MATH 3332: Data Analytic Tools

HU-HTAKM
Website: https://htakm.github.io/htakm_test/

Last major change: December 21, 2025
Last small update: December 21, 2025

Contents

Chapter 1

Introduction

Machine Learning is a type of artificial intelligence that focuses on the development of algorithms and models to perform tasks without being explicitly programmed to do so. Machine learning algorithms create a model based on sample data, known as training data. There are three common types of machine learning:

1. Supervised learning: Classification, Regression
   Data: N input-output pairs

   $$(\mathbf{x}_i, \mathbf{y}_i) \qquad \text{for } \mathbf{x}_i \in X, \mathbf{y}_i \in Y,\ i = 1, \cdots, N.$$

   Goal: Find a function map $f : X \to Y$ such that:

   $$f(\mathbf{x}_i) \approx \mathbf{y}_i \qquad \text{for } i = 1, \cdots, N.$$

   If a new input $\mathbf{x}$ is provided, $f(\mathbf{x})$ should accurately predict the label of $\mathbf{x}$.

2. Unsupervised learning: Clustering, Self-supervised Learning
   Data: N inputs without labels

   $$\mathbf{x}_i \qquad \text{for } \mathbf{x}_i \in X,\ i = 1, \cdots, N.$$

   Goal: Different applications have their own goals. For example, in the case of denoising, find a function map f such that:

   $$f(\mathbf{x}_i + \boldsymbol{\varepsilon}_i) = \mathbf{x}_i \qquad \text{for } i = 1, \cdots, N,$$

   where $\boldsymbol{\varepsilon}_i$ are noise vectors.

3. Reinforcement learning (Reinforcement learning algorithms are usually iterative algorithms).

In general, we want to find a good function f that maps the training data well while generalizing to other inputs.

Definition 1.1. The set of all 'good' candidate functions (models) is called the **hypothesis space**.

In our case studies, we will follow Pedro Domingos' definition of machine learning:

$$\text{Learning} = \text{Representation} + \text{Evaluation} + \text{Optimization}.$$

1. Representation: Mainly focuses on 'vector' representations.
   (a) How can we effectively represent the input data $\mathbf{x}_i$?
   (b) How can we represent the function f?
2. Evaluation: Evaluate the problem.
   (a) How do we define 'the best' function in the hypothesis space?
   We need to define a function $f' : f \to \mathbb{R}$ to compare.
   (b) How do we define 'the best' representation of the input data?
3. Optimization: Find the optimal model.
   (a) How can we obtain the optimal solution numerically using a computer?
   (b) Is convex optimization feasible? (Some problems involve non-convex optimization.)

Chapter 2

Vector spaces, metrics, limits, and convergence

2.1 Vector spaces (linear space)

Definition 2.1. A **vector space** over $\mathbb{R}$ is a set V together with two operations:

1. Addition: For all $\mathbf{u}, \mathbf{v} \in V$, $\mathbf{u} + \mathbf{v} \in V$.
2. Scalar multiplication: For all $\alpha \in \mathbb{R}$ and $\mathbf{v} \in V$, $\alpha\mathbf{v} \in V$.

These two operations satisfy the following eight properties for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $\alpha, \beta \in \mathbb{R}$:

$(+1)$	$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$,	(Addition Commutativity)
$(+2)$	$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$,	(Addition Associativity)
$(+3)$	There exists $\mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$,	(Zero Exists)
$(+4)$	For every $\mathbf{u} \in V$, there exists $\mathbf{u}' \in V$ such that $\mathbf{u} + \mathbf{u}' = \mathbf{0}$ $(\mathbf{u}' = -\mathbf{u})$,	(Additive Inverse Exists)
$(\cdot 1)$	$(\alpha\beta)\mathbf{u} = \alpha(\beta\mathbf{u})$,	(Multiplication Associativity)
$(\cdot 2)$	$1 \cdot \mathbf{u} = \mathbf{u}$,	(Unity)
$(\cdot 3)$	$\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$,	(Distributivity 1)
$(\cdot 4)$	$(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}$.	(Distributivity 2)

Remark 2.1.1. A vector space over the complex domain $\mathbb{C}$ can be defined in a similar way.

Example 2.1. The set of real numbers $\mathbb{R}$, with the standard addition and multiplication of real numbers, forms a vector space.

Example 2.2. The n-dimensional Euclidean space $\mathbb{R}^n$, with the following operations, forms a vector space over $\mathbb{R}$:

$$+ : \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \ \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix} \in \mathbb{R}^n,$$

$$\bullet : \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \text{ and } \alpha \in \mathbb{R}, \ \alpha \cdot \mathbf{x} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{pmatrix} \in \mathbb{R}^n.$$

Remark 2.1.2. Many types of input data can be modeled as vectors in $\mathbb{R}^n$, such as:

1. Digital signals of length n,
2. Stock prices over n time intervals,
3. n different features or attributes of a single object.

Example 2.3. All real $m \times n$ matrices $\mathbb{R}^{m\times n}$ with

$$+: \quad \text{For all } \mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}, \mathbf{B} = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n},$$

$$\text{we have } \mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n},$$

$$\bullet: \quad \text{For all } \mathbf{B} = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n} \text{ and } \alpha \in \mathbb{R}, \text{ we have } \alpha \cdot \mathbf{B} = \begin{pmatrix} \alpha b_{11} & \dots & \alpha b_{1n} \\ \vdots & \ddots & \vdots \\ \alpha b_{m1} & \dots & \alpha b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n}$$

is a vector space over $\mathbb{R}$.

Remark 2.1.3. This vector space is equivalent to $\mathbb{R}^{mn}$:

$$\begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} \xrightarrow{\text{vectorization}} \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \\ x_{21} \\ \vdots \\ x_{2n} \\ \vdots \\ x_{mn} \end{pmatrix} \in \mathbb{R}^{mn}.$$

Remark 2.1.4. An $m \times n$ matrix can be used to represent a black-and-white digital image.

Example 2.4. All real 3-arrays of size $m \times n \times \ell$, $\mathbb{R}^{m\times n\times \ell}$, with

$+:$ For all $X = (x_{ijk})_{i,j,k}, Y = (y_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$, we have $X + Y = (x_{ijk} + y_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$,

$\bullet:$ For all $X = (x_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$ and $\alpha \in \mathbb{R}$, we have $\alpha \cdot X = (\alpha x_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$

is a vector space over $\mathbb{R}$.

Remark 2.1.5. Similar to matrices, this vector space is equivalent to $\mathbb{R}^{mn\ell}$.

Remark 2.1.6. Many types of data can be modeled by this vector space $\mathbb{R}^{m\times n\times \ell}$, such as:

1. Color images with 3 color channels (RGB) ($\ell = 3$),
2. Black-and-white videos with ℓ frames, each of size $m \times n$.

More complex data, such as color videos, are represented as 4-arrays. These arrays are usually collectively called **tensors**.

Example 2.5. Consider the set of all strings. We can quickly see that strings do not satisfy the commutativity property:

$$\text{'Standing'} = \text{'Stand'} + \text{'ing'} \neq \text{'ing'} + \text{'Stand'} = \text{'ingStand'}.$$

Therefore, vector spaces cannot be used to model text data.

Example 2.6. The set of all continuous functions on $[a, b]$, denoted by:

$$\mathcal{C}[a, b] = \{f : f \text{ is a continuous function on } [a, b]\},$$

with, for all $t \in [a, b]$,

$$\begin{aligned} + &: \quad \text{For all } f, g \in \mathcal{C}[a, b], \text{ we have } (f + g)(t) = f(t) + g(t) \in \mathcal{C}[a, b], \\ \bullet &: \quad \text{For all } f \in \mathcal{C}[a, b] \text{ and } \alpha \in \mathbb{R}, \text{ we have } (\alpha f)(t) = \alpha f(t) \in \mathcal{C}[a, b]. \end{aligned}$$

is a vector space over $\mathbb{R}$. It is referred to as a **function space**.

Remark 2.1.7. $\mathcal{C}[a, b]$ can be considered as a hypothesis space of a learner with one input and one output. Given a dataset of $x_i \in [a, b]$ and $y_i \in \mathbb{R}$, find the function $f \in \mathcal{C}[a, b]$ such that $f(x_i) \approx y_i$ for all i.

Example 2.7. The infinite sequences:

$$\ell_\infty = \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ \vdots \end{pmatrix} : \text{There exists } c < \infty \text{ such that } |a_i| \leq c \text{ for any } i \right\},$$

with

$$\begin{aligned} + &: \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \vdots \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \end{pmatrix} \in \ell_\infty, \text{ we have } \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \\ \vdots \end{pmatrix} \in \ell_\infty, \\ \bullet &: \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \vdots \end{pmatrix} \in \ell_\infty \text{ and } \alpha \in \mathbb{R}, \text{ we have } \alpha \cdot \mathbf{x} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \\ \vdots \end{pmatrix} \in \ell_\infty. \end{aligned}$$

is a vector space over $\mathbb{R}$.

2.2 Metrics on vector space

We can convert some types of input data into vector space. However, how do we determine the distance between two vectors? Our goal is to define the distance in order to perform calculus.
Let V be a vector space and $\mathbf{u}, \mathbf{v} \in V$. We want to find a function such that:

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \text{dist}(\mathbf{0}, \mathbf{u} - \mathbf{v}) = \text{length of } \mathbf{u} - \mathbf{v}.$$

Therefore, to define the distance, we only need to define the length of vectors.

Definition 2.2. Let V be a vector space over $\mathbb{R}$. A **norm** on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y} \in V$,

1. $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$.
2. $\|\alpha\mathbf{x}\| = |\alpha| \|\mathbf{x}\|$ for $\alpha \in \mathbb{R}$.
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

The ordered pair $(V, \|\cdot\|)$ is called a **normed vector space**.

Remark 2.2.1. If it is clear from the context which norm is intended, then it is common to denote the normed vector space by V.

Remark 2.2.2. $\|\mathbf{x}\|$ represents the **length** of $\mathbf{x}$.

Remark 2.2.3. The distance between $\mathbf{x}$ and $\mathbf{y}$ can now be defined as $\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

Example 2.8. If we set $V = \mathbb{R}$, the following can be norms on $\mathbb{R}$ for all $x \in \mathbb{R}$:

1. $\|x\| = |x|$,
2. $\|x\| = \frac{1}{2}|x|$,
3. $\|x\| = c|x|$ for some $c > 0$.

Remark 2.2.4. In fact, there are infinitely many norms on the same vector space.

Example 2.9. For $\mathbf{x} \in \mathbb{R}^n$, the **Manhattan norm** (1-norm or L_1-norm) defined by:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

is a norm on $\mathbb{R}^n$. The induced distance $\|\mathbf{x} - \mathbf{y}\|_1$ for $\mathbf{x}, \mathbf{y} \in V$ is called the **Manhattan distance**.

Proof.

1. For any $\mathbf{x}, \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \geq 0.$$

Moreover,

$$\begin{aligned} \|\mathbf{x}\|_1 = 0 &\iff \sum_{i=1}^{n} |x_i| = 0 \\ &\iff |x_i| = 0 \iff x_i = 0 \qquad \text{for } i = 1, \cdots, n, \\ &\iff \mathbf{x} = \mathbf{0}. \end{aligned}$$

2. For any $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

$$\|\alpha\mathbf{x}\|_1 = \sum_{i=1}^{n} |\alpha x_i| = |\alpha| \sum_{i=1}^{n} |x_i| = |\alpha| \|\mathbf{x}\|_1.$$

3. Using the Triangle Inequality, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{x} + \mathbf{y}\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} (|x_i| + |y_i|) = \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1.$$

Therefore, by definition, the 1-norm is a norm on $\mathbb{R}^n$. □

Example 2.10. For $\mathbf{x} \in \mathbb{R}^n$, the Euclidean norm (2-norm or L_2-norm) defined by:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

is a norm on $\mathbb{R}^n$. The induced distance $\|\mathbf{x}-\mathbf{y}\|_2$ for $\mathbf{x}, \mathbf{y} \in V$ is called the **Euclidean distance**.

Proof.

1. For $\mathbf{x} \in \mathbb{R}^n$,
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \geq 0.$$
Moreover,
$$\begin{aligned} \|\mathbf{x}\|_2 = 0 &\iff \sum_{i=1}^{n} x_i^2 = 0 \\ &\iff x_i^2 = 0 \iff x_i = 0 \qquad \text{for } i = 1, \cdots, n, \\ &\iff \mathbf{x} = \mathbf{0}. \end{aligned}$$

2. For any $\alpha \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$,
$$\|\alpha\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} (\alpha x_i)^2} = \sqrt{\alpha^2}\sqrt{\sum_{i=1}^{n} x_i^2} = |\alpha| \|\mathbf{x}\|_2 .$$

3. This is a bit more complicated, but it is similar to the proof of the Cauchy-Schwarz inequality.
For $\mathbf{x} \in \mathbb{R}^n$, if $\mathbf{y} = \mathbf{0}$, then:
$$\|\mathbf{x}+\mathbf{y}\|_2 = \|\mathbf{x}\|_2 = \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$$
If $\mathbf{y} \neq 0$, then for any $t \in \mathbb{R}$,
$$\|\mathbf{x}+t\mathbf{y}\|_2^2 = \sum_{i=1}^{n} (x_i + ty_i)^2 = \left(\sum_{i=1}^{n} x_i^2\right) + t\left(2\sum_{i=1}^{n} x_i y_i\right) + t^2\left(\sum_{i=1}^{n} y_i^2\right).$$
By (1), $\|\mathbf{x}+t\mathbf{y}\|_2^2 \geq 0$. Therefore, since $\sum_{i=1}^{n} y_i^2 > 0$, $\|\mathbf{x}+t\mathbf{y}\|_2^2$ is a quadratic function with at most one real root. Using the discriminant, we have:
$$\begin{aligned} \Delta &\leq 0, \\ \left(2\sum_{i=1}^{n} x_i y_i\right)^2 - 4\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) &\leq 0, \\ \left(\sum_{i=1}^{n} x_i y_i\right)^2 &\leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right), \\ \sum_{i=1}^{n} x_i y_i &\leq \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}. \end{aligned} \qquad \text{(Cauchy-Schwarz Inequality for } \mathbb{R}^n)$$
Consequently,
$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|_2^2 &= \sum_{i=1}^{n} (x_i + y_i)^2 \\ &= \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \\ &\leq \sum_{i=1}^{n} x_i^2 + 2\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} + \sum_{i=1}^{n} y_i^2 = \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2, \\ \|\mathbf{x}+\mathbf{y}\|_2 &\leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2 . \end{aligned}$$

Therefore, by definition, the 2-norm is a norm on $\mathbb{R}^n$. □

Example 2.11. For $\mathbf{x} \in \mathbb{R}^n$, the p-norm or L_p-norm with $p \geq 1$ defined by:

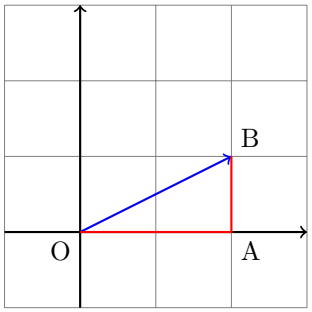
$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

is a norm on $\mathbb{R}^n$.

Example 2.12. For $\mathbf{x} \in \mathbb{R}^n$, the maximum norm ($L_\infty$-norm) defined by:

$$\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \|\mathbf{x}\|_p = \max_{1 \leq i \leq n} |x_i|$$

is a norm on $\mathbb{R}^n$.



Red: Manhattan distance from O to B

$|\text{OA}| + |\text{AB}|$

Blue: Euclidean distance from O to B

$|\text{OB}|$

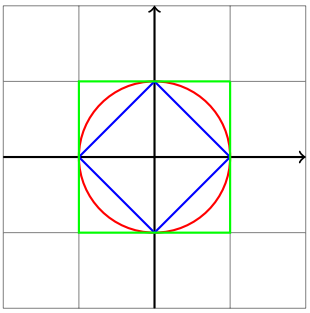
Figure 2.1: Difference between Manhattan distance and Euclidean distance

Definition 2.3. The **open unit ball** of a norm $\|\cdot\|$ is defined by:

$$B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| < 1\}.$$

Example 2.13. For p-norm, the open unit ball of $\|\cdot\|_p$ is defined by:

$$B_p = \left\{ \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} < 1 \right\}.$$



$B_2 = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 < 1\}$

$B_1 = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 < 1\}$

$B_\infty = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\infty < 1\}$

Figure 2.2: Unit balls of p-norm for different p

Theorem 2.4. If $0 < q \leq p < \infty$, then for any $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_q .$$

Remark 2.4.1. There are other norms on $\mathbb{R}^n$, not just p-norms.

For a set of matrices $\mathbb{R}^{m\times n}$, we can view them as $\mathbb{R}^{mn}$.

Example 2.14. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, the **vector p-norm** with $p\geq 1$ defined by:

$$\|\mathbf{A}\|_{p,\text{vec}}=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p\right)^{\frac{1}{p}}$$

is a norm on $\mathbb{R}^{m\times n}$.

Example 2.15. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, when $p=2$, the **Frobenius norm** (vector 2-norm) defined by:

$$\|\mathbf{A}\|_F=\|\mathbf{A}\|_{2,\text{vec}}=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2}$$

is a norm on $\mathbb{R}^{m\times n}$.

We can also view $\mathbb{R}^{m\times n}$ as a linear transformation $\mathbb{R}^n\to\mathbb{R}^m$. Given a vector $\mathbf{x}\in\mathbb{R}^n$, one can consider the function:

$$f(\mathbf{x})=\mathbf{A}\mathbf{x}\in\mathbb{R}^m$$

as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Example 2.16. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, the **matrix p-norm** with $p\geq 1$ defined by:

$$\|\mathbf{A}\|_p=\max_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}\in\mathbb{R}^n}}\frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_p=1}}\|\mathbf{A}\mathbf{x}\|_p$$

is a norm on $\mathbb{R}^{m\times n}$.

Example 2.17. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, when $p=1$, the matrix 1-norm defined by:

$$\|\mathbf{A}\|_1=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_1=1}}\|\mathbf{A}\mathbf{x}\|_1=\max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|$$

is a norm on $\mathbb{R}^{m\times n}$.

