcal{L} = \mathbf{tr}(\boldsymbol{\Phi}^\top \boldsymbol{A}\, \boldsymbol{\Phi}) - \mathbf{tr}\big(\boldsymbol{\Lambda}^\top (\boldsymbol{\Phi}^\top \boldsymbol{B}\, \boldsymbol{\Phi} - \boldsymbol{I})\big),$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose entries are the Lagrange multipliers.
Equating the derivative of $\mathcal{L}$ to zero gives us:

$$\mathbb{R}^{d \times d} \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Phi}} = 2\, \boldsymbol{A}\boldsymbol{\Phi} - 2\, \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{\Lambda} \overset{\text{set}}{=} \mathbf{0} \implies \boldsymbol{A}\boldsymbol{\Phi} = \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{\Lambda},$$

which is an eigenvalue problem $(\boldsymbol{A}, \boldsymbol{B})$ according to Eq. (67).

Optimization Forms of (Generalized) Eigenvalue Problem

- **Optimization Form 3:** Consider the following optimization problem with the variable $\boldsymbol{\phi} \in \mathbb{R}^d$:

$$\begin{aligned} \underset{\boldsymbol{\phi}}{\text{minimize}} \quad & ||\boldsymbol{X} - \boldsymbol{\phi}\,\boldsymbol{\phi}^\top \boldsymbol{X}||_F^2, \\ \text{subject to} \quad & \boldsymbol{\phi}^\top \boldsymbol{\phi} = 1, \end{aligned} \tag{70}$$

where $\boldsymbol{X} \in \mathbb{R}^{d \times n}$. The objective function in Eq. (70) is simplified as:

$$\begin{aligned} &||\boldsymbol{X} - \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X}||_F^2 = \mathbf{tr}\big((\boldsymbol{X} - \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X})^\top (\boldsymbol{X} - \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X})\big) \\ &= \mathbf{tr}\big((\boldsymbol{X}^\top - \boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top)(\boldsymbol{X} - \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X})\big) = \mathbf{tr}\big(\boldsymbol{X}^\top \boldsymbol{X} - 2\boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X} + \boldsymbol{X}^\top \boldsymbol{\phi} \underbrace{\boldsymbol{\phi}^\top \boldsymbol{\phi}}_{1} \boldsymbol{\phi}^\top \boldsymbol{X}\big) \\ &= \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X}) = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{X}) \\ &= \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - \mathbf{tr}(\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top) = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top). \end{aligned}$$

The Lagrangian is:

$$\mathcal{L} = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - \mathbf{tr}(\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top) - \lambda(\boldsymbol{\phi}^\top \boldsymbol{\phi} - 1),$$

where λ is the Lagrange multiplier. Equating the derivative of $\mathcal{L}$ to zero gives:

$$\mathbb{R}^d \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{\phi}} = 2\,\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\phi} - 2\,\lambda\,\boldsymbol{\phi} \overset{\text{set}}{=} \mathbf{0} \implies \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\phi} = \lambda\,\boldsymbol{\phi} \implies \boldsymbol{A}\,\boldsymbol{\phi} = \lambda\,\boldsymbol{\phi},$$

which is an eigenvalue problem for $\boldsymbol{A}$ according to Eq. (66), where $\boldsymbol{\phi}$ is the eigenvector and λ is the eigenvalue.

Optimization Forms of (Generalized) Eigenvalue Problem

- **Optimization Form 4:** Consider the following optimization problem with the variable $\boldsymbol{\Phi} \in \mathbb{R}^{d \times d}$:

$$\begin{aligned} \underset{\boldsymbol{\Phi}}{\text{minimize}} \quad & ||\boldsymbol{X} - \boldsymbol{\Phi}\,\boldsymbol{\Phi}^\top \boldsymbol{X}||_F^2, \\ \text{subject to} \quad & \boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{I}, \end{aligned} \tag{71}$$

where $\boldsymbol{X} \in \mathbb{R}^{d \times n}$. Similar to Eq. (70), the objective function in Eq. (71) is simplified as:

$$||\boldsymbol{X} - \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \boldsymbol{X}||_F^2 = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top)$$

The Lagrangian is:

$$\mathcal{L} = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - \mathbf{tr}(\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top) - \mathbf{tr}\big(\boldsymbol{\Lambda}^\top (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} - \boldsymbol{I})\big),$$

$$\mathbb{R}^{d \times d} \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Phi}} = 2\,\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Phi} - 2\,\boldsymbol{\Phi}\boldsymbol{\Lambda} \overset{\text{set}}{=} \mathbf{0} \implies \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda} \implies \boldsymbol{A}\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda},$$

which is an eigenvalue problem for $\boldsymbol{A}$ according to Eq. (67). The columns of $\boldsymbol{\Phi}$ are the eigenvectors of $\boldsymbol{A}$ and the diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues.