Example 2.18. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, when $p=2$, the matrix 2-norm can be described with the following properties:

$$\|\mathbf{A}\|_2^2=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_2=1}}\|\mathbf{A}\mathbf{x}\|_2^2=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_2=1}}\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}=\text{max eigenvalue of }\mathbf{A}^T\mathbf{A},$$

$$\|\mathbf{A}\|_2=\sqrt{\text{max eigenvalue of }\mathbf{A}^T\mathbf{A}}=\text{max singular value of }\mathbf{A}.$$

Therefore, the matrix 2-norm is also called the **operator norm** of $\mathbf{A}$.

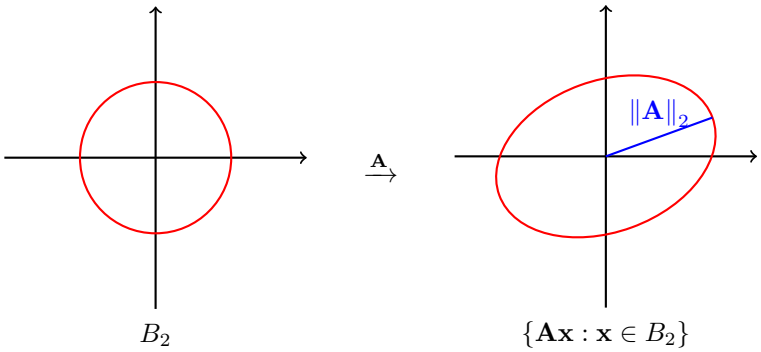


Figure 2.3: Meaning of matrix 2-norm when $n=2$ and $m=2$

In addition, the matrix p-norm can be generalized with two different p-norms.

Example 2.19. For $\mathbf{A} \in \mathbb{R}^{m\times n}$, the matrix norm induced by p-norm in $\mathbb{R}^n$ and q-norm in $\mathbb{R}^m$ defined by:

$$\|\mathbf{A}\|_{p\to q} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \|\mathbf{x}\|_p=1}} \|\mathbf{A}\mathbf{x}\|_q$$

is a norm on $\mathbb{R}^{m\times n}$. Moreover, the matrix p-norm is a special case of this norm.

Matrix norms do not need to be defined using vector norms.

Example 2.20. For $\mathbf{A} \in \mathbb{R}^{m\times n}$, the **nuclear norm** (trace norm or Ky Fan n-norm) defined by:

$$\|\mathbf{A}\|_* = \mathrm{Tr}(\sqrt{\mathbf{A}^T\mathbf{A}})$$

is a norm on $\mathbb{R}^{m\times n}$.

How about norms for continuous functions?

Example 2.21. For $f \in \mathcal{C}[a,b]$, where the set is defined as:

$$\mathcal{C}[a,b] = \{f : f \text{ is a continuous function on } [a,b]\},$$

the **maximum norm** (Chebyshev norm) defined by:

$$\|f\|_\infty = \max_{t\in[a,b]} |f(t)|$$

is a norm on $\mathcal{C}[a,b]$.

Example 2.22. For $f \in \mathcal{C}[a,b]$, the **p-norm** with $p \geq 1$ defined by:

$$\|f\|_p = \left(\int_a^b |f(t)|^p \, dt\right)^{\frac{1}{p}}$$

is a norm on $\mathcal{C}[a,b]$.

Is there a norm for the set of vectors with infinite dimensions?

Example 2.23. For $\mathbf{x} \in \ell_\infty$, where the set is defined as:

$$\ell_\infty = \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ \vdots \end{pmatrix} : \text{There exists } c < \infty \text{ such that } |a_i| \leq c \text{ for any } i \right\},$$

the **supremum norm** defined by:

$$\|\mathbf{x}\|_\infty = \sup_i |x_i|$$

is a norm on ℓ_∞.

Example 2.24. For $\mathbf{x} \in \ell_p$ with $p \geq 1$, where the set is defined as:

$$\ell_p = \{\mathbf{x} \in \ell_\infty : \|\mathbf{x}\|_p < \infty\} \subset \ell_\infty,$$

the p-norm defined by:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^\infty |x_i|^p\right)^{\frac{1}{p}}$$

is a norm on ℓ_p. However, it is important to note that it is not a norm on ℓ_∞.

Remark 2.4.2. For the same vector space, we can define infinitely many norms on it. The simplest would be by adding or subtracting different norms.

2.3 Limit and convergence on normed vector space

In this section, assume that V is a vector space over $\mathbb{R}$ with norm $\|\cdot\|$.
In data analysis, many algorithms are iterative algorithms, which generate n sequences of vectors:

$$\{\mathbf{x}_i^{(k)}\} \subset V, \qquad i = 1, \cdots, n,$$

where V is endowed with a norm $\|\cdot\|$. As the iteration continues, it is important that the output stays within the hypothesis space. If the algorithm generates an output that is outside the expected domain, then it is not considered a good solution.

Definition 2.5. Let $\mathbf{x} \in V$. We say the sequence $\{\mathbf{x}^{(k)}\} \in V$ **converges** to $\mathbf{x}$, denoted by $\mathbf{x}^{(k)} \to \mathbf{x}$, if:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\| = 0.$$

More rigorously, the sequence $\{\mathbf{x}^{(k)}\}$ **converges** to $\mathbf{x}$ if, for any $\varepsilon > 0$, there exists N such that for any $n \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}\| < \varepsilon.$$

Example 2.25. Consider $V = \mathbb{R}^n$ with $\|\cdot\|_2$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix} \frac{1}{k} \\ \vdots \\ \frac{n}{k} \end{pmatrix}, \qquad \mathbf{x} = \mathbf{0}.$$

Then we have:

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \|\mathbf{x}^{(k)}\|_2 = \sqrt{\sum_{i=1}^{n} \left(\frac{i}{k}\right)^2} = \frac{1}{k}\sqrt{\sum_{i=1}^{n} i^2},$$

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \lim_{k\to\infty} \frac{1}{k}\sqrt{\sum_{i=1}^{n} i^2} = 0.$$

Therefore, $\mathbf{x}^{(k)} \to \mathbf{x}$.

Example 2.26. Consider $V = \mathcal{C}[0,1]$ with $\|\cdot\|_\infty$. Let:

$$f^{(k)}(t) = \frac{\sin(2\pi k t)}{k^2} \in \mathcal{C}[0,1].$$

Let 0 be the zero function. We can easily find that $0 \in \mathcal{C}[0,1]$. Then we have:

$$\|f^{(k)} - 0\|_\infty = \|f^{(k)}\|_\infty = \max_{t\in[0,1]} \left|\frac{\sin(2\pi k t)}{k^2}\right| = \frac{1}{k^2},$$

$$\lim_{k\to\infty} \|f^{(k)} - 0\|_\infty = \lim_{k\to\infty} \frac{1}{k^2} = 0.$$

Therefore, $f^{(k)} \to 0$.

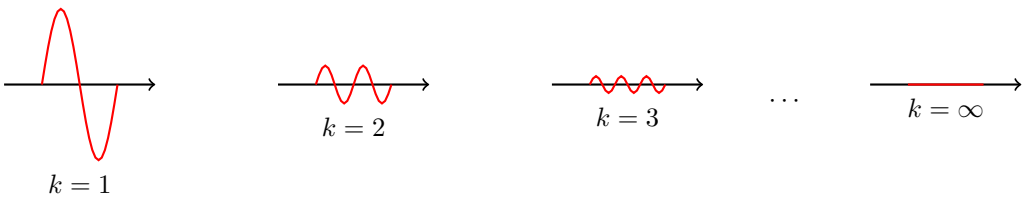


Figure 2.4: As k increases, the wave keeps shrinking in amplitude.

Remark 2.5.1. Convergence depends on the norm used.

Example 2.27. Consider $V = \ell_p$ for any p with $\|\cdot\|_p$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix} \frac{1}{k} \\ \vdots \\ \frac{1}{k} \\ 0 \\ \vdots \end{pmatrix} \begin{matrix} \\ \\ k\text{-terms} \\ \\ \\ \end{matrix} = \sum_{i=1}^{k} \frac{1}{k}\mathbf{e}_i \in \ell_p, \qquad \mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix} = \mathbf{0} \in \ell_p.$$

When $p = 2$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \|\mathbf{x}^{(k)}\|_2 = \sqrt{\sum_{i=1}^{k} \frac{1}{k^2}} = \frac{1}{\sqrt{k}},$$
$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \lim_{k\to\infty} \frac{1}{\sqrt{k}} = 0.$$

Therefore, $\mathbf{x}^{(k)} \to \mathbf{x}$ with the norm $\|\cdot\|_2$.

When $p = \infty$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty = \|\mathbf{x}^{(k)}\|_\infty = \frac{1}{k},$$
$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty = \lim_{k\to\infty} \frac{1}{k} = 0.$$

Therefore, $\mathbf{x}^{(k)} \to \mathbf{x}$ with the norm $\|\cdot\|_\infty$.

When $p = 1$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_1 = \|\mathbf{x}^{(k)}\|_1 = \sum_{i=1}^{k} \frac{1}{k} = 1,$$
$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_1 = \lim_{k\to\infty} 1 = 1 \neq 0.$$

Therefore, $\mathbf{x}^{(k)} \not\to \mathbf{x}$ with the norm $\|\cdot\|_1$.

Remark 2.5.2. The limit may not be in the same vector space. If this happens, the normed vector space is called **incomplete**.

Example 2.28. Consider $V = \ell_1$ with $\|\cdot\|_\infty$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{k} \\ 0 \\ \vdots \end{pmatrix} = \sum_{i=1}^{k} \frac{1}{i}\mathbf{e}_i \in \ell_1, \qquad \mathbf{x} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{k} \\ \frac{1}{k+1} \\ \vdots \end{pmatrix} = \sum_{i=1}^{\infty} \frac{1}{i}\mathbf{e}_i.$$

Then we have:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty = \lim_{k\to\infty} \left\| \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{k+1} \\ \vdots \end{pmatrix} \right\|_\infty = \lim_{k\to\infty} \frac{1}{k+1} = 0.$$

However, $\sum_{i=1}^{\infty} \frac{1}{i} = \infty$, and thus $\mathbf{x} \notin \ell_1$. Therefore, we cannot say that $\mathbf{x}^{(k)} \to \mathbf{x}$ with the norm $\|\cdot\|_\infty$.

With the definition of convergence, we can define the completeness of a vector space.

Definition 2.6. The sequence $\{\mathbf{x}^{(k)}\} \subset V$ is a Cauchy sequence if, for any $\varepsilon > 0$, there exists N such that for any $n, m \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\| < \varepsilon.$$

Theorem 2.7. If $\mathbf{x}^{(k)} \to \mathbf{x}$ in $(V, \|\cdot\|)$, then $\{\mathbf{x}^{(k)}\}$ is a Cauchy sequence.

Proof.
If $\mathbf{x}^{(k)} \to \mathbf{x}$, then for all $\varepsilon > 0$, there exists N such that for all $n \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}\| < \frac{\varepsilon}{2}.$$

Therefore, for all $n, m \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\| \leq \|\mathbf{x}^{(n)} - \mathbf{x}\| + \|\mathbf{x}^{(m)} - \mathbf{x}\| < \varepsilon.$$

□

Remark 2.7.1. The converse is not necessarily true.

Definition 2.8. A normed vector space $(V, \|\cdot\|)$ is **complete** if the limit of all Cauchy sequences in V is in V.

Definition 2.9. A complete normed vector space is called a **Banach space**.

Example 2.29. $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$, or $\mathbb{R}^{m\times n\times \ell}$ with any norm is a Banach space.

Example 2.30. $\mathcal{C}[a, b]$ with $\|\cdot\|_\infty$ is a Banach space.

Example 2.31. For $p \geq 1$, including $p = \infty$, $(\ell_p, \|\cdot\|_p)$ is a Banach space.

Remark 2.9.1. We can always include all the limits of the Cauchy sequences to convert an incomplete normed vector space into a complete one.

Example 2.32. $(\ell_1, \|\cdot\|_\infty)$ is an incomplete normed vector space. Its completion is ℓ_∞.

Example 2.33. For $p \geq 1$, $(\mathcal{C}[a, b], \|\cdot\|_p)$ is an incomplete normed vector space. Its completion is $L^p[a, b]$.

In practical cases, how do we check the convergence of a given sequence?

Example 2.34. From an iterative algorithm, we generate a sequence of vectors $\{\mathbf{x}^{(k)}\}$. Our goal is to check if this sequence converges. Pick a threshold $\varepsilon > 0$. We can check computationally that for large n, m,

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\| < \varepsilon.$$

Practically, to reduce computational cost, we usually only check for large n,

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| < \varepsilon.$$

2.4 Finite Dimensional Vector Spaces

In most cases, we deal with finite-dimensional vector spaces.

Remark 2.9.2. Every finite-dimensional vector space with any norm is complete. That is, any finite-dimensional vector space is Banach.

Remark 2.9.3. For a finite-dimensional vector space V, all norms are equivalent. For any two norms $\|\cdot\|_A$ and $\|\cdot\|_B$, there exist $c_1, c_2 > 0$ such that:

$$c_1 \|\mathbf{x}\|_A \leq \|\mathbf{x}\|_B \leq c_2 \|\mathbf{x}\|_A , \qquad \text{for all } \mathbf{x} \in V.$$

Theorem 2.10. The limit of the same finite-dimensional sequence under any norm is the same. That means, given two finite-dimensional normed vector spaces $(V, \|\cdot\|_A)$ and $(V, \|\cdot\|_B)$, for any sequence $\{\mathbf{x}^{(k)}\}$ and $\mathbf{x} \in V$,

$$\mathbf{x}^{(k)} \to \mathbf{x} \text{ in } \|\cdot\|_A \iff \mathbf{x}^{(k)} \to \mathbf{x} \text{ in } \|\cdot\|_B .$$

Proof.
Since $\mathbf{x}^{(k)} \to \mathbf{x}$ in $\|\cdot\|_A$,

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_A = 0.$$

Therefore, there exist $c_1, c_2 > 0$ such that:

$$c_1 \|\mathbf{x}^{(k)} - \mathbf{x}\|_A \leq \|\mathbf{x}^{(k)} - \mathbf{x}\|_B \leq c_2 \|\mathbf{x}^{(k)} - \mathbf{x}\|_A.$$

Taking $k \to \infty$, we have:

$$0 \leq c_1 \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_A \leq \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_B \leq c_2 \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_A = 0.$$

By the Squeeze Theorem, we find that:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_B = 0.$$

To conclude, $\mathbf{x}^{(k)} \to \mathbf{x}$ in $\|\cdot\|_B$. The proof for the converse is similar. □

Example 2.35. Consider $V = \mathbb{R}^n$ with $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$.

1. $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent because for $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\, \|\mathbf{x}\|_2 .$$

2. $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equivalent because for $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\, \|\mathbf{x}\|_\infty .$$

3. $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are equivalent because for $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\, \|\mathbf{x}\|_\infty .$$

Remark 2.10.1. Based on the theorem, the convergence speed depends on the norms used.

Example 2.36. Consider $V = \mathbb{R}^2$. Let:

$$\mathbf{x}^{(k)} = \frac{1}{k}\begin{pmatrix}\cos\left(\frac{(2k-1)\pi}{4}\right)\\ \sin\left(\frac{(2k-1)\pi}{4}\right)\end{pmatrix} \in \mathbb{R}^2.$$

We can easily find that $\mathbf{x}^{(k)} \to \mathbf{0}$ with the norms $\|\cdot\|_1$ and $\|\cdot\|_2$. However,

$$\|\mathbf{x}^{(k)} - \mathbf{0}\|_1 = \frac{\sqrt{2}}{k}, \qquad \|\mathbf{x}^{(k)} - \mathbf{0}\|_2 = \frac{1}{k}.$$

To achieve ε-precision:

$$\|\mathbf{x}^{(k)} - \mathbf{0}\|_1 < \varepsilon \iff \frac{\sqrt{2}}{k_1} < \varepsilon \implies k_1 > \frac{\sqrt{2}}{\varepsilon},$$
$$\|\mathbf{x}^{(k)} - \mathbf{0}\|_2 < \varepsilon \iff \frac{1}{k_2} < \varepsilon \implies k_2 > \frac{1}{\varepsilon}.$$

The norm $\|\cdot\|_1$ takes about $\sqrt{2}$ times as many iterations as the norm $\|\cdot\|_2$ to check convergence using ε as the threshold.

Remark 2.10.2. In the case of infinite-dimensional vector spaces, not all norms are equivalent.

Example 2.37. Consider $V = \ell_1$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\end{pmatrix} \begin{matrix}\\ \\ k\text{-terms}\\ \\ \\ \end{matrix} = \sum_{i=1}^{k} \mathbf{e}_i \in \ell_1.$$

Consider the norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$. We find that:

$$\|\mathbf{x}^{(k)}\|_\infty = 1, \qquad \|\mathbf{x}^{(k)}\|_1 = k, \qquad \lim_{k\to\infty} \frac{\|\mathbf{x}^{(k)}\|_1}{\|\mathbf{x}^{(k)}\|_\infty} = \lim_{k\to\infty} k = \infty.$$

Therefore, the two norms are not equivalent.

Read Appendix A (Clustering, K-means, K-medians) to see the case study for Chapter 2.

Chapter 3

Inner products, Hilbert Spaces

3.1 Inner product

How do we describe whether two vectors are correlated? We cannot use norms to describe this because they are scaling-sensitive. As such, we define the inner product to quantify the relationship between two vectors.

Definition 3.1. Let V be a vector space over $\mathbb{R}$. An **inner product** over $\mathbb{R}$ is a binary operator $\langle \cdot, \cdot \rangle : (V, V) \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}$.
2. $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle$ for all $\alpha, \beta \in \mathbb{R}$.
3. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.

Remark 3.1.1. Property (2) and (3) are equivalent to the following: for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle \mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle + \beta \langle \mathbf{x}, \mathbf{z} \rangle .$$

Remark 3.1.2. Inner products can also be defined over $\mathbb{C}$, but property (3) would change to:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}.$$

Example 3.1. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the **dot product** for $\mathbb{R}^n$ defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T \mathbf{y}$$

is the standard inner product in $\mathbb{R}^n$.

Example 3.2. For a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the "weighted" inner product defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}$$

is an inner product on $\mathbb{R}^n$. The standard inner product is a special case of this inner product where $\mathbf{A} = \mathbf{I}$. However, if $\mathbf{A}$ is a symmetric positive semi-definite matrix, then the weighted inner product is not a true inner product. Instead, it is a **semi-inner-product**.

Proof.