Optimization Forms of (Generalized) Eigenvalue Problem

- **Optimization Form 5:** Consider the following optimization problem [5] with the variable $\boldsymbol{\phi} \in \mathbb{R}^d$:

$$\underset{\boldsymbol{\phi}}{\text{maximize}} \quad \frac{\boldsymbol{\phi}^\top \boldsymbol{A}\, \boldsymbol{\phi}}{\boldsymbol{\phi}^\top \boldsymbol{B}\, \boldsymbol{\phi}}. \tag{72}$$

According to the generalized Rayleigh-Ritz quotient method [6], this optimization problem can be restated as:

$$\begin{aligned} \underset{\boldsymbol{\phi}}{\text{maximize}} \quad & \boldsymbol{\phi}^\top \boldsymbol{A}\, \boldsymbol{\phi}, \\ \text{subject to} \quad & \boldsymbol{\phi}^\top \boldsymbol{B}\, \boldsymbol{\phi} = 1, \end{aligned} \tag{73}$$

The Lagrangian is:

$$\mathcal{L} = \boldsymbol{\phi}^\top \boldsymbol{A}\, \boldsymbol{\phi} - \lambda(\boldsymbol{\phi}^\top \boldsymbol{B}\, \boldsymbol{\phi} - 1),$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 2\, \boldsymbol{A}\, \boldsymbol{\phi} - 2\, \lambda\, \boldsymbol{B}\, \boldsymbol{\phi} \overset{\text{set}}{=} \boldsymbol{0} \implies 2\, \boldsymbol{A}\, \boldsymbol{\phi} = 2\, \lambda\, \boldsymbol{B}\, \boldsymbol{\phi} \implies \boldsymbol{A}\, \boldsymbol{\phi} = \lambda\, \boldsymbol{B}\, \boldsymbol{\phi},$$

which is a generalized eigenvalue problem $(\boldsymbol{A}, \boldsymbol{B})$ according to Eq. (66).

Singular Value Decomposition

- **Singular Value Decomposition (SVD)** [10] is one of the most well-known and effective matrix decomposition methods. There are different methods for obtaining this decomposition, one of which is Jordan's algorithm [10].
- SVD has two different forms, i.e., complete and incomplete.
- Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{\alpha \times \beta}$. The **complete SVD** decomposes the matrix as:

$$\begin{aligned} &\mathbb{R}^{\alpha \times \beta} \ni \boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \\ &\boldsymbol{U} \in \mathbb{R}^{\alpha \times \alpha}, \quad \boldsymbol{V} \in \mathbb{R}^{\beta \times \beta}, \quad \boldsymbol{\Sigma} \in \mathbb{R}^{\alpha \times \beta}, \end{aligned} \tag{74}$$

  where the columns of $\boldsymbol{U}$ and the columns of $\boldsymbol{V}$ are called *left singular vectors* and *right singular vectors*, respectively.
- In complete SVD, $\boldsymbol{\Sigma}$ is a *rectangular* diagonal matrix whose main diagonal includes the *singular values*. In the cases with $\alpha > \beta$ and $\alpha < \beta$, this matrix is in the following forms:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_\beta \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 & \cdots & 0 \\ 0 & 0 & \sigma_\alpha & 0 & \cdots & 0 \end{bmatrix},$$

  respectively. In other words, the number of singular values is $\min(\alpha, \beta)$.

Singular Value Decomposition

- The **incomplete SVD** decomposes the matrix as:

$$\mathbb{R}^{\alpha\times\beta} \ni \boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \tag{75}$$
$$\boldsymbol{U} \in \mathbb{R}^{\alpha\times k}, \quad \boldsymbol{V} \in \mathbb{R}^{\beta\times k}, \quad \boldsymbol{\Sigma} \in \mathbb{R}^{k\times k},$$

  where [11]:

$$k := \min(\alpha, \beta), \tag{76}$$

  and the columns of $\boldsymbol{U}$ and the columns of $\boldsymbol{V}$ are called *left singular vectors* and *right singular vectors*, respectively.