1. For $\mathbf{x} \in \mathbb{R}^n$:
$$\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0.$$
Moreover:
$$\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = 0 \iff \mathbf{x}^T \mathbf{A} \mathbf{x} = 0 \iff \mathbf{x} = \mathbf{0}.$$
2. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$:
$$\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle_{\mathbf{A}} = (\alpha\mathbf{x} + \beta\mathbf{y})^T \mathbf{A} \mathbf{z} = \alpha \mathbf{x}^T \mathbf{A} \mathbf{z} + \beta \mathbf{y}^T \mathbf{A} \mathbf{z} = \alpha \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{A}} + \beta \langle \mathbf{y}, \mathbf{z} \rangle_{\mathbf{A}} .$$
3. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:
$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y} = (\mathbf{x}^T \mathbf{A} \mathbf{y})^T = \mathbf{y}^T \mathbf{A}^T \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbf{A}} .$$

□

Example 3.3. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, the inner product defined by:

$$\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij} = \mathrm{Tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{Tr}(\mathbf{B}^T\mathbf{A}) = \mathrm{Tr}(\mathbf{A}\mathbf{B}^T) = \mathrm{Tr}(\mathbf{B}\mathbf{A}^T)$$

is the standard inner product in $\mathbb{R}^{m\times n}$.

Example 3.4. For $\mathbf{x}, \mathbf{y} \in \ell_2$, the inner product defined by:

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{\infty} x_i y_i$$

is the standard inner product in ℓ_2.

Example 3.5. For $f, g \in \mathcal{C}[a, b]$, the inner product defined by:

$$\langle f, g\rangle = \int_a^b f(t)g(t)\,dt$$

is the standard inner product in $\mathcal{C}[a, b]$.

3.2 Properties of inner products

Let V be a vector space over $\mathbb{R}$ with an inner product $\langle\cdot,\cdot\rangle$.
Using the definition of inner products, we can discuss some properties of inner products.

Theorem 3.2. For any $\mathbf{x} \in V$, we have:

$$\langle \mathbf{x}, \mathbf{0}\rangle = \langle \mathbf{0}, \mathbf{x}\rangle = 0.$$

Proof.

$$\langle \mathbf{x}, \mathbf{0}\rangle = \langle \mathbf{x}, \mathbf{x} - \mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{x}, \mathbf{x}\rangle = 0.$$

□

In the previous chapter, we proved the Cauchy-Schwarz Inequality in $\mathbb{R}^n$ when verifying whether the 2-norm in $\mathbb{R}^n$ is a norm. We can generalize this inequality to all inner products.

Theorem 3.3. (**Cauchy-Schwarz Inequality**) For all $\mathbf{x}, \mathbf{y} \in V$:

$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 \le \langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle .$$

Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ for some $\alpha \in \mathbb{R}$.

Proof.
For $\mathbf{x} \in V$, if $\mathbf{y} = \mathbf{0}$, then:

$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 = |\langle \mathbf{x}, \mathbf{0}\rangle|^2 = 0 = \langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle .$$

If $\mathbf{y} \neq \mathbf{0}$, then for any $\lambda \in \mathbb{R}$:

$$0 \le \langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + \lambda \langle \mathbf{x}, \mathbf{y}\rangle + \lambda \langle \mathbf{y}, \mathbf{x}\rangle + \lambda^2 \langle \mathbf{y}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + \lambda(2\langle \mathbf{x}, \mathbf{y}\rangle) + \lambda^2 \langle \mathbf{y}, \mathbf{y}\rangle .$$

Since $\langle \mathbf{y}, \mathbf{y}\rangle > 0$, $\langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}\rangle$ is a quadratic function with at most one real root. Using the discriminant, we have:

$$\Delta = (2\langle \mathbf{x}, \mathbf{y}\rangle)^2 - 4\langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle \le 0,$$
$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 \le \langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle .$$

For the equality case, notice that $\Delta = 0$ if and only if:

$$\langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + 2\lambda\langle \mathbf{x}, \mathbf{y}\rangle + \lambda^2\langle \mathbf{y}, \mathbf{y}\rangle = \left(\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} + \lambda\sqrt{\langle \mathbf{y}, \mathbf{y}\rangle}\right)^2 - 2\lambda\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle} + 2\lambda\langle \mathbf{x}, \mathbf{y}\rangle = 0.$$

Solving this equation gives $\lambda = -\sqrt{\frac{\langle \mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{y}, \mathbf{y}\rangle}}$, and we find $\mathbf{y} = \sqrt{\frac{\langle \mathbf{y}, \mathbf{y}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}}\mathbf{x}$ by definition.
Substituting $\mathbf{y} = \alpha\mathbf{x}$ for any $\alpha \in \mathbb{R}$ shows that $\Delta = 0$ if and only if $\mathbf{y} = \alpha\mathbf{x}$. □

With the Cauchy-Schwarz Inequality, we can construct a norm using an inner product.

Theorem 3.4. The **norm induced by the inner product** is a norm defined by, for any $\mathbf{x} \in V$:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

Proof.

1. For any $\mathbf{x} \in V$:
$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \geq 0.$$
Moreover:
$$\|\mathbf{x}\| = 0 \iff \langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}.$$

2. For any $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$:
$$\|\alpha \mathbf{x}\| = \sqrt{\langle \alpha \mathbf{x}, \alpha \mathbf{x} \rangle} = \sqrt{\alpha^2} \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |\alpha| \, \|\mathbf{x}\| .$$

3. For any $\mathbf{x}, \mathbf{y} \in V$:
$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + 2 \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 |\langle \mathbf{x}, \mathbf{y} \rangle| && (c \leq |c|) \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 \|\mathbf{x}\| \, \|\mathbf{y}\| = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2, && \text{(Cauchy-Schwarz Inequality)} \\ \|\mathbf{x} + \mathbf{y}\| &\leq \|\mathbf{x}\| + \|\mathbf{y}\| . \end{aligned}$$

□

Remark 3.4.1. Inner product spaces are a subset of normed vector spaces. Not all normed vector spaces can define an inner product.

Remark 3.4.2. The Cauchy-Schwarz Inequality can be rewritten as, for $\mathbf{x}, \mathbf{y} \in V$:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\| .$$

Example 3.6. Consider $V = \mathbb{R}^n$ with the standard inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The induced norm is:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2} = \|\mathbf{x}\|_2 , \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

Remark 3.4.3. For $V = \mathbb{R}^n$, among all p-norms, only the 2-norm can be induced by an inner product.

Example 3.7. For a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, consider $V = \mathbb{R}^n$ with the weighted inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The induced norm is:

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j}.$$

If $\mathbf{A}$ is symmetric positive semi-definite, then the induced norm is not a true norm. Instead, it is a semi-norm (discussed in Appendix C).

Example 3.8. Consider $V = \mathbb{R}^{m\times n}$ with the standard inner product:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}, \qquad \text{for } \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}.$$

The induced norm is:

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \|\mathbf{A}\|_F = \|\mathbf{A}\|_{2,\text{vec}}, \qquad \text{for } \mathbf{A} \in \mathbb{R}^{m\times n}.$$

Example 3.9. Consider $V = \ell_2$ with the standard inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \ell_2.$$

The induced norm is:

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{\infty} x_i^2} = \|\mathbf{x}\|_2, \qquad \text{for } \mathbf{x} \in \ell_2.$$

Example 3.10. Consider $V = \mathcal{C}[a, b]$ with the standard inner product:

$$\langle f, g \rangle = \int_a^b f(t) g(t)\, dt, \qquad \text{for } f, g \in \mathcal{C}[a, b].$$

The induced norm is:

$$\|f\| = \sqrt{\int_a^b (f(t))^2\, dt} = \|f\|_2, \qquad \text{for } f \in \mathcal{C}[a, b].$$

What conditions are required to define an inner product based on the normed vector space?

Theorem 3.5. Let $(V, \|\cdot\|)$ be a normed vector space over $\mathbb{R}$. The norm $\|\cdot\|$ is induced by an inner product if and only if the parallelogram law holds. This means that for all $\mathbf{x}, \mathbf{y} \in V$,

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 .$$

Proof.

$\Longrightarrow$ Suppose that the norm is induced by an inner product $\langle\cdot,\cdot\rangle$. This means for any $\mathbf{x}, \mathbf{y} \in V$:

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle = \|\mathbf{x}\|^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 ,\\ \|\mathbf{x}-\mathbf{y}\|^2 &= \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle = \|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 .\end{aligned}$$

Therefore,

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 .$$

$\Longleftarrow$ Suppose that the parallelogram law holds. We may define a binary operator as:

$$\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2), \qquad \text{for } \mathbf{x}, \mathbf{y} \in V.$$

We check whether this is an inner product.

1. For any $\mathbf{x} \in V$,
$$\langle \mathbf{x}, \mathbf{x}\rangle = \frac{1}{4}(\|2\mathbf{x}\|^2 + \|\mathbf{0}\|^2) = \|\mathbf{x}\|^2 \geq 0.$$
Moreover,
$$\langle \mathbf{x}, \mathbf{x}\rangle = \|\mathbf{x}\|^2 = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}.$$

2. Since proving homogeneity extending to $\mathbb{R}$ is out of scope, we will only prove additivity here. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, by the parallelogram law,
$$\begin{aligned}\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 ,\\ &= 2\|\mathbf{y}\|^2 + 2\|\mathbf{x}+\mathbf{z}\|^2 - \|-\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 ,\\ \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}-\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 ,\\ &= 2\|\mathbf{y}\|^2 + 2\|\mathbf{x}-\mathbf{z}\|^2 - \|-\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 .\end{aligned}$$
Combining the formulas, we have:
$$\begin{aligned}\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|-\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 ,\\ \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 - \frac{1}{2}\|-\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 ,\\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|-\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 .\end{aligned}$$
Therefore,
$$\langle \mathbf{x}+\mathbf{y}, \mathbf{z}\rangle = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2) = \frac{1}{4}(\|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2) = \langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle .$$

3. For any $\mathbf{x}, \mathbf{y} \in V$,
$$\langle \mathbf{y}, \mathbf{x}\rangle = \frac{1}{4}(\|\mathbf{y}+\mathbf{x}\|^2 - \|\mathbf{y}-\mathbf{x}\|^2) = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2) = \langle \mathbf{x}, \mathbf{y}\rangle .$$

Therefore, since additionally $\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} = \|\mathbf{x}\|$, the binary operator is an inner product that induces the norm.

□

Similar to normed vector spaces, we can define the completeness of a vector space with an inner product.

Definition 3.6. A complete inner product space is called a **Hilbert space**.

Example 3.11. $\mathbb{R}^n$ with the standard inner product or the weighted inner product for any symmetric positive definite $\mathbf{A} \in \mathbb{R}^{n\times n}$:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}, \qquad \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

are Hilbert spaces.

Example 3.12. $\mathbb{R}^{m\times n}$ with the standard inner product:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{Tr}(\mathbf{A}^T \mathbf{B}), \qquad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n},$$

is a Hilbert space.

Example 3.13. ℓ_2 with the standard inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \mathbf{x}, \mathbf{y} \in \ell_2,$$

is a Hilbert space.

Example 3.14. $\mathcal{C}[a, b]$ with the standard inner product:

$$\langle f, g \rangle = \int_a^b f(t) g(t)\, dt, \qquad f, g \in \mathcal{C}[a, b],$$

is not a Hilbert space. Its completion is $L^2(a, b)$.

3.3 Orthogonality

By the Cauchy-Schwarz Inequality, for all non-zero $\mathbf{x}, \mathbf{y} \in V$:

$$-\|\mathbf{x}\|\|\mathbf{y}\| \leq \langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|\|\mathbf{y}\| \iff -1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} \leq 1.$$

Considering when the Cauchy-Schwarz Inequality achieves equality:

1. If $\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} = 1$, then $\mathbf{y} = \alpha\mathbf{x}$ with $\alpha > 0$. (If $\alpha \leq 0$, then $\langle \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{x} \rangle = \alpha \|\mathbf{x}\|^2 \leq 0$.)

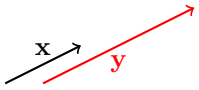


Figure 3.1: $\mathbf{y} = 2\mathbf{x}$

2. If $\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} = -1$, then $\mathbf{y} = \alpha\mathbf{x}$ with $\alpha < 0$. (If $\alpha \geq 0$, then $\langle \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{x} \rangle = \alpha \|\mathbf{x}\|^2 \geq 0$.)

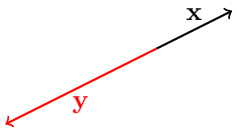


Figure 3.2: $\mathbf{y} = -2\mathbf{x}$

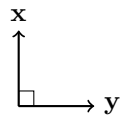
Definition 3.7. The **angle** between nonzero $\mathbf{x}, \mathbf{y} \in V$ is defined by:

$$\angle(\mathbf{x}, \mathbf{y}) = \arccos\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}\right).$$

With angles defined, we can define orthogonality.

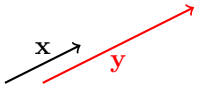
Definition 3.8. For $\mathbf{x}, \mathbf{y} \in V$:

1. If $\left|\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}\right| = 1$, then $\mathbf{x}$ and $\mathbf{y}$ are the **most correlated**.
2. If $\left|\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}\right| = 0$ ($\langle \mathbf{x}, \mathbf{y} \rangle = 0$), then $\mathbf{x}$ and $\mathbf{y}$ are the **least correlated**. We say $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal**.

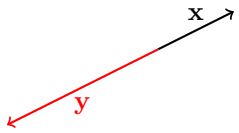


$\angle(\mathbf{x}, \mathbf{y}) = \frac{\pi}{2}$

(a) The least correlated



$\angle(\mathbf{x}, \mathbf{y}) = 0$



$\angle(\mathbf{x}, \mathbf{y}) = \pi$

(b) The most correlated

Based on orthogonality, we have the following theorem.

Theorem 3.9. (**Pythagorean Theorem**) For $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if and only if:

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

Proof.
If $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, then $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Therefore:

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

If $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, then:

$$0 = \|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{y}, \mathbf{y} \rangle = 2\langle \mathbf{x}, \mathbf{y} \rangle.$$

Therefore, $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and thus $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. □

Read Appendix B (Kernel K-means/Kernel Trick) and C (Metric Learning) to see the case studies for Chapter 3.

Chapter 4

Linear Functions and Differentiation

In Chapter 2, we observed that the norm does not preserve vector addition but does preserve scalar multiplication. In Chapter 3, by fixing one of the vectors, we demonstrated linearity. Here, we aim to investigate the behavior of linear functions in vector spaces.

4.1 Linear Functions

Definition 4.1. Let V be a vector space over $\mathbb{R}$. A function $f : V \to \mathbb{R}$ is **linear** if, for all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}).$$

Example 4.1. The mean of a vector (not to be confused with mean vectors): for all $\mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a linear function.

Proof.
For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta y_i) = \frac{\alpha}{n}\sum_{i=1}^{n} x_i + \frac{\beta}{n}\sum_{i=1}^{n} y_i = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}).$$

□

Example 4.2. The maximum entry of a vector: for all $\mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) = \max_{1 \le i \le n} x_i$$

is not a linear function.

Proof.
Assume that $\mathbf{x} = \begin{pmatrix}1\\0\end{pmatrix}$, $\mathbf{y} = \begin{pmatrix}0\\1\end{pmatrix}$, $\alpha = 1$, $\beta = 1$. We have:

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = f\left(\begin{pmatrix}1\\1\end{pmatrix}\right) = 1, \qquad \alpha f(\mathbf{x}) + \beta f(\mathbf{y}) = f\left(\begin{pmatrix}1\\0\end{pmatrix}\right) + f\left(\begin{pmatrix}0\\1\end{pmatrix}\right) = 2 \neq f(\alpha\mathbf{x} + \beta\mathbf{y}),$$

which violates the definition of a linear function. □

Example 4.3. Let V be a normed vector space with an inner product $\langle \cdot, \cdot \rangle$ and let $\mathbf{a} \in V$. The function $f : V \to \mathbb{R}$ defined by:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle \qquad \text{for } \mathbf{x} \in V$$

is a linear function.

Proof.
For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \langle \mathbf{a}, \alpha\mathbf{x} + \beta\mathbf{y} \rangle = \alpha \langle \mathbf{a}, \mathbf{x} \rangle + \beta \langle \mathbf{a}, \mathbf{y} \rangle = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}).$$

□

Example 4.4. A functional $F : \mathcal{C}[-1,1] \to \mathbb{R}$ defined by:

$$F(f) = f(0) \qquad \text{for } f \in \mathcal{C}[-1,1]$$

is a linear function.

Proof.
For all $f, g \in \mathcal{C}[-1,1]$ and $\alpha, \beta \in \mathbb{R}$,

$$F(\alpha f + \beta g) = (\alpha f + \beta g)(0) = \alpha f(0) + \beta g(0) = \alpha F(f) + \beta F(g).$$

□

Example 4.5. A functional $F : \mathcal{C}[a,b] \to \mathbb{R}$ defined by:

$$F(f) = \int_a^b f(t)\,dt \qquad \text{for } f \in \mathcal{C}[a,b]$$

is a linear function.

Proof.
For all $f, g \in \mathcal{C}[a,b]$ and $\alpha, \beta \in \mathbb{R}$,

$$F(\alpha f + \beta g) = \int_a^b (\alpha f + \beta g)(t)\,dt = \int_a^b (\alpha f(t) + \beta g(t))\,dt = \alpha \int_a^b f(t)\,dt + \beta \int_a^b g(t)\,dt = \alpha F(f) + \beta F(g).$$

□

Theorem 4.2. For any vector space V, any norm function $\|\cdot\|$ on V is not a linear function.

Proof.
By the absolute homogeneity property of the norm, for all $\mathbf{x} \in V$,

$$\|-\mathbf{x}\| = \|\mathbf{x}\|.$$

Assume that the norm function is linear. Then:

$$\|-\mathbf{x}\| = \|-\mathbf{x} + 0\mathbf{x}\| = -\|\mathbf{x}\| + 0\|\mathbf{x}\| = -\|\mathbf{x}\|,$$

which results in a contradiction. Therefore, any norm function is not a linear function. □

The following properties can be easily derived from the definition.

Theorem 4.3. A linear function f has the following properties:

1. Homogeneity: For all $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$,
$$f(\alpha\mathbf{x}) = \alpha f(\mathbf{x}).$$
2. Additivity: For any $\mathbf{x}, \mathbf{y} \in V$,
$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}).$$

Proof.