- In incomplete SVD, $\boldsymbol{\Sigma}$ is a *square* diagonal matrix whose main diagonal includes the *singular values*. The matrix $\boldsymbol{\Sigma}$ is in the form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_k \end{bmatrix}.$$

Singular Value Decomposition

- Note that in both complete and incomplete SVD, the left singular vectors are orthonormal and the right singular vectors are also orthonormal; therefore, $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrices so:

$$\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}, \tag{77}$$

$$\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}. \tag{78}$$

  If these orthogonal matrices are not truncated and thus are square matrices, e.g. for complete SVD, there are also:

$$\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}, \tag{79}$$

$$\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}. \tag{80}$$

Relation of SVD and EVD

- In both complete and incomplete SVD of matrix $\boldsymbol{A}$, the left and right singular vectors are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$, respectively, and the singular values are the square root of eigenvalues of either $\boldsymbol{A}\boldsymbol{A}^\top$ or $\boldsymbol{A}^\top\boldsymbol{A}$.
- Proof: There is:

$$\boldsymbol{A}\boldsymbol{A}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\underbrace{\boldsymbol{V}^\top\boldsymbol{V}}_{\boldsymbol{I}}\boldsymbol{\Sigma}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top,$$

  which is the eigen-decomposition [12] of $\boldsymbol{A}\boldsymbol{A}^\top$ where the columns of $\boldsymbol{U}$ are the eigenvectors and the diagonal of $\boldsymbol{\Sigma}^2$ are the eigenvalues so the diagonal of $\boldsymbol{\Sigma}$ are the square root of eigenvalues.
- Also:

$$\boldsymbol{A}^\top\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}\underbrace{\boldsymbol{U}^\top\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top,$$

  which is the eigenvalue decomposition of $\boldsymbol{A}^\top\boldsymbol{A}$ where the columns of $\boldsymbol{V}$ are the eigenvectors and the diagonal of $\boldsymbol{\Sigma}^2$ are the eigenvalues, so the diagonal of $\boldsymbol{\Sigma}$ are the square root of eigenvalues.

Acknowledgement

- Some slides are based on our textbook: "Elements of Dimensionality Reduction and Manifold Learning" [1]
- Some slides of this slide deck (expectation parts) are inspired by the lectures of Prof. Mu Zhu at University of Waterloo, Department of Statistics and Actuarial Science.
- For more information on optimization, refer to my course "Optimization Techniques" in University of Guelph. Link in my YouTube channel: `https://www.youtube.com/playlist?list=PLPrxGIUWsqP3ZBM4Zy5YqfCh1BqM5sJov`
- More information on Rayleigh-Ritz quotient, eigenvalue problem, and SVD: see our tutorial paper "Eigenvalue and generalized eigenvalue problems: Tutorial" [12]
- More information on expectation: see our two tutorial papers "The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial" [13] and "Sampling algorithms, from survey sampling to Monte Carlo methods: Tutorial and literature review" [14]

References

[1] B. Ghojogh, M. Crowley, F. Karray, and A. Ghodsi, *Elements of Dimensionality Reduction and Manifold Learning*.
Springer Nature, 2023.

[2] S. Garcia, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 3, pp. 417–435, 2012.

[3] I. Triguero, J. Derrac, S. Garcia, and F. Herrera, "A taxonomy and experimental study on prototype generation for nearest neighbor classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 1, pp. 86–100, 2011.

[4] B. Ghojogh, "Data reduction algorithms in machine learning and data science," 2021.

[5] B. N. Parlett, "The symmetric eigenvalue problem," *Classics in Applied Mathematics*, vol. 20, 1998.

[6] E. Croot, "The Rayleigh principle for finding eigenvalues," tech. rep., Georgia Institute of Technology, School of Mathematics, 2005.
Online, Accessed: March 2019.

[7] I. Jolliffe, *Principal component analysis*.
Springer, 2011.

References (cont.)

[8] J. H. Wilkinson, *The algebraic eigenvalue problem*, vol. 662. Oxford Clarendon, 1965.

[9] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3. The Johns Hopkins University Press, 2012.

[10] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM review*, vol. 35, no. 4, pp. 551–566, 1993.

[11] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[12] B. Ghojogh, F. Karray, and M. Crowley, "Eigenvalue and generalized eigenvalue problems: Tutorial," *arXiv preprint arXiv:1903.11240*, 2019.

[13] B. Ghojogh and M. Crowley, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial," *arXiv preprint arXiv:1905.12787*, 2019.

[14] B. Ghojogh, H. Nekoei, A. Ghojogh, F. Karray, and M. Crowley, "Sampling algorithms, from survey sampling to Monte Carlo methods: Tutorial and literature review," *arXiv preprint arXiv:2011.00901*, 2020.