1. Setting $\beta = 0$, for any $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$,
$$f(\alpha\mathbf{x}) = f(\alpha\mathbf{x} + 0\mathbf{y}) = \alpha f(\mathbf{x}) + 0 f(\mathbf{y}) = \alpha f(\mathbf{x}), \qquad \text{for } \mathbf{y} \in V.$$
2. Setting $\alpha = \beta = 1$, for any $\mathbf{x}, \mathbf{y} \in V$,
$$f(\mathbf{x} + \mathbf{y}) = f(1\mathbf{x} + 1\mathbf{y}) = 1f(\mathbf{x}) + 1f(\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}).$$

□

Remark 4.3.1. By induction, for $\mathbf{x}_1, \cdots, \mathbf{x}_k \in V$ and $\alpha_1, \cdots, \alpha_k \in \mathbb{R}$,

$$f(\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k) = \alpha_1 f(\mathbf{x}_1) + \cdots + \alpha_k f(\mathbf{x}_k).$$

Let H be a Hilbert space. From Example 4.3, we have shown that for all $\mathbf{a} \in H$, the function:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle \qquad \text{for } \mathbf{x} \in H$$

is a linear function. Is it true that for all linear functions $f : H \to \mathbb{R}$, there exists a fixed vector $\mathbf{a} \in H$ such that the function can be written in inner product form? The answer is yes!

Theorem 4.4. (**Riesz Representation Theorem**) Let H be a Hilbert space with an inner product $\langle \cdot, \cdot \rangle$. The function $f : H \to \mathbb{R}$ is linear and bounded if and only if:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle, \qquad \text{for } \mathbf{x} \in H$$

for some unique $\mathbf{a} \in H$, called the **Riesz representation** of f.

Proof (When $H = \mathbb{R}^n$).
For all $\mathbf{x} \in \mathbb{R}^n$, we have:

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n.$$

Therefore, we have:

$$f(\mathbf{x}) = f(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) = x_1 f(\mathbf{e}_1) + \cdots + x_n f(\mathbf{e}_n) = \left\langle \begin{pmatrix} f(\mathbf{e}_1) \\ \vdots \\ f(\mathbf{e}_n) \end{pmatrix}, \mathbf{x} \right\rangle \Longrightarrow \mathbf{a} = \begin{pmatrix} f(\mathbf{e}_1) \\ \vdots \\ f(\mathbf{e}_n) \end{pmatrix}.$$

Suppose that $\mathbf{a}$ is not unique. This means there exist $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ such that:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle = \langle \mathbf{b}, \mathbf{x} \rangle \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

If we choose $\mathbf{x} = \mathbf{e}_i$ for $i = 1, \cdots, n$,

$$\begin{aligned} f(\mathbf{e}_i) = \langle \mathbf{a}, \mathbf{e}_i \rangle = \langle \mathbf{b}, \mathbf{e}_i \rangle \Longrightarrow a_i = b_i, \qquad & \text{for } i = 1, \cdots, n, \\ \Longrightarrow \mathbf{a} = \mathbf{b}. & \end{aligned}$$

This results in a contradiction. Therefore, $\mathbf{a}$ is unique. □

Remark 4.4.1. The "boundedness" implied in this theorem does not mean the codomain of the function is a bounded set. Given a function $f : X \to Y$ that maps between two normed vector spaces, if X has a norm $\|\cdot\|_X$ and Y has a norm $\|\cdot\|_Y$, then there exists some $M > 0$ such that:

$$\|f(\mathbf{x})\|_Y \leq M \|\mathbf{x}\|_X, \qquad \text{for } \mathbf{x} \in X.$$

In this theorem, if H induces a norm $\|\cdot\|_H$, then there exists some $M > 0$ such that:

$$|f(\mathbf{x})| \leq M \|\mathbf{x}\|_H, \qquad \text{for } \mathbf{x} \in H.$$

Remark 4.4.2. The smallest value M is often called the **operator norm**. Recall Example 2.18. It can be generalized to:

$$M = \|f\|_{op} = \sup\{\|f(\mathbf{x})\|_Y : \mathbf{x} \in X \text{ with } \|\mathbf{x}\|_X \leq 1\},$$

which depends on the choice of norms.

Remark 4.4.3. Any linear function defined on a finite-dimensional normed vector space is always bounded.

Example 4.6. The mean of a vector $\mathbf{x} \in \mathbb{R}^n$ can be represented by:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \left\langle \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \mathbf{x} \right\rangle.$$

Example 4.7. The trace of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be represented by:

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = \langle \mathbf{I}, \mathbf{A} \rangle.$$

Remark 4.4.4. In infinite-dimensional Hilbert spaces, there exist linear but unbounded functions.

Example 4.8. We consider $L^2(-1,1)$, which is the completion of $\mathcal{C}[-1,1]$ under:

$$\langle f,g\rangle = \int_{-1}^{1} f(t)g(t)\,dt, \qquad \|f\|_2 = \sqrt{\langle f,f\rangle}.$$

Let $F(f) = f(0)$ for all $f \in L^2(-1,1)$. Consider that:

$$f(t) = \begin{cases} 1, & t \neq 0, \\ +\infty, & t = 0 \end{cases}, \qquad \text{for } t \in (-1,1).$$

Therefore, $F(f) = f(0) = +\infty$. There is, in fact, no inner product representation for $F(f) = f(0)$.

Example 4.9. We consider $G : L^2(-1,1) \to \mathbb{R}$ defined by:

$$G(f) = \int_{-1}^{1} f(t)\,dt, \qquad \text{for } f \in L^2(-1,1).$$

For any $f, g \in L^2(-1,1)$ and $\alpha, \beta \in \mathbb{R}$,

$$G(\alpha f + \beta g) = \int_{-1}^{1} (\alpha f + \beta g)(t)\,dt = \alpha \int_{-1}^{1} f(t)\,dt + \beta \int_{-1}^{1} g(t)\,dt = \alpha G(f) + \beta G(g).$$

Therefore, we can find that G is linear. We can also see that for any $f \in L^2(-1,1)$,

$$G(f) = \int_{-1}^{1} f(t)\,dt = \int_{-1}^{1} 1 \cdot f(t)\,dt = \langle 1, f\rangle \leq \|f\|_2 \sqrt{\int_{-1}^{1} 1^2\,dt} = \sqrt{2}\,\|f\|_2 .$$

Therefore, G is also bounded. In fact, we have found that:

$$G(f) = \langle 1, f\rangle .$$

Example 4.10. We consider ℓ_2, which is the completion of the space of all finite sequences under:

$$\langle \mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x},\mathbf{x}\rangle}.$$

Let $H : \ell_2 \to \mathbb{R}$ be defined by:

$$H(\mathbf{x}) = x_1, \qquad \text{for } \mathbf{x} \in \ell_2.$$

For any $\mathbf{x}, \mathbf{y} \in \ell_2$ and $\alpha, \beta \in \mathbb{R}$,

$$H(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha x_1 + \beta y_1 = \alpha H(\mathbf{x}) + \beta H(\mathbf{y}).$$

Therefore, we can find that H is linear. We can also see that for any $\mathbf{x} \in \ell_2$,

$$H(\mathbf{x}) = x_1 \leq \sqrt{\sum_{i=1}^{\infty} x_i^2} = \|\mathbf{x}\|_2 .$$

Therefore, H is also bounded. In fact, we have found that:

$$H(\mathbf{x}) = \left\langle \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \mathbf{x} \right\rangle .$$

Additionally, we have a function that is similar to a linear function but shifted in the output space.

Definition 4.5. Let V be a vector space over $\mathbb{R}$. A function $f : V \to \mathbb{R}$ is **affine** if it can be written as:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad \text{for } \mathbf{x} \in V,$$

where $g : V \to \mathbb{R}$ is linear and $b \in \mathbb{R}$.

We can also derive the following properties from the definition.

Theorem 4.6. An affine function f has the following properties:

1. For any $\alpha, \beta \in \mathbb{R}$ such that $\alpha + \beta = 1$,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}).$$

2. If H is a Hilbert space and f is bounded, then f is affine if and only if:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b,$$

for some $\mathbf{a} \in H$ and $b \in \mathbb{R}$.

Proof.

1. There exists a linear function g such that:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad \text{for } \mathbf{x} \in V,$$

where $b \in \mathbb{R}$. If $\alpha + \beta = 1$,

$$\begin{aligned} f(\alpha\mathbf{x} + \beta\mathbf{y}) &= g(\alpha\mathbf{x} + \beta\mathbf{y}) + b, \\ &= \alpha g(\mathbf{x}) + \beta g(\mathbf{y}) + (\alpha + \beta)b, && (\alpha + \beta = 1) \\ &= \alpha(g(\mathbf{x}) + b) + \beta(g(\mathbf{y}) + b), \\ &= \alpha f(\mathbf{x}) + \beta f(\mathbf{y}). \end{aligned}$$

2. $\implies$ If f is affine, then there exists a linear function g such that:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad g(\mathbf{x}) = f(\mathbf{x}) - b, \qquad \text{for } \mathbf{x} \in V,$$

where $b \in \mathbb{R}$. Since f is bounded, there exists some $M_f > 0$ such that:

$$|f(\mathbf{x})| \leq M_f \|\mathbf{x}\|_H, \qquad b = |f(\mathbf{0})| \leq M_f \|\mathbf{0}\|_H = 0.$$

By the Riesz Representation Theorem, there exists some $\mathbf{a} \in H$ such that:

$$g(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle, \qquad \text{for } \mathbf{x} \in H.$$

Therefore,

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b, \qquad \text{for } \mathbf{x} \in H.$$

$\impliedby$ If there exists some $\mathbf{a} \in H$ and $b \in \mathbb{R}$ such that:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b, \qquad \text{for } \mathbf{x} \in H,$$

then, by Example 4.3, $\langle \mathbf{a}, \mathbf{x} \rangle$ is a linear function. Therefore, f is affine.

□

4.2 Hyperplane

Definition 4.7. Let V be a vector space over $\mathbb{R}$. A subset $U \subset V$ is a **(linear) subspace** of V if, for any $\mathbf{u}, \mathbf{v} \in U$ and $\alpha, \beta \in \mathbb{R}$,

$$\alpha\mathbf{u} + \beta\mathbf{v} \in U.$$

Definition 4.8. Let V be a vector space over $\mathbb{R}$. A set W is an **affine subspace** if:

$$W = \mathbf{a} + U = \{\mathbf{a} + \mathbf{u} : \mathbf{u} \in U\},$$

where $\mathbf{a} \in V$ and U is a linear subspace of V.

Let H be a Hilbert space and $\mathbf{a} \in H$. We denote:

$$S_{\mathbf{a},b} = \{\mathbf{x} \in H : \langle \mathbf{a}, \mathbf{x} \rangle = b\} \subset H.$$

Consider $b = 0$. For all $\mathbf{x}, \mathbf{y} \in S_{\mathbf{a},0}$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle \mathbf{a}, \alpha\mathbf{x} + \beta\mathbf{y} \rangle = \alpha \langle \mathbf{a}, \mathbf{x} \rangle + \beta \langle \mathbf{a}, \mathbf{y} \rangle = 0.$$

This means that $\alpha\mathbf{x} + \beta\mathbf{y} \in S_{\mathbf{a},0}$. Therefore, $S_{\mathbf{a},0}$ is a linear subspace of H.
For a general b, let $\mathbf{x}_0 \in S_{\mathbf{a},b}$. Then we have:

$$\langle \mathbf{a}, \mathbf{x}_0 \rangle = b.$$

For all $\mathbf{x} \in S_{\mathbf{a},b}$,

$$\begin{aligned} \langle \mathbf{a}, \mathbf{x} - \mathbf{x}_0 \rangle = \langle \mathbf{a}, \mathbf{x} \rangle - \langle \mathbf{a}, \mathbf{x}_0 \rangle = b - b = 0 &\implies \mathbf{x} - \mathbf{x}_0 \in S_{\mathbf{a},0}, \\ &\implies \mathbf{x} \in \mathbf{x}_0 + S_{\mathbf{a},0}, \\ &\implies S_{\mathbf{a},b} \subset \mathbf{x}_0 + S_{\mathbf{a},0}. \end{aligned}$$

For all $\mathbf{x} \in S_{\mathbf{a},0}$,

$$\begin{aligned} \langle \mathbf{a}, \mathbf{x} + \mathbf{x}_0 \rangle = \langle \mathbf{a}, \mathbf{x} \rangle + \langle \mathbf{a}, \mathbf{x}_0 \rangle = 0 + b = b &\implies \mathbf{x} + \mathbf{x}_0 \in S_{\mathbf{a},b}, \\ &\implies S_{\mathbf{a},0} + \mathbf{x}_0 \subset S_{\mathbf{a},b}. \end{aligned}$$

Therefore, we have $S_{\mathbf{a},b} = \mathbf{x}_0 + S_{\mathbf{a},0}$, and thus $S_{\mathbf{a},b}$ is an affine subspace. We can define such a set as a hyperplane.

Definition 4.9. A **hyperplane** S in the Hilbert space H is defined by:

$$S = \{\mathbf{x} \in H : \langle \mathbf{a}, \mathbf{x} \rangle = b\} \subset H,$$

where $\mathbf{a} \in H$ and $b \in \mathbb{R}$.

Remark 4.9.1. If H has a dimension of n, then the hyperplane S has a dimension of $n - 1$ or a codimension of 1.

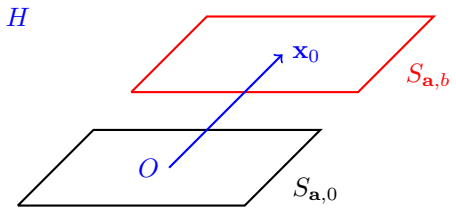


Figure 4.1: Hyperplanes

For any vectors that do not lie on the hyperplane, we can project those vectors onto the hyperplane.

Definition 4.10. Consider a hyperplane $S \subset H$. For any $\mathbf{y} \in H$, the vector on S that is closest to $\mathbf{y}$ is called the **projection** of $\mathbf{y}$ onto S, denoted by $P_S\mathbf{y}$. Equivalently:

$$P_S\mathbf{y} = \operatorname*{argmin}_{\mathbf{x}\in S} \|\mathbf{y} - \mathbf{x}\| .$$

How do we find the explicit form of such an equation?

Theorem 4.11. Given a hyperplane $S \subset H$, the vector $\mathbf{z} \in H$ is a solution of:

$$\min_{\mathbf{x}\in S} \|\mathbf{y} - \mathbf{x}\|$$

if and only if $\mathbf{z} \in S$ and $\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle = 0$ for all $\mathbf{x} \in S$.

Proof.

$\implies$ Assume that $\mathbf{z}$ is a solution of the minimization equation. Then $\mathbf{z} \in S$. For all $\mathbf{x} \in S$ and $t \in \mathbb{R}$,

$$\begin{aligned}\langle \mathbf{a}, \mathbf{z} + t(\mathbf{x} - \mathbf{z})\rangle &= \langle \mathbf{a}, \mathbf{z}\rangle + t(\langle \mathbf{a}, \mathbf{x}\rangle - \langle \mathbf{a}, \mathbf{z}\rangle),\\ &= b + t(b - b),\\ &= b.\end{aligned}$$

Therefore, $\mathbf{z} + t(\mathbf{x} - \mathbf{z}) \in S$. We have:

$$\begin{aligned}\|\mathbf{z} - \mathbf{y}\|^2 \le \|\mathbf{z} + t(\mathbf{x} - \mathbf{z}) - \mathbf{y}\|^2 &= \|\mathbf{z} - \mathbf{y}\|^2 + 2t\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle + t^2 \|\mathbf{x} - \mathbf{z}\|^2 ,\\ -t^2 \|\mathbf{x} - \mathbf{z}\|^2 &\le 2t\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle .\end{aligned}$$

Assume that $t > 0$. As $t \to 0^+$,

$$\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle \ge -\lim_{t\to 0^+} \frac{t}{2} \|\mathbf{x} - \mathbf{z}\|^2 = 0.$$

Assume that $t < 0$. As $t \to 0^-$,

$$\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle \le -\lim_{t\to 0^-} \frac{t}{2} \|\mathbf{x} - \mathbf{z}\|^2 = 0.$$

Therefore, $\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle = 0$ for all $\mathbf{x} \in S$.

$\impliedby$ Assume that $\mathbf{z} \in S$ and $\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle = 0$ for all $\mathbf{x} \in S$. This means that:

$$\begin{aligned}\|\mathbf{x} - \mathbf{y}\|^2 &= \|(\mathbf{x} - \mathbf{z}) + (\mathbf{z} - \mathbf{y})\|^2 ,\\ &= \|\mathbf{x} - \mathbf{z}\|^2 + 2\langle \mathbf{x} - \mathbf{z}, \mathbf{z} - \mathbf{y}\rangle + \|\mathbf{z} - \mathbf{y}\|^2 ,\\ &= \|\mathbf{x} - \mathbf{z}\|^2 + \|\mathbf{z} - \mathbf{y}\|^2 ,\\ &\ge \|\mathbf{z} - \mathbf{y}\|^2 .\end{aligned}$$

Therefore, we can find that:

$$\mathbf{z} = \operatorname*{argmin}_{\mathbf{x}\in S} \|\mathbf{x} - \mathbf{y}\|^2 .$$

□

Using the last theorem, we can obtain the explicit form of the projection.

Theorem 4.12. Let H be a Hilbert space and S be the hyperplane defined by:

$$S = \{\mathbf{x} \in H : \langle \mathbf{a}, \mathbf{x} \rangle = b\},$$

where $\mathbf{a} \in H$ and $b \in \mathbb{R}$. Given $\mathbf{y} \in H$, we have a unique solution:

$$P_S \mathbf{y} = \operatorname*{argmin}_{\mathbf{x} \in S} \|\mathbf{y} - \mathbf{x}\| = \mathbf{y} - \left(\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \right) \mathbf{a}.$$

Proof.
Let $\mathbf{z} = \mathbf{y} - \left(\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \right) \mathbf{a}$.

$$\begin{aligned} \langle \mathbf{a}, \mathbf{z} \rangle &= \langle \mathbf{a}, \mathbf{y} \rangle - \frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \langle \mathbf{a}, \mathbf{a} \rangle, \\ &= \langle \mathbf{a}, \mathbf{y} \rangle - (\langle \mathbf{a}, \mathbf{y} \rangle - b), \\ &= b. \end{aligned}$$

Therefore, $\mathbf{z} \in S$. For any $\mathbf{x} \in S$,

$$\begin{aligned} \langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z} \rangle &= \left\langle -\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \mathbf{a}, \mathbf{x} - \mathbf{z} \right\rangle, \\ &= -\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} (\langle \mathbf{a}, \mathbf{x} \rangle - \langle \mathbf{a}, \mathbf{z} \rangle), \\ &= -\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} (b - b), \\ &= 0. \end{aligned}$$

Therefore, by Theorem 4.11, $\mathbf{z}$ is a solution to:

$$\min_{\mathbf{x} \in S} \|\mathbf{y} - \mathbf{x}\|.$$

We prove that this solution is unique.
Suppose that it has two distinct solutions $\mathbf{z}_1$ and $\mathbf{z}_2$. Then it implies that $\mathbf{z}_1, \mathbf{z}_2 \in S$. By Theorem 4.11,

$$\langle \mathbf{z}_1 - \mathbf{y}, \mathbf{z}_2 - \mathbf{z}_1 \rangle = 0, \qquad \langle \mathbf{z}_2 - \mathbf{y}, \mathbf{z}_1 - \mathbf{z}_2 \rangle = \langle \mathbf{y} - \mathbf{z}_2, \mathbf{z}_2 - \mathbf{z}_1 \rangle = 0.$$

Adding the two identities, we have:

$$\begin{aligned} \langle \mathbf{z}_1 - \mathbf{y}, \mathbf{z}_2 - \mathbf{z}_1 \rangle + \langle \mathbf{y} - \mathbf{z}_2, \mathbf{z}_2 - \mathbf{z}_1 \rangle = \langle \mathbf{z}_1 - \mathbf{z}_2, \mathbf{z}_2 - \mathbf{z}_1 \rangle = 0 &\iff - \|\mathbf{z}_1 - \mathbf{z}_2\|^2 = 0, \\ &\iff \mathbf{z}_1 = \mathbf{z}_2. \end{aligned}$$

This results in a contradiction. Therefore, the solution $\mathbf{z}$ is unique. □

4.3 Linear approximation and Differentiation

Recall that for a function $f : \mathbb{R} \to \mathbb{R}$, its derivative at x is defined as:

$$f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}.$$

This is equivalent to:

$$\lim_{y \to x} \frac{|f(y) - (f(x) + f'(x)(y - x))|}{|y - x|} = 0.$$

Let $g(y) = f(x) + f'(x)(y - x)$. This function satisfies the following properties:

1. It is affine.
2. It passes through $(x, f(x))$.
3. $$\lim_{y \to x} \frac{|f(y) - g(y)|}{|y - x|} = 0,$$
   i.e., the error $f(y) - g(y) = o(|y - x|)$.

We can extend this to functions $f : H \to \mathbb{R}$, where H is a Hilbert space. Consider $\mathbf{x} \in H$. We want to find a function $g : H \to \mathbb{R}$ such that:

1. It is affine and bounded.
2. It passes through $(\mathbf{x}, f(\mathbf{x}))$.
3. $$\lim_{\mathbf{y} \to \mathbf{x}} \frac{|f(\mathbf{y}) - g(\mathbf{y})|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

By Theorem 4.6(2), let $g(\mathbf{y}) = \langle \mathbf{v}, \mathbf{y} \rangle + b$, where $\mathbf{v} \in H$ and $b \in \mathbb{R}$. We may define $\mathbf{v}$ and b such that:

$$g(\mathbf{y}) = \langle \mathbf{v}, \mathbf{x} \rangle + b = f(\mathbf{x}).$$

Therefore, the error limit becomes:

$$\lim_{\mathbf{y} \to \mathbf{x}} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

We have obtained the following definition.

Definition 4.13. Let H be a Hilbert space. Let $f : H \to \mathbb{R}$ be a function and $\mathbf{x} \in H$. Then, f is said to be **Fréchet differentiable** at $\mathbf{x}$ if there exists $\mathbf{v} \in H$ such that:

$$\lim_{\mathbf{y} \to \mathbf{x}} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

If f is differentiable at $\mathbf{x}$, then $\mathbf{v}$ is called the gradient at $\mathbf{x}$, denoted by $\nabla f(\mathbf{x})$.

Remark 4.13.1. With the definition of convergence for vectors, we can rewrite the formula as:

$$\lim_{\|\mathbf{y} - \mathbf{x}\| \to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

Remark 4.13.2. Fréchet differentiation depends on the inner product on H.

Example 4.11. Consider $f(\mathbf{x}) = \|\mathbf{x}\|^2$, where $\|\cdot\|$ is the norm on H. At $\mathbf{x} \in H$, for any $\mathbf{y} \in H$,

$$\begin{aligned} f(\mathbf{y}) = \|\mathbf{y}\|^2 &= \|\mathbf{x} + (\mathbf{y} - \mathbf{x})\|^2 \\ &= \|\mathbf{x}\|^2 + 2\langle \mathbf{x}, \mathbf{y} - \mathbf{x}\rangle + \|\mathbf{y} - \mathbf{x}\|^2 \\ &= f(\mathbf{x}) + \langle 2\mathbf{x}, \mathbf{y} - \mathbf{x}\rangle + \|\mathbf{y} - \mathbf{x}\|^2 . \end{aligned}$$

We find that:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle 2\mathbf{x}, \mathbf{y} - \mathbf{x}\rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{\|\mathbf{y} - \mathbf{x}\|^2}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

Therefore, $\nabla f(\mathbf{x}) = 2\mathbf{x}$.

Example 4.12. Consider $f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x}\rangle$ for some $\mathbf{a} \in H$. At $\mathbf{x} \in H$, for any $\mathbf{y} \in H$,

$$\begin{aligned} f(\mathbf{y}) = \langle \mathbf{a}, \mathbf{y}\rangle &= \langle \mathbf{a}, \mathbf{x}\rangle + \langle \mathbf{a}, \mathbf{y} - \mathbf{x}\rangle \\ &= f(\mathbf{x}) + \langle \mathbf{a}, \mathbf{y} - \mathbf{x}\rangle + 0. \end{aligned}$$

We find that:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{a}, \mathbf{y} - \mathbf{x}\rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

Therefore, $\nabla f(\mathbf{x}) = \mathbf{a}$.

Example 4.13. Consider $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\|^2$, where $\|\cdot\|$ is the norm on H and $\mathbf{a} \in H$. At $\mathbf{x} \in H$, for any $\mathbf{y} \in H$,

$$\begin{aligned} f(\mathbf{y}) &= \|\mathbf{y} - \mathbf{a}\|^2 \\ &= \|\mathbf{x} - \mathbf{a} + \mathbf{y} - \mathbf{x}\|^2 \\ &= \|\mathbf{x} - \mathbf{a}\|^2 + 2\langle \mathbf{x} - \mathbf{a}, \mathbf{y} - \mathbf{x}\rangle + \|\mathbf{y} - \mathbf{x}\|^2 \\ &= f(\mathbf{x}) + \langle 2(\mathbf{x} - \mathbf{a}), \mathbf{y} - \mathbf{x}\rangle + \|\mathbf{y} - \mathbf{x}\|^2 . \end{aligned}$$

We find that:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle 2(\mathbf{x} - \mathbf{a}), \mathbf{y} - \mathbf{x}\rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{\|\mathbf{y} - \mathbf{x}\|^2}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

Therefore, $\nabla f(\mathbf{x}) = 2(\mathbf{x} - \mathbf{a})$.

There are some properties of Fréchet differentiation.

Theorem 4.14. Fréchet differentiation has the following properties:

1. Fréchet differentiation is linear in f, i.e., for any $\alpha, \beta \in \mathbb{R}$ and $f, g : H \to \mathbb{R}$,

   $$\nabla(\alpha f + \beta g)(\mathbf{x}) = \alpha \nabla f(\mathbf{x}) + \beta \nabla g(\mathbf{x}), \qquad \text{for } \mathbf{x} \in H,$$

   provided that $\nabla f(\mathbf{x})$ and $\nabla g(\mathbf{x})$ exist.

2. (**Chain rule**) Let $f : H \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. Then $g \circ f : H \to \mathbb{R}$ and:

   $$\nabla(g \circ f)(\mathbf{x}) = g'(f(\mathbf{x})) \cdot \nabla f(\mathbf{x}), \qquad \text{for } \mathbf{x} \in H,$$

   provided that g and f are differentiable at $f(\mathbf{x})$ and $\mathbf{x}$, respectively.

Proof.

1. By definition, for any $\mathbf{x} \in H$,

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} = 0, \qquad \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|g(\mathbf{y}) - (g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} = 0.$$

We check if $\alpha\nabla f(\mathbf{x}) + \beta\nabla g(\mathbf{x})$ is the gradient of $\alpha f + \beta g$ at $\mathbf{x}$.

$$\begin{aligned}
0 &\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|(\alpha f + \beta g)(\mathbf{y}) - ((\alpha f + \beta g)(\mathbf{x}) + \langle \alpha\nabla f(\mathbf{x}) + \beta\nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|\alpha(f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)) + \beta(g(\mathbf{y}) - (g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle))|}{\|\mathbf{y}-\mathbf{x}\|} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \left(|\alpha| \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} + |\beta| \frac{|g(\mathbf{y}) - (g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} \right) = 0. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Triangle Inequality)}
\end{aligned}$$

Therefore, $\nabla(\alpha f + \beta g)(\mathbf{x}) = \alpha\nabla f(\mathbf{x}) + \beta\nabla g(\mathbf{x})$.

2. By definition, if f is differentiable at $\mathbf{x} \in H$ and g is differentiable at $t \in \mathbb{R}$,

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} = 0, \qquad \lim_{s\to t} \frac{g(s) - (g(t) + g'(t)(s-t))}{|s-t|} = 0.$$

We check if $g'(f(\mathbf{x})) \cdot \nabla f(\mathbf{x})$ is the gradient of $g \circ f$ at $\mathbf{x}$.

$$\begin{aligned}
0 &\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|g(f(\mathbf{y})) - (g(f(\mathbf{x})) + \langle g'(f(\mathbf{x})) \cdot \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|g(f(\mathbf{y})) - (g(f(\mathbf{x})) + g'(f(\mathbf{x}))(f(\mathbf{y}) - f(\mathbf{x})) - g'(f(\mathbf{x}))(f(\mathbf{y}) - f(\mathbf{x})) + \langle g'(f(\mathbf{x})) \cdot \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \left(\frac{|g(f(\mathbf{y})) - (g(f(\mathbf{x})) + g'(f(\mathbf{x}))(f(\mathbf{y}) - f(\mathbf{x}))|}{\|\mathbf{y}-\mathbf{x}\|} + |g'(f(\mathbf{x}))| \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle)|}{\|\mathbf{y}-\mathbf{x}\|} \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Triangle Inequality)} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|g(f(\mathbf{y})) - (g(f(\mathbf{x})) + g'(f(\mathbf{x}))(f(\mathbf{y}) - f(\mathbf{x}))|}{\|\mathbf{y}-\mathbf{x}\|} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|g(f(\mathbf{y})) - (g(f(\mathbf{x})) + g'(f(\mathbf{x}))(f(\mathbf{y}) - f(\mathbf{x}))|}{|f(\mathbf{y}) - f(\mathbf{x})|} \left(\frac{|f(\mathbf{y}) - f(\mathbf{x})|}{\|\mathbf{y}-\mathbf{x}\|} \right).
\end{aligned}$$

Now the only thing left to prove is that $f(\mathbf{y}) \to f(\mathbf{x})$ when $\|\mathbf{y}-\mathbf{x}\| \to 0$. Recall that:

$$f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle) = o(\|\mathbf{y}-\mathbf{x}\|).$$

Rearrange the terms:

$$\begin{aligned}
0 \le |f(\mathbf{y}) - f(\mathbf{x})| &= |\langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + o(\|\mathbf{y}-\mathbf{x}\|)| \\
&\le |\langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle| + o(\|\mathbf{y}-\mathbf{x}\|) && \text{(Triangle Inequality)} \\
&\le \|\nabla f(\mathbf{x})\| \, \|\mathbf{y}-\mathbf{x}\| + o(\|\mathbf{y}-\mathbf{x}\|) && \text{(Cauchy-Schwarz Inequality)} \\
&\le \|\nabla f(\mathbf{x})\| \, \|\mathbf{y}-\mathbf{x}\| + c \, \|\mathbf{y}-\mathbf{x}\|, && \text{for some } c \in \mathbb{R} \\
0 \le \frac{|f(\mathbf{y}) - f(\mathbf{x})|}{\|\mathbf{y}-\mathbf{x}\|} &\le \|\nabla f(\mathbf{x})\| + c.
\end{aligned}$$

This means as $\|\mathbf{y}-\mathbf{x}\| \to 0$, $f(\mathbf{y})$ has to converge to $f(\mathbf{x})$. By applying $s = f(\mathbf{y})$ and $t = f(\mathbf{x})$,

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|g(f(\mathbf{y})) - (g(f(\mathbf{x})) + g'(f(\mathbf{x}))(f(\mathbf{y}) - f(\mathbf{x}))|}{|f(\mathbf{y}) - f(\mathbf{x})|} = 0.$$

Therefore, $\nabla(g \circ f)(\mathbf{x}) = g'(f(\mathbf{x})) \cdot \nabla f(\mathbf{x})$.

□

Example 4.14. Consider $f(\mathbf{x}) = \|\mathbf{x}\|$, where $\|\cdot\|$ is the norm on H. Find its derivative for any $\mathbf{x} \in H$. Let:

$$f_1(\mathbf{x}) = \|\mathbf{x}\|^2, \qquad \text{for } \mathbf{x} \in H, \qquad f_2(t) = \sqrt{t}, \qquad \text{for } t \in \mathbb{R}_{\geq 0}, \qquad f(\mathbf{x}) = f_2(f_1(\mathbf{x})).$$

When $\mathbf{x} \neq 0$, both f_1 and f_2 are differentiable. By the chain rule,

$$\nabla f(\mathbf{x}) = f_2'(f_1(\mathbf{x})) \cdot \nabla f_1(\mathbf{x}) = \frac{1}{2\sqrt{\|\mathbf{x}\|^2}}(2\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

When $\mathbf{x} = \mathbf{0}$, $f_1(\mathbf{0}) = 0$. f_2 is not differentiable at $f_1(\mathbf{0})$. We cannot apply the chain rule.
In fact, f is not differentiable at $\mathbf{x} = \mathbf{0}$.

We can find Fréchet differentiation element-wise if the Hilbert space we are considering is $\mathbb{R}^n$.

Theorem 4.15. For functions $f : \mathbb{R}^n \to \mathbb{R}$, where $\mathbb{R}^n$ is with the standard inner product, if f is differentiable at $\mathbf{x} \in \mathbb{R}^n$, then:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{pmatrix}.$$

Proof.
Since f is differentiable at $\mathbf{x}$,

$$\lim_{\mathbf{y} \to \mathbf{x}} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

Choose $\mathbf{y} = \mathbf{x} + t\mathbf{e}_i$ where $t \in \mathbb{R}$ and let $g(t) = f(\mathbf{x} + t\mathbf{e}_i)$. We have:

$$\lim_{t \to 0} \frac{|f(\mathbf{x} + t\mathbf{e}_i) - (f(\mathbf{x}) + t \langle \nabla f(\mathbf{x}), \mathbf{e}_i \rangle)|}{|t|} = 0,$$
$$\lim_{t \to 0} \frac{|g(t) - (g(0) + \langle \nabla f(\mathbf{x}), \mathbf{e}_i \rangle (t - 0))|}{|t - 0|} = 0.$$

Therefore, by definition, $g'(0) = \langle \nabla f(\mathbf{x}), \mathbf{e}_i \rangle$. Moreover,

$$\langle \nabla f(\mathbf{x}), \mathbf{e}_i \rangle = \frac{\mathrm{d}}{\mathrm{d}t} g(t)\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{e}_i)\bigg|_{t=0} = \frac{\partial}{\partial x_i} f(\mathbf{x}).$$

Thus,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{pmatrix}.$$

□

Remark 4.15.1. Fréchet differentiation is consistent with the standard differentiation in multi-variable calculus if $\mathbb{R}^n$ is with the standard inner product.

Based on the last theorem, we used $\mathbf{y} = \mathbf{x} + t\mathbf{e}_i$. Can we extend $\mathbf{e}_i$ to any vector?

Lemma 4.16. Let $f : H \to \mathbb{R}$. Assume that f is differentiable at $\mathbf{x} \in H$. For any $\mathbf{u} \in H$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u})\bigg|_{t=0} = \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle .$$

Proof.
The proof is similar to the last theorem. Since f is differentiable at $\mathbf{x}$,

$$\lim_{\mathbf{y} \to \mathbf{x}} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

Choose $\mathbf{y} = \mathbf{x} + t\mathbf{u}$ where $t \in \mathbb{R}$ and let $g(t) = f(\mathbf{x} + t\mathbf{u})$. We have:

$$\lim_{t \to 0} \frac{|f(\mathbf{x} + t\mathbf{u}) - (f(\mathbf{x}) + t \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle)|}{|t|} = 0,$$
$$\lim_{t \to 0} \frac{|g(t) - (g(0) + \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle (t - 0))|}{|t - 0|} = 0.$$

Therefore, by definition, $g'(0) = \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle$. Moreover,

$$\langle \nabla f(\mathbf{x}), \mathbf{u} \rangle = \frac{\mathrm{d}}{\mathrm{d}t} g(t)\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u})\bigg|_{t=0} .$$

□

Remark 4.16.1. If $\|\mathbf{u}\| = 1$, then $\langle \nabla f(\mathbf{x}), \mathbf{u} \rangle$ is the **directional derivative** along the $\mathbf{u}$ direction.

We can generalize this to all t.

Theorem 4.17. Let $f : H \to \mathbb{R}$. For any $\mathbf{x}, \mathbf{u} \in H$ and $t \in \mathbb{R}$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u}) = \langle \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \rangle ,$$

if f is differentiable at $\mathbf{x} + t\mathbf{u}$.

Proof.
Let $g(t) = f(\mathbf{x} + t\mathbf{u})$. Fix an arbitrary t_0, we have:

$$g(t_0 + t) = f(\mathbf{x} + t_0\mathbf{u} + t\mathbf{u}).$$

From Lemma 4.16, if f is differentiable at $\mathbf{z} \in H$, then:

$$\frac{\mathrm{d}}{\mathrm{d}t} g'(t_0) = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t_0\mathbf{u} + t\mathbf{u})\bigg|_{t=0} = \langle \nabla f(\mathbf{x} + t_0\mathbf{u}), \mathbf{u} \rangle .$$

Therefore, since t_0 is arbitrary,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u}) = \langle \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \rangle .$$

□

Recall the Taylor expansion for functions $f : \mathbb{R} \to \mathbb{R}$:

$$f(y) = f(x) + f'(x)(y - x) + o((y - x)^k), \qquad \text{for } y \in \mathbb{R}.$$

Similarly, we can extend it to $f : H \to \mathbb{R}$ by Fréchet differentiation.

Theorem 4.18. Assume that $f : H \to \mathbb{R}$ is differentiable at $\mathbf{x} \in H$. The **Taylor expansion** of f is:

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|), \qquad \text{for } \mathbf{y} \in H.$$

Up to this point, we have only considered the Hilbert space H. What about functions on Banach spaces, which may not have an inner product? Let $f : V \to \mathbb{R}$, where V is a Banach space. Let $\mathbf{x} \in V$. We want to find $g : V \to \mathbb{R}$ such that:

1. It is affine.
2. $g(\mathbf{x}) = f(\mathbf{x})$.
3. The error is $o(\|\mathbf{y} - \mathbf{x}\|)$.

We cannot use the inner product. Instead, we let $g(\mathbf{x}) = L\mathbf{x} + \mathbf{a}$ for all $\mathbf{x} \in V$, where $L : V \to \mathbb{R}$ is linear and $\mathbf{a} \in \mathbb{R}$. Therefore,

$$g(\mathbf{y}) = L\mathbf{y} + \mathbf{a} = L\mathbf{y} - L\mathbf{x} + L\mathbf{x} + \mathbf{a} = L(\mathbf{y} - \mathbf{x}) + f(\mathbf{x}).$$

The error limit becomes:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + L(\mathbf{y} - \mathbf{x}))|}{\|\mathbf{y} - \mathbf{x}\|}.$$

We now have a more general definition for differentiation.

Definition 4.19. Let V be a Banach space. Let $f : V \to \mathbb{R}$ be a function and $\mathbf{x} \in V$. Then, f is said to be **(Fréchet) differentiable** at $\mathbf{x}$ if there exists a linear function $L : V \to \mathbb{R}$ such that:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|\to 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + L(\mathbf{y} - \mathbf{x}))|}{\|\mathbf{y} - \mathbf{x}\|} = 0.$$

The linear function L is called the **differentiation** of f at $\mathbf{x}$, defined by:

$$Df(\mathbf{x}) = L.$$

We will continue discussing this differentiation in the next chapter. Read Appendix D (Linear Regression), E (Kernel Ridge Regression), F (Linear Classification), G (Solvability and Optimality), and H (Gradient Descent) to see the case studies for Chapter 4.

Chapter 5

Linear Transformations or Linear Operators

In the last chapter, we discussed linear functions $f : V \to \mathbb{R}$ in more detail. What about operators that map from one vector space to another? This chapter focuses on such operators, allowing us to generalize the differentiation of functions.

5.1 Linear Transformations or Linear Operators

Definition 5.1. Let V, W be two vector spaces. A map $L : V \to W$ is a **linear transformation / linear operator** if, for all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$L(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha L\mathbf{x} + \beta L\mathbf{y}.$$

Remark 5.1.1. Parentheses are often omitted. For example, $L\mathbf{x}$ means $L(\mathbf{x})$.

Example 5.1. Let $f : V \to \mathbb{R}$ be a linear function. Then f is a linear transformation.

Example 5.2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Define $L : \mathbb{R}^n \to \mathbb{R}^m$ by:

$$L\mathbf{x} = \mathbf{A}\mathbf{x}, \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

It is a linear transformation because, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$,

$$\mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{A}\mathbf{x} + \beta\mathbf{A}\mathbf{y}.$$

In fact, for any linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$, there exists $\mathbf{A} \in \mathbb{R}^{m\times n}$ such that:

$$L\mathbf{x} = \mathbf{A}\mathbf{x}, \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

Example 5.3. Let $\mathbf{a}_1, \cdots, \mathbf{a}_k \in H$. Define $L : H \to \mathbb{R}^k$ by:

$$L\mathbf{x} = \begin{pmatrix} \langle \mathbf{a}_1, \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{a}_k, \mathbf{x} \rangle \end{pmatrix} \in \mathbb{R}^k, \qquad \text{for } \mathbf{x} \in H.$$

Then L is a linear transformation.

Example 5.4. Let:

$$V = \{f : f \text{ and } f' \text{ are continuous on } [a,b]\},$$
$$W = \{f : f \text{ is continuous on } [a,b]\}.$$

Define $D : V \to W$ by:

$$Df = f', \qquad \text{for } f \in V.$$

Then D is a linear transformation.

Example 5.5. Consider $\mathcal{C}[-1,1]$ and $\mathcal{C}[0,2]$. Define $T : \mathcal{C}[-1,1] \to \mathcal{C}[0,2]$ by:

$$(Tf)(t) = f(t-1), \qquad \text{for any } t \in [0,2], \qquad \text{for } f \in \mathcal{C}[-1,1].$$

Then T is linear.

We have the following properties:

Theorem 5.2. Let V, W be two non-trivial vector spaces. If V and W are finite-dimensional, then for any linear transformation $T : V \to W$, there exists a unique matrix $\mathbf{A}$ such that:

$$T\mathbf{v} = \mathbf{A}\mathbf{v}, \qquad \text{for } \mathbf{v} \in V.$$

Remark 5.2.1. Due to this fact, matrices are often used to represent linear transformations.

Definition 5.3. Let V, W be two vector spaces, and let $\mathbf{A}, \mathbf{B} : V \to W$ be linear operators. We define $\mathbf{A} + \mathbf{B} : V \to W$ by:

$$(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}, \qquad \text{for } \mathbf{x} \in V.$$

For any $\alpha \in \mathbb{R}$, we define $\alpha\mathbf{A} : V \to W$ by:

$$(\alpha\mathbf{A})\mathbf{x} = \alpha\mathbf{A}\mathbf{x}.$$

We may define some norms for the linear operators.

Theorem 5.4. For any linear operator $\mathbf{A} : V \to W$,

$$\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{A}\mathbf{x}\|_W .$$

This is a norm. If $\|\mathbf{A}\| < +\infty$, then $\mathbf{A}$ is said to be bounded.

Proof.

1. We can easily find that:
$$\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{A}\mathbf{x}\|_W \geq 0.$$
Moreover,
$$\begin{aligned} \|\mathbf{A}\| = 0 &\iff \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{A}\mathbf{x}\|_W = 0 \\ &\iff \|\mathbf{A}\mathbf{x}\|_W = 0, \qquad \text{for } \mathbf{x} \in V \text{ satisfying } \|\mathbf{x}\|_V = 1 \\ &\iff \left\| \mathbf{A} \frac{\mathbf{y}}{\|\mathbf{y}\|_V} \right\|_W = 0 \iff \mathbf{A}\mathbf{y} = \mathbf{0}, \qquad \text{for } \mathbf{y} \in V \\ &\iff \mathbf{A} = \mathbf{O}. \end{aligned}$$

2. For any $\mathbf{A} : V \to W$, $\alpha \in \mathbb{R}$,
$$\|\alpha\mathbf{A}\| = \sup_{\|\mathbf{x}\|_V = 1} \|(\alpha\mathbf{A})\mathbf{x}\|_W = |\alpha| \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{A}\mathbf{x}\|_W = |\alpha| \, \|\mathbf{A}\| .$$

3. For any linear operators $\mathbf{A}, \mathbf{B} : V \to W$,
$$\|\mathbf{A} + \mathbf{B}\| = \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}\|_W \leq \sup_{\|\mathbf{x}\|_V = 1} (\|\mathbf{A}\mathbf{x}\|_W + \|\mathbf{B}\mathbf{x}\|_W) \leq \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{A}\mathbf{x}\|_W + \sup_{\|\mathbf{x}\|_V = 1} \|\mathbf{B}\mathbf{x}\|_W = \|\mathbf{A}\| + \|\mathbf{B}\| .$$

Therefore, it is a norm. □

In addition, it satisfies additional properties.

Theorem 5.5. For any linear operator $\mathbf{A} : V \to W$,

$$\|\mathbf{Ax}\|_W \leq \|\mathbf{A}\| \|\mathbf{x}\|_V , \qquad \text{for } \mathbf{x} \in V.$$

Proof.
If $\mathbf{x} = \mathbf{0}$, then:

$$\|\mathbf{Ax}\|_W = 0 \leq 0 = \|\mathbf{A}\| \|\mathbf{x}\|_V .$$

If $\mathbf{x} \neq 0$, then:

$$\|\mathbf{A}\| = \sup_{\|\mathbf{z}\|_V = 1} \|\mathbf{Az}\|_W \geq \left\| \mathbf{A} \frac{\mathbf{x}}{\|\mathbf{x}\|_V} \right\|_W = \frac{1}{\|\mathbf{x}\|_V} \|\mathbf{Ax}\|_W .$$

After rearranging, we have $\|\mathbf{A}\| \|\mathbf{x}\|_V \geq \|\mathbf{Ax}\|_W$. □

Theorem 5.6. For any linear operators $\mathbf{A} : V_2 \to V_3$ and $\mathbf{B} : V_1 \to V_2$,

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\| .$$

Proof.
By Theorem 5.5,

$$\|\mathbf{AB}\| = \sup_{\|\mathbf{x}\|_{V_1} = 1} \|\mathbf{ABx}\|_{V_3} = \|\mathbf{A}\| \sup_{\|\mathbf{x}\|_{V_1} = 1} \|\mathbf{Bx}\|_{V_2} = \|\mathbf{A}\| \|\mathbf{B}\| .$$

□

Example 5.6. Consider $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$. We endow them with the 2-norm. For any $\mathbf{A} \in \mathbb{R}^{m \times n}$, we call the corresponding operator norm the **2-norm**, defined by:

$$\|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|\mathbf{Ax}\|_2 = \sqrt{\sup_{\|\mathbf{x}\|_2 = 1} \|\mathbf{Ax}\|_2^2} = \sqrt{\sup_{\mathbf{x}^T\mathbf{x} = 1} \mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{x}} = \sqrt{\text{Max eigenvalue of } \mathbf{A}^T\mathbf{A}}.$$

Example 5.7. Consider $\mathbf{A} \in \mathbb{R}^{m \times n}$ as a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let $\mathbb{R}^n$ and $\mathbb{R}^m$ be endowed with 1-norms. We call the resulting operator norm the **1-norm**, defined by:

$$\|\mathbf{A}\|_1 = \sup_{\|\mathbf{x}\|_1 = 1} \|\mathbf{Ax}\|_1 .$$

Theorem 5.7. Let $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_n) \in \mathbb{R}^{m \times n}$. Then:

$$\|\mathbf{A}\|_1 = \max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1 .$$

Proof.
For any $\mathbf{x} \in \mathbb{R}^n$ satisfying $\|\mathbf{x}\|_1 = 1$,

$$\|\mathbf{Ax}\|_1 = \left\| \sum_{i=1}^n x_i \mathbf{a}_i \right\|_1 \leq \sum_{i=1}^n \|x_i \mathbf{a}_i\|_1 = \sum_{i=1}^n |x_i| \|\mathbf{a}_i\|_1 \leq \sum_{j=1}^n |x_j| \max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1 = \|\mathbf{x}\|_1 \max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1 = \max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1 .$$

By taking the supremum over $\mathbf{x}$, we get $\|\mathbf{A}\|_1 \leq \max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1$.
Let $i_m = \text{argmax}_{1 \leq i \leq n} \|\mathbf{a}_i\|_1$. Then:

$$\max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1 = \|\mathbf{a}_{i_m}\|_1 = \|\mathbf{A}\mathbf{e}_{i_m}\|_1 \leq \sup_{\|\mathbf{x}\|_1} \|\mathbf{Ax}\|_1 = \|\mathbf{A}\|_1 .$$

Therefore, $\|\mathbf{A}\|_1 = \max_{1 \leq i \leq n} \|\mathbf{a}_i\|_1$. □

Example 5.8. Consider $\mathbf{A} \in \mathbb{R}^{m\times n}$ as a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let $\mathbb{R}^n$ and $\mathbb{R}^m$ be endowed with infinity norms. We call the resulting operator norm the **∞-norm**, defined by:

$$\|\mathbf{A}\|_\infty = \sup_{\|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty .$$

Theorem 5.8. Let $\mathbf{A} = (\mathbf{a}^{(1)}, \cdots, \mathbf{a}^{(m)})^T \in \mathbb{R}^{m\times n}$. Then:

$$\|\mathbf{A}\|_\infty = \max_{1\le i\le m} \|\mathbf{a}^{(i)}\|_1.$$

Proof.
For any $\mathbf{x} \in \mathbb{R}^n$ satisfying $\|\mathbf{x}\|_\infty = 1$,

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_\infty &= \max_{1\le i\le m} \left|\left\langle \mathbf{a}^{(i)}, \mathbf{x}\right\rangle\right| \\
&= \max_{1\le i\le m} \left|\sum_{j=1}^n a_j^{(i)} x_j\right| \\
&\le \max_{1\le i\le m} \sum_{j=1}^n \left|a_j^{(i)}\right| |x_j| \\
&\le \max_{1\le j\le n} |x_j| \max_{1\le i\le m} \sum_{j=1}^n \left|a_j^{(i)}\right| = \|\mathbf{x}\|_\infty \max_{1\le i\le m} \|\mathbf{a}^{(i)}\|_1 = \max_{1\le i\le m} \|\mathbf{a}^{(i)}\|_1.
\end{aligned}$$

By taking the supremum over $\mathbf{x}$, we get $\|\mathbf{A}\|_\infty \le \max_{1\le i\le m}\|\mathbf{a}^{(i)}\|_1$.
Without loss of generality, assume that:

$$\|\mathbf{a}^{(1)}\|_1 = \max_{1\le i\le m} \|\mathbf{a}^{(i)}\|_1.$$

We construct $\mathbf{x}$ such that $x_i = \operatorname{sgn}(a_i^{(1)})$, which has $\|\mathbf{x}\|_\infty = 1$. Then $\left|\left\langle \mathbf{a}^{(1)}, \mathbf{x}\right\rangle\right| = \|\mathbf{a}^{(1)}\|_1$.
Therefore, $\|\mathbf{A}\|_\infty = \max_{1\le i\le m}\|\mathbf{a}^{(i)}\|_1$. □

Example 5.9. Let $T : \mathcal{C}[-1,1] \to \mathcal{C}[0,2]$ be the translation operator defined by:

$$(Tf)(t) = f(t-1), \qquad \text{for any } t \in [0,2], \qquad \text{for } f \in \mathcal{C}[-1,1].$$

We have found that T is linear. The norm on $\mathcal{C}[a,b]$ is:

$$\|f\|_\infty = \max_{t\in[a,b]} |f(t)| .$$

Then the norm of T is:

$$\|T\| = \sup_{\|f\|_\infty=1} \|Tf\|_\infty = \sup_{\max_{t\in[-1,1]}|f(t)|=1} \max_{t\in[0,2]} |f(t-1)| = 1.$$

Example 5.10. Let:

$$V = \{f : f \text{ and } f' \text{ are continuous on } [a,b]\}, \qquad W = \{f : f \text{ is continuous on } [a,b]\}$$

both equipped with the norm $\|\cdot\|_\infty$. Consider the differentiation operator $D : V \to W$ defined by:

$$Df = f', \qquad \text{for } f \in V.$$

For $t \in [0,1]$, consider $f_k(t) = \sin(2\pi kt)$, where $k \in \mathbb{N}$. Then $f_k'(t) = 2\pi k\cos(2\pi kt)$. We have $f_k \in V$ and:

$$\|Df_k\|_\infty = 2\pi k \|\cos(2\pi kt)\|_\infty = 2\pi k.$$

Hence:

$$\|D\| = \sup_{\|f\|_\infty=1} \|Df\|_\infty \ge \lim_{k\to\infty} \|Df_k\|_\infty = \lim_{k\to\infty} 2\pi k = \infty.$$

Therefore, $\|D\| = +\infty$.

We can group all linear operators from one vector space to another into a set.

Definition 5.9. Let V, W be two normed vector spaces with $\|\cdot\|_V$ and $\|\cdot\|_W$, respectively. The normed vector space of all linear and bounded operators is defined as:

$$\mathscr{L}(V,W) = \{\mathbf{A} : \mathbf{A} \text{ is linear from } V \to W \text{ and } \|\mathbf{A}\| < +\infty\}.$$

Remark 5.9.1. If both V and W are Banach spaces, then $\mathscr{L}(V,W)$ is also a Banach space.

Example 5.11. If $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ with any norm, then $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m\times n}$.

Example 5.12. Let $\mathbf{a}_1, \cdots, \mathbf{a}_k \in H$, where H is a Hilbert space. Define $L : H \to \mathbb{R}^k$ by:

$$L\mathbf{x} = \begin{pmatrix} \langle \mathbf{a}_1, \mathbf{x}\rangle \\ \vdots \\ \langle \mathbf{a}_k, \mathbf{x}\rangle \end{pmatrix}, \qquad \text{for } \mathbf{x} \in H.$$

We know L is linear, but is it bounded?

$$\|L\| = \sup_{\|\mathbf{x}\|=1} \|L\mathbf{x}\| = \sup_{\|\mathbf{x}\|=1} \sqrt{\sum_{i=1}^{k} (\langle \mathbf{a}_i, \mathbf{x}\rangle)^2} \le \sup_{\|\mathbf{x}\|=1} \sqrt{\sum_{i=1}^{k} \|\mathbf{a}_i\|^2 \|\mathbf{x}\|^2} = \sqrt{\sum_{i=1}^{k} \|\mathbf{a}_i\|^2} < +\infty.$$

Therefore, $L \in \mathscr{L}(H, \mathbb{R}^k)$.

Example 5.13. Let $T : L^2(-1,1) \to L^2(0,2)$ be the translation operator defined by:

$$(Tf)(t) = f(t-1), \qquad \text{for all } t \in (0,2), \qquad \text{for } f \in L^2(-1,1).$$

We prove that T is linear and bounded.

1. For any $f, g \in L^2(-1,1)$ and $\alpha, \beta \in \mathbb{R}$,

$$T(\alpha f + \beta g)(t) = (\alpha f + \beta g)(t-1) = \alpha f(t-1) + \beta g(t-1) = \alpha (Tf)(t) + \beta (Tg)(t).$$

   Therefore, T is linear.

2. $$\|T\| = \sup_{\|f\|_2=1} \|Tf\|_2 = \sup_{\int_{-1}^{1} |f(t)|^2\, dt = 1} \sqrt{\int_0^2 |f(t-1)|^2\, dt} = 1.$$

   Therefore, T is bounded.

Thus, $T \in \mathscr{L}(L^2(-1,1), L^2(0,2))$.

Consider the Hilbert spaces H_1 and H_2. Both have an inner product. If there is a linear operator acting on vectors in H_1 to H_2, is there a corresponding linear operator acting on vectors in H_2 to H_1?

Definition 5.10. Let H_1, H_2 be two Hilbert spaces with $\langle \cdot, \cdot \rangle_{H_1}$ and $\langle \cdot, \cdot \rangle_{H_2}$, respectively. Let $\mathbf{A} \in \mathscr{L}(H_1, H_2)$. The **adjoint operator** $\mathbf{A}^* : H_2 \to H_1$ is the unique bounded linear operator satisfying:

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle_{H_2} = \langle \mathbf{x}, \mathbf{A}^* \mathbf{y} \rangle_{H_1}, \qquad \text{for } \mathbf{x} \in H_1, \mathbf{y} \in H_2.$$

Example 5.14. Consider $\mathbf{A} \in \mathbb{R}^{m \times n} = \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$. Then for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T \mathbf{A}\mathbf{x} = (\mathbf{A}^T \mathbf{y})^T \mathbf{x} = \langle \mathbf{x}, \mathbf{A}^T \mathbf{y} \rangle .$$

We know that $\mathbf{A}^T \in \mathscr{L}(\mathbb{R}^m, \mathbb{R}^n)$. Therefore, the adjoint of $\mathbf{A}$ is $\mathbf{A}^T$.

Remark 5.10.1. The adjoint is an extension of the matrix transpose.

Example 5.15. Let H be a Hilbert space. Let $f : H \to \mathbb{R}$ be linear and bounded. We have:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle_H, \qquad \text{for } \mathbf{a} \in H.$$

For all $\mathbf{x} \in H$ and $y \in \mathbb{R}$,

$$\langle f(\mathbf{x}), y \rangle_{\mathbb{R}} = f(\mathbf{x}) y = y \langle \mathbf{a}, \mathbf{x} \rangle_H = \langle \mathbf{x}, y\mathbf{a} \rangle_H .$$

We may check whether $g(y) = y\mathbf{a}$ is linear and bounded for all $y \in \mathbb{R}$.

1. For any $y_1, y_2, \alpha, \beta \in \mathbb{R}$,

$$\begin{aligned} g(\alpha y_1 + \beta y_2) &= (\alpha y_1 + \beta y_2)\mathbf{a} \\ &= \alpha y_1 \mathbf{a} + \beta y_2 \mathbf{a} \\ &= \alpha g(y_1) + \beta g(y_2). \end{aligned}$$

2.

$$\|g\| = \sup_{|y|=1} \|g(y)\|_H = \sup_{|y|=1} \|y\mathbf{a}\|_H = \|\mathbf{a}\|_H < \infty.$$

Thus, $g \in \mathscr{L}(\mathbb{R}, H)$ and $f^*(y) = y\mathbf{a}$.

Example 5.16. Let $\mathbf{a}_1, \cdots, \mathbf{a}_k \in H$, where H is a Hilbert space. Define $L : H \to \mathbb{R}^k$ by:

$$L\mathbf{x} = \begin{pmatrix} \langle \mathbf{a}_1, \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{a}_k, \mathbf{x} \rangle \end{pmatrix}, \qquad \text{for } \mathbf{x} \in H.$$

From Example 5.12, $L \in \mathscr{L}(H, \mathbb{R}^k)$. For all $\mathbf{x} \in H$ and $\mathbf{y} \in \mathbb{R}^k$,

$$\langle L\mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^k} = \sum_{i=1}^{k} \langle \mathbf{a}_i, \mathbf{x} \rangle_H \, y_i = \sum_{i=1}^{k} \langle y_i \mathbf{a}_i, \mathbf{x} \rangle_H = \left\langle \mathbf{x}, \sum_{i=1}^{k} y_i \mathbf{a}_i \right\rangle_H .$$

Therefore, we can define $L^* : \mathbb{R}^k \to H$ by:

$$L^*\mathbf{y} = \sum_{i=1}^{k} y_i \mathbf{a}_i, \qquad \text{for } \mathbf{y} \in \mathbb{R}^k.$$

It remains to show that $L^* \in \mathscr{L}(\mathbb{R}^k, H)$.

1. For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$ and $\alpha, \beta \in \mathbb{R}$,

$$L^*(\alpha\mathbf{u} + \beta\mathbf{v}) = \sum_{i=1}^{k} (\alpha u_i + \beta v_i)\mathbf{a}_i = \alpha \sum_{i=1}^{k} u_i \mathbf{a}_i + \beta \sum_{i=1}^{k} v_i \mathbf{a}_i = \alpha L^*(\mathbf{u}) + \beta L^*(\mathbf{v}).$$

   Therefore, L^* is linear.

2.

$$\begin{aligned} \|L^*\| &= \sup_{\|\mathbf{y}\|=1} \|L^*\mathbf{y}\| = \sup_{\|\mathbf{y}\|=1} \left\| \sum_{i=1}^{k} y_i \mathbf{a}_i \right\| \\ &\le \sup_{\|\mathbf{y}\|=1} \sum_{i=1}^{k} |y_i| \, \|\mathbf{a}_i\| = \sup_{\|\mathbf{y}\|=1} \left\langle \begin{pmatrix} |y_1| \\ \vdots \\ |y_k| \end{pmatrix}, \begin{pmatrix} \|\mathbf{a}_1\| \\ \vdots \\ \|\mathbf{a}_k\| \end{pmatrix} \right\rangle \\ &\le \sup_{\|\mathbf{y}\|=1} \left\| \begin{pmatrix} |y_1| \\ \vdots \\ |y_k| \end{pmatrix} \right\| \left\| \begin{pmatrix} \|\mathbf{a}_1\| \\ \vdots \\ \|\mathbf{a}_k\| \end{pmatrix} \right\| = \left\| \begin{pmatrix} \|\mathbf{a}_1\| \\ \vdots \\ \|\mathbf{a}_k\| \end{pmatrix} \right\| = \sqrt{\sum_{i=1}^{k} \|\mathbf{a}_i\|^2} < \infty. \end{aligned}$$

   Therefore, $L^* \in \mathscr{L}(\mathbb{R}^k, H)$ and thus L^* is the adjoint of L.

Example 5.17. Let $T : L^2(-1, 1) \to L^2(0, 2)$ be the translation operator defined by:

$$(Tf)(t) = f(t - 1), \qquad \text{for all } t \in (0, 2), \qquad \text{for } f \in L^2(-1, 1).$$

From Example 5.13, we have $T \in \mathscr{L}(L^2(-1, 1), L^2(0, 2))$. Let $f \in L^2(-1, 1)$ and $g \in L^2(0, 2)$. We have:

$$\langle Tf, g \rangle = \int_0^2 (Tf)(t) g(t) \, dt = \int_0^2 f(t - 1) g(t) \, dt = \int_{-1}^1 f(s) g(s + 1) \, ds.$$

We may define $T^*g(t) = g(t + 1)$ for any $t \in (-1, 1)$ and $g \in L^2(0, 2)$.
Similarly, $T^* \in \mathscr{L}(L^2(0, 2), L^2(-1, 1))$, and thus $T^*g(t) = g(t + 1)$.

5.2 Linear Approximation and Differentiation of Transformations

Recall the definition of differentiation in the previous chapter. We can generalize it to linear transformations. Let V, W be two Banach spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_W$, respectively. Let $F : V \to W$ be a map (not necessarily linear). At $\mathbf{x} \in V$, the linear approximation passing through $(\mathbf{x}, F(\mathbf{x}))$ is given by:

$$F(\mathbf{y}) \approx F(\mathbf{x}) + L(\mathbf{y} - \mathbf{x}), \qquad \text{for } \mathbf{y} \in V,$$

where $L \in \mathscr{L}(V, W)$. If the error is $o(\|\mathbf{y} - \mathbf{x}\|_V)$, then we call L the differentiation of F at $\mathbf{x}$.

Definition 5.11. Let V, W be Banach spaces. Let $F : V \to W$ be a map and $\mathbf{x} \in V$. Then, F is said to be **(Fréchet) differentiable** at $\mathbf{x}$ if there exists $L \in \mathscr{L}(V, W)$ such that:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|F(\mathbf{y}) - (F(\mathbf{x}) + L(\mathbf{y} - \mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} = 0.$$

L is called the **differentiation** of F at $\mathbf{x}$, defined by:

$$DF(\mathbf{x}) = L.$$

Remark 5.11.1. $DF(\mathbf{x})$ is a function, not a vector in W.

Remark 5.11.2. We can substitute $\mathbf{y} = \mathbf{x} + \mathbf{h}$. Then the definition becomes:

$$\lim_{\|\mathbf{h}\|_V \to 0} \frac{\|F(\mathbf{x} + \mathbf{h}) - (F(\mathbf{x}) + L\mathbf{h})\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} = 0.$$

$L\mathbf{h}$ is called the **differential** of F at $\mathbf{x}$, defined by:

$$DF(\mathbf{x})(\mathbf{h}) = L\mathbf{h}.$$

Remark 5.11.3. Let V, W_1, W_2 be Banach spaces, and let $f_1 : V \to W_1$, $f_2 : V \to W_2$. Define:

$$F(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x})) \in W_1 \times W_2, \qquad \text{for } \mathbf{x} \in V.$$

If f_1 and f_2 are differentiable, then:

$$DF(\mathbf{x}) = (Df_1(\mathbf{x}), Df_2(\mathbf{x})).$$

Example 5.18. Let $\mathbf{A} \in \mathscr{L}(V, W)$. Then, for all $\mathbf{x} \in V$,

$$\lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|\mathbf{A}\mathbf{y} - (\mathbf{A}\mathbf{x} + \mathbf{A}(\mathbf{y} - \mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} = 0.$$

Therefore, $D\mathbf{A}\mathbf{x} = \mathbf{A} \in \mathscr{L}(V, W)$.

This example is related to the following theorem.

Theorem 5.12. Let V, W be Banach spaces. Then, for any linear $F : V \to W$,

$$DF(\mathbf{x}) = F, \qquad \text{for } \mathbf{x} \in V.$$

Proof.
If F is linear,

$$\lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|F(\mathbf{y}) - (F(\mathbf{x}) + F(\mathbf{y} - \mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} = \lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|F(\mathbf{y}) - (F(\mathbf{x}) + F(\mathbf{y}) - F(\mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} = 0.$$

□

Example 5.19. Let H be a Hilbert space and $F : H \to \mathbb{R}$. Then, the differentiation of F at $\mathbf{x} \in H$ is:

$$DF(\mathbf{x})(\mathbf{y}) = \langle \nabla F(\mathbf{x}), \mathbf{y} \rangle, \qquad \text{for } \mathbf{y} \in H.$$

Remark 5.12.1. The gradient is defined only for $F : H \to \mathbb{R}$.

Example 5.20. Let $\mathbf{A} \in \mathscr{L}(H_1, H_2)$, where H_1, H_2 are Hilbert spaces. Let $f : H_1 \to \mathbb{R}$ be defined by:

$$f(\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{H_2}^2, \qquad \text{for } \mathbf{x} \in H_1,$$

where $\mathbf{b} \in H_2$. How do we find $\nabla f(\mathbf{x})$?

$$\begin{aligned} f(\mathbf{y}) &= \frac{1}{2} \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_{H_2}^2 \\ &= \frac{1}{2} \|(\mathbf{A}\mathbf{x} - \mathbf{b}) + (\mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{x})\|_{H_2}^2 \\ &= \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{H_2}^2 + \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}(\mathbf{y} - \mathbf{x}) \rangle_{H_2} + \frac{1}{2} \|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_{H_2}^2 \\ &= f(\mathbf{x}) + \langle \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{b}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_{H_2}^2. \end{aligned}$$

We may find that as $\|\mathbf{y} - \mathbf{x}\|_{H_1} \to 0$,

$$\frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{b}), \mathbf{y} - \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|_{H_1}} = \frac{\frac{1}{2} \|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_{H_2}^2}{\|\mathbf{y} - \mathbf{x}\|_{H_1}} \leq \frac{1}{2} \|\mathbf{A}\|^2 \frac{\|\mathbf{y} - \mathbf{x}\|_{H_1}^2}{\|\mathbf{y} - \mathbf{x}\|_{H_1}} \to 0.$$

Therefore, $\nabla f(\mathbf{x}) = \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{b})$ and $Df(\mathbf{x})(\mathbf{y}) = \langle \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{b}), \mathbf{y} \rangle$.

It has some properties.

Theorem 5.13. Let V, W be Banach spaces. For any $F, G : V \to W$ and $\alpha, \beta \in \mathbb{R}$,

$$D(\alpha F + \beta G)(\mathbf{x}) = \alpha DF(\mathbf{x}) + \beta DG(\mathbf{x}), \qquad \text{for } \mathbf{x} \in V.$$

Proof.
By definition, for all $\mathbf{x} \in V$,

$$\begin{aligned} 0 &\leq \lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|(\alpha F + \beta G)(\mathbf{y}) - ((\alpha F + \beta G)(\mathbf{x}) + (\alpha DF(\mathbf{x}) + \beta DG(\mathbf{x}))(\mathbf{y} - \mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} \\ &\leq |\alpha| \lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|F(\mathbf{y}) - (F(\mathbf{x}) + DF(\mathbf{x})(\mathbf{y} - \mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} + |\beta| \lim_{\|\mathbf{y}-\mathbf{x}\|_V \to 0} \frac{\|G(\mathbf{y}) - (G(\mathbf{x}) + DG(\mathbf{x})(\mathbf{y} - \mathbf{x}))\|_W}{\|\mathbf{y} - \mathbf{x}\|_V} \\ &\leq |\alpha| (0) + |\beta| (0) = 0. \end{aligned}$$

Therefore, $D(\alpha F + \beta G)(\mathbf{x}) = \alpha DF(\mathbf{x}) + \beta DG(\mathbf{x})$. □

The differentiation has an important property that is fundamental to many other properties.

Theorem 5.14. (**Chain Rule**) Let V_1, V_2, V_3 be Banach spaces. Let $F : V_1 \to V_2$ and $G : V_2 \to V_3$. Then, the differentiation of $G \circ F : V_1 \to V_3$ at $\mathbf{x} \in V_1$ is given by:

$$D(G \circ F)(\mathbf{x}) = DG(F(\mathbf{x})) \circ DF(\mathbf{x}),$$

if F is differentiable at $\mathbf{x}$ and G is differentiable at $F(\mathbf{x})$.

Proof.
Since F is differentiable at $\mathbf{x}$ and G is differentiable at $F(\mathbf{x})$, we have:

$$\lim_{\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0} \frac{\|F(\mathbf{y}) - (F(\mathbf{x}) + DF(\mathbf{x})(\mathbf{y}-\mathbf{x}))\|_{V_2}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} = 0, \tag{1}$$

$$\lim_{\|F(\mathbf{y})-F(\mathbf{x})\|_{V_2} \to 0} \frac{\|G(F(\mathbf{y})) - (G(F(\mathbf{x})) + DG(F(\mathbf{x}))(F(\mathbf{y})-F(\mathbf{x})))\|_{V_3}}{\|F(\mathbf{y})-F(\mathbf{x})\|_{V_2}} = 0. \tag{2}$$

By (1), there exists $\delta > 0$ such that if $\|\mathbf{y}-\mathbf{x}\|_{V_1} < \delta$,

$$\|F(\mathbf{y}) - (F(\mathbf{x}) + DF(\mathbf{x})(\mathbf{y}-\mathbf{x}))\|_{V_2} < \|\mathbf{y}-\mathbf{x}\|_{V_1}.$$

Therefore, since $DF(\mathbf{x})$ is bounded, $\|DF(\mathbf{x})\| = M < \infty$,

$$\begin{aligned}
\|F(\mathbf{y}) - F(\mathbf{x})\|_{V_2} &= \|F(\mathbf{y}) - F(\mathbf{x}) - DF(\mathbf{x})(\mathbf{y}-\mathbf{x}) + DF(\mathbf{x})(\mathbf{y}-\mathbf{x})\|_{V_2} \\
&\le \|F(\mathbf{y}) - (F(\mathbf{x}) + DF(\mathbf{x})(\mathbf{y}-\mathbf{x}))\|_{V_2} + \|DF(\mathbf{x})(\mathbf{y}-\mathbf{x})\|_{V_2} \\
&\le \|\mathbf{y}-\mathbf{x}\|_{V_1} + M\|\mathbf{y}-\mathbf{x}\|_{V_1}, \\
\frac{\|F(\mathbf{y}) - F(\mathbf{x})\|_{V_2}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} &\le 1 + M.
\end{aligned}$$

From this, we also have that $\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0$ implies that $\|F(\mathbf{y}) - F(\mathbf{x})\|_{V_2} \to 0$. Therefore, since $DG(F(\mathbf{x}))$ is bounded,

$$\begin{aligned}
0 &\le \lim_{\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0} \frac{\|G(F(\mathbf{y})) - (G(F(\mathbf{x})) + DG(F(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y}-\mathbf{x})))\|_{V_3}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} \\
&\le \lim_{\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0} \frac{\|G(F(\mathbf{y})) - (G(F(\mathbf{x})) + DG(F(\mathbf{x}))(F(\mathbf{y})-F(\mathbf{x})))\|_{V_3}}{\|F(\mathbf{y})-F(\mathbf{x})\|_{V_2}} \left(\frac{\|F(\mathbf{y}) - F(\mathbf{x})\|_{V_2}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} \right) \\
&\quad + \lim_{\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0} \frac{\|DG(F(\mathbf{x}))(F(\mathbf{y})-F(\mathbf{x})) - DG(F(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y}-\mathbf{x}))\|_{V_3}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} \\
&\le (1+M) \lim_{\|F(\mathbf{y})-F(\mathbf{x})\|_{V_2} \to 0} \frac{\|G(F(\mathbf{y})) - (G(F(\mathbf{x})) + DG(F(\mathbf{x}))(F(\mathbf{y})-F(\mathbf{x})))\|_{V_3}}{\|F(\mathbf{y})-F(\mathbf{x})\|_{V_2}} \\
&\quad + \|DG(F(\mathbf{x}))\| \lim_{\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0} \frac{\|F(\mathbf{y}) - (F(\mathbf{x}) + DF(\mathbf{x})(\mathbf{y}-\mathbf{x}))\|_{V_2}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} = 0.
\end{aligned}$$

Therefore, by the Squeeze Theorem,

$$\lim_{\|\mathbf{y}-\mathbf{x}\|_{V_1} \to 0} \frac{\|G(F(\mathbf{y})) - (G(F(\mathbf{x})) + DG(F(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y}-\mathbf{x})))\|_{V_3}}{\|\mathbf{y}-\mathbf{x}\|_{V_1}} = 0,$$

and thus $D(G \circ F)(\mathbf{x}) = DG(F(\mathbf{x})) \circ DF(\mathbf{x})$. □

We can use the chain rule to derive many properties of differentiation.

Theorem 5.15. (**Theorem 4.14**) Let H be a Hilbert space. Let $F : H \to \mathbb{R}$ and $G : \mathbb{R} \to \mathbb{R}$. Then, the gradient of $G \circ F : H \to \mathbb{R}$ at $\mathbf{x} \in H$ is given by:

$$\nabla(G \circ F)(\mathbf{x}) = G'(F(\mathbf{x})) \cdot \nabla F(\mathbf{x}),$$

if F is differentiable at $\mathbf{x}$ and G is differentiable at $F(\mathbf{x})$.

Alternative Proof.
For $\mathbf{y} \in H$ and $\alpha \in \mathbb{R}$,

$$DG(F(\mathbf{x}))(\alpha) = G'(F(\mathbf{x}))\alpha, \qquad DF(\mathbf{x})(\mathbf{y}) = \langle \nabla F(\mathbf{x}), \mathbf{y} \rangle .$$

Therefore, by the chain rule,

$$\begin{aligned} D(G \circ F)(\mathbf{x})(\mathbf{y}) &= DG(F(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y})) \\ &= G'(F(\mathbf{x})) \langle \nabla F(\mathbf{x}), \mathbf{y} \rangle \\ &= \langle G'(F(\mathbf{x})) \cdot \nabla F(\mathbf{x}), \mathbf{y} \rangle . \end{aligned}$$

Therefore, $\nabla(G \circ F)(\mathbf{x}) = G'(F(\mathbf{x})) \cdot \nabla F(\mathbf{x})$. □

Theorem 5.16. Let H_1, H_2 be Hilbert spaces. Let $F : H_1 \to H_2$ and $G : H_2 \to \mathbb{R}$. Then, the gradient of $G \circ F : H_1 \to \mathbb{R}$ at $\mathbf{x} \in H_1$ is given by:

$$\nabla(G \circ F)(\mathbf{x}) = (DF(\mathbf{x}))^* \nabla G(F(\mathbf{x})),$$

if F is differentiable at $\mathbf{x}$ and G is differentiable at $F(\mathbf{x})$.

Proof.
For $\mathbf{y} \in H_1$, by the chain rule,

$$\begin{aligned} D(G \circ F)(\mathbf{x})(\mathbf{y}) &= DG(F(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y})) \\ &= \langle \nabla G(F(\mathbf{x})), DF(\mathbf{x})(\mathbf{y}) \rangle \\ &= \langle (DF(\mathbf{x}))^* \nabla G(F(\mathbf{x})), \mathbf{y} \rangle . \end{aligned}$$

Therefore, $\nabla(G \circ F)(\mathbf{x}) = (DF(\mathbf{x}))^* \nabla G(F(\mathbf{x}))$. □

Remark 5.16.1. If G is linear and bounded, then there exists $\mathbf{a} \in H_2$ such that:

$$G(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle , \qquad \nabla G(\mathbf{x}) = \mathbf{a}, \qquad \text{for } \mathbf{x} \in H_2.$$

Therefore, $\nabla(G \circ F)(\mathbf{x}) = (DF(\mathbf{x}))^* \mathbf{a}$.

Remark 5.16.2. If F is linear and bounded ($F \in \mathscr{L}(H_1, H_2)$), then:

$$DF(\mathbf{x}) = F, \qquad \text{for } \mathbf{x} \in H_1.$$

Therefore, $\nabla(G \circ F)(\mathbf{x}) = F^* \nabla G(F(\mathbf{x}))$.

We can also derive the product rule for differentiation.

Theorem 5.17. Let H be a Hilbert space, and let $f_1, f_2 : H \to \mathbb{R}$. If f_1 and f_2 are differentiable at $\mathbf{x} \in H$, then:

$$\nabla(f_1 f_2)(\mathbf{x}) = f_2(\mathbf{x}) \cdot \nabla f_1(\mathbf{x}) + f_1(\mathbf{x}) \cdot \nabla f_2(\mathbf{x}).$$

Proof.
Define $F : H \to \mathbb{R}^2$ and $G : \mathbb{R}^2 \to \mathbb{R}$ by:

$$F(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{pmatrix}, \qquad \text{for } \mathbf{x} \in H, \qquad\qquad G\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right) = \alpha\beta, \qquad \text{for } \alpha, \beta \in \mathbb{R}.$$

Then, $f(\mathbf{x}) = G(F(\mathbf{x})) = (G \circ F)(\mathbf{x})$. For DG, for any $\alpha, \beta \in \mathbb{R}$,

$$DG\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right)(\mathbf{y}) = \left\langle \nabla G\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right), \mathbf{y} \right\rangle = \left\langle \begin{pmatrix} \beta \\ \alpha \end{pmatrix}, \mathbf{y} \right\rangle.$$

For DF, since f_1 and f_2 are Fréchet differentiable at $\mathbf{x} \in H$, for any $\mathbf{z} \in H$,

$$\begin{aligned}
f_1(\mathbf{z}) &= f_1(\mathbf{x}) + \langle \nabla f_1(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + o(\|\mathbf{z} - \mathbf{x}\|) = f_1(\mathbf{x}) + \langle \nabla f_1(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + \varepsilon_1, \\
f_2(\mathbf{z}) &= f_2(\mathbf{x}) + \langle \nabla f_2(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + o(\|\mathbf{z} - \mathbf{x}\|) = f_2(\mathbf{x}) + \langle \nabla f_2(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + \varepsilon_2, \\
F(\mathbf{z}) &= F(\mathbf{x}) + \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle \\ \langle \nabla f_2(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = F(\mathbf{x}) + \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle \\ \langle \nabla f_2(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle \end{pmatrix} + \boldsymbol{\varepsilon}.
\end{aligned}$$

Since $\|\boldsymbol{\varepsilon}\|_2 \le \|\boldsymbol{\varepsilon}\|_1 = |\varepsilon_1| + |\varepsilon_2| \sim o(\|\mathbf{z} - \mathbf{x}\|)$, for $\mathbf{y} \in H$,

$$DF(\mathbf{x})(\mathbf{y}) = \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{y} \rangle \\ \langle \nabla f_2(\mathbf{x}), \mathbf{y} \rangle \end{pmatrix}.$$

By the chain rule, for $\mathbf{y} \in H$,

$$\begin{aligned}
D(f_1 f_2)(\mathbf{x})(\mathbf{y}) = DG(F(\mathbf{x})) \circ DF(\mathbf{x}) &= \left\langle \begin{pmatrix} f_2(\mathbf{x}) \\ f_1(\mathbf{x}) \end{pmatrix}, \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{y} \rangle \\ \langle \nabla f_2(\mathbf{x}), \mathbf{y} \rangle \end{pmatrix} \right\rangle = \langle f_2(\mathbf{x}) \cdot \nabla f_1(\mathbf{x}) + f_1(\mathbf{x}) \cdot \nabla f_2(\mathbf{x}), \mathbf{y} \rangle, \\
\nabla(f_1 f_2)(\mathbf{x}) &= f_2(\mathbf{x}) \cdot \nabla f_1(\mathbf{x}) + f_1(\mathbf{x}) \cdot \nabla f_2(\mathbf{x}).
\end{aligned}$$

□

Theorem 5.18. Let $F : \mathbb{R}^n \to \mathbb{R}^m$, where $\mathbb{R}^n$ and $\mathbb{R}^m$ are equipped with the standard inner product. Then:

$$DF(\mathbf{x}) = \left(\frac{\partial (F(\mathbf{x}))_i}{\partial x_j} \right)_{i=1,j=1}^{m,n} \in \mathbb{R}^{m \times n}.$$

Proof.
Denote $(F(\mathbf{x}))_i = f_i(\mathbf{x}) \in \mathbb{R}$. By differentiability,

$$f_i(\mathbf{y}) = f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \varepsilon_i, \qquad \text{for } \mathbf{y} \in \mathbb{R}^n,$$

where $|\varepsilon_i| = o(\|\mathbf{y} - \mathbf{x}\|_2)$. Therefore,

$$F(\mathbf{y}) = F(\mathbf{x}) + \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ \vdots \\ \langle \nabla f_m(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix} = F(\mathbf{x}) + \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ \vdots \\ \langle \nabla f_m(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \end{pmatrix} + \boldsymbol{\varepsilon}.$$

Since $\|\boldsymbol{\varepsilon}\|_2 \le \|\boldsymbol{\varepsilon}\|_1 \sim o(\|\mathbf{y} - \mathbf{x}\|_2)$,

$$DF(\mathbf{x})(\mathbf{y}) = \begin{pmatrix} \langle \nabla f_1(\mathbf{x}), \mathbf{y} \rangle \\ \vdots \\ \langle \nabla f_m(\mathbf{x}), \mathbf{y} \rangle \end{pmatrix} = \begin{pmatrix} (\nabla f_1(\mathbf{x}))^T \mathbf{y} \\ \vdots \\ (\nabla f_m(\mathbf{x}))^T \mathbf{y} \end{pmatrix} = \begin{pmatrix} (\nabla f_1(\mathbf{x}))^T \\ \vdots \\ (\nabla f_m(\mathbf{x}))^T \end{pmatrix} \mathbf{y}, \qquad DF(\mathbf{x}) = \begin{pmatrix} (\nabla f_1(\mathbf{x}))^T \\ \vdots \\ (\nabla f_m(\mathbf{x}))^T \end{pmatrix}.$$

□

Remark 5.18.1. Differentiation is an extension of the Jacobian matrix.

Can we extend the multiplication property to Banach spaces?

Theorem 5.19. (**Number Multiplication Rule**) Let V be a Banach space, and let $f_1, f_2 : V \to \mathbb{R}$. If f_1 and f_2 are differentiable at $\mathbf{x} \in V$, then:

$$D(f_1 f_2)(\mathbf{x})(\mathbf{y}) = f_2(\mathbf{x}) \cdot Df_1(\mathbf{x})(\mathbf{y}) + f_1(\mathbf{x}) \cdot Df_2(\mathbf{x})(\mathbf{y}), \qquad \text{for } \mathbf{y} \in V.$$

Proof.
Since f_1 and f_2 are differentiable at $\mathbf{x}$, for any $\mathbf{h} \in V$,

$$\begin{aligned} f_1(\mathbf{x}+\mathbf{h}) &= f_1(\mathbf{x}) + Df_1(\mathbf{x})(\mathbf{h}) + \varepsilon_1, \\ f_2(\mathbf{x}+\mathbf{h}) &= f_2(\mathbf{x}) + Df_2(\mathbf{x})(\mathbf{h}) + \varepsilon_2, \end{aligned}$$

where $|\varepsilon_i| = o(\|\mathbf{h}\|)$. Expanding the product,

$$\begin{aligned} (f_1 f_2)(\mathbf{x}+\mathbf{h}) = (f_1 f_2)(\mathbf{x}) + (f_1(\mathbf{x}) \cdot Df_2(\mathbf{x}) + f_2(\mathbf{x}) \cdot Df_1(\mathbf{x}))(\mathbf{h}) + (Df_1(\mathbf{x}) \cdot Df_2(\mathbf{x}))(\mathbf{h}) \\ + \varepsilon_1(f_2(\mathbf{x}) + Df_2(\mathbf{x})(\mathbf{h})) + \varepsilon_2(f_1(\mathbf{x}) + Df_1(\mathbf{x})(\mathbf{h})) + \varepsilon_1\varepsilon_2. \end{aligned}$$

Let $R(\mathbf{h}) = (Df_1(\mathbf{x})Df_2(\mathbf{x}))(\mathbf{h}) + \varepsilon_1(f_2(\mathbf{x}) + Df_2(\mathbf{x})(\mathbf{h})) + \varepsilon_2(f_1(\mathbf{x}) + Df_1(\mathbf{x})(\mathbf{h})) + \varepsilon_1\varepsilon_2$. As $\|\mathbf{h}\| \to 0$,

$$\begin{aligned} \frac{|(Df_1(\mathbf{x})Df_2(\mathbf{x}))(\mathbf{h})|}{\|\mathbf{h}\|} &\le \frac{\|Df_1(\mathbf{x})\| \, \|Df_2(\mathbf{x})\| \, \|\mathbf{h}\|^2}{\|\mathbf{h}\|} = \|Df_1(\mathbf{x})\| \, \|Df_2(\mathbf{x})\| \, \|\mathbf{h}\| \to 0, \\ \frac{|\varepsilon_1(f_2(\mathbf{x}) + Df_2(\mathbf{x})(\mathbf{h}))|}{\|\mathbf{h}\|} &\le \frac{|\varepsilon_1|}{\|\mathbf{h}\|}(|f_2(\mathbf{x})| + \|Df_2(\mathbf{x})\| \, \|\mathbf{h}\|) \to 0, && (|\varepsilon_1| = o(\|\mathbf{h}\|)) \\ \frac{|\varepsilon_2(f_1(\mathbf{x}) + Df_1(\mathbf{x})(\mathbf{h}))|}{\|\mathbf{h}\|} &\le \frac{|\varepsilon_2|}{\|\mathbf{h}\|}(|f_1(\mathbf{x})| + \|Df_1(\mathbf{x})\| \, \|\mathbf{h}\|) \to 0, && (|\varepsilon_2| = o(\|\mathbf{h}\|)) \\ |\varepsilon_1\varepsilon_2| &= o(\|\mathbf{h}\|). && (|\varepsilon_1\varepsilon_2| = o(\|\mathbf{h}\|^2)) \end{aligned}$$

Therefore, $|R(\mathbf{h})| = o(\|\mathbf{h}\|)$. We have:

$$(f_1 f_2)(\mathbf{x}+\mathbf{h}) = (f_1 f_2)(\mathbf{x}) + (f_1(\mathbf{x}) \cdot Df_2(\mathbf{x}) + f_2(\mathbf{x}) \cdot Df_1(\mathbf{x}))(\mathbf{h}) + o(\|\mathbf{h}\|).$$

By definition, $D(f_1 f_2)(\mathbf{x})(\mathbf{y}) = f_2(\mathbf{x}) \cdot Df_1(\mathbf{x})(\mathbf{y}) + f_1(\mathbf{x}) \cdot Df_2(\mathbf{x})(\mathbf{y})$. □

Theorem 5.20. (**Scalar Multiplication Rule**) Let V be a Banach space. Let $f : V \to \mathbb{R}$ and $F : V \to V$. If f and F are differentiable at $\mathbf{x} \in V$, then:

$$D(fF)(\mathbf{x})(\mathbf{y}) = f(\mathbf{x}) \cdot DF(\mathbf{x})(\mathbf{y}) + Df(\mathbf{x})(\mathbf{y})F(\mathbf{x}), \qquad \text{for } \mathbf{y} \in V.$$

Proof.
Define $G : V \to \mathbb{R} \times V$ and $\widetilde{F} : \mathbb{R} \times V \to V$ by:

$$\begin{aligned} G(\mathbf{x}) &= (f(\mathbf{x}), F(\mathbf{x})), && \text{for } \mathbf{x} \in V, \\ \widetilde{F}(\lambda, \mathbf{v}) &= \lambda \mathbf{v}, && \text{for } \lambda \in \mathbb{R}, \mathbf{v} \in V, \\ fF &= \widetilde{F} \circ G. \end{aligned}$$

We check the differentiability of $\widetilde{F}$. For $h \in \mathbb{R}$ and $\mathbf{k} \in V$, since $\widetilde{F}$ is bilinear,

$$\begin{aligned} \widetilde{F}(\lambda + h, \mathbf{v} + \mathbf{k}) &= \lambda \mathbf{v} + \lambda \mathbf{k} + h\mathbf{v} + h\mathbf{k} \\ &= \widetilde{F}(\lambda, \mathbf{v}) + \lambda \mathbf{k} + h\mathbf{v}. \end{aligned}$$

We have that $\|h\mathbf{v}\|_V = |h| \|\mathbf{v}\|_V$ and as $(h, \mathbf{v}) \to (0, \mathbf{0})$,

$$\begin{aligned} \frac{|h| \|\mathbf{v}\|_V}{\sqrt{|h|^2 + \|\mathbf{v}\|_V^2}} &\leq \frac{|h| \|\mathbf{v}\|_V}{|h|} = \|\mathbf{v}\|_V \to 0, \\ \frac{|h| \|\mathbf{v}\|_V}{\sqrt{|h|^2 + \|\mathbf{v}\|_V^2}} &\leq \frac{|h| \|\mathbf{v}\|_V}{\|\mathbf{v}\|_V} = |h| \to 0, \\ \implies \frac{|h| \|\mathbf{v}\|_V}{\sqrt{|h|^2 + \|\mathbf{v}\|_V^2}} &\leq \min(|h|, \|\mathbf{v}\|_V) \to 0. \end{aligned}$$

Thus, $\|h\mathbf{v}\|_V = o(\sqrt{h^2 + \|\mathbf{v}\|_V^2})$ and:

$$D\widetilde{F}(\lambda, \mathbf{v})(h, \mathbf{k}) = \lambda \mathbf{k} + h\mathbf{v}.$$

By the chain rule,

$$\begin{aligned} D(fF)(\mathbf{x})(\mathbf{y}) &= D\widetilde{F}(G(\mathbf{x}))(DG(\mathbf{x})(\mathbf{y})) \\ &= D\widetilde{F}(f(\mathbf{x}), F(\mathbf{x}))(Df(\mathbf{x})(\mathbf{y}), DF(\mathbf{x})(\mathbf{y})) \\ &= f(\mathbf{x}) \cdot DF(\mathbf{x})(\mathbf{y}) + Df(\mathbf{x})(\mathbf{y})F(\mathbf{x}). \end{aligned}$$

□

Remark 5.20.1. You can also prove the number multiplication rule with a similar proof.

Theorem 5.21. (**Matrix Multiplication Rule**) Let V be a Banach space. Let $F : V \to \mathbb{R}^{m\times n}$ and $G : V \to \mathbb{R}^{n\times p}$. If F and G are differentiable at $\mathbf{x} \in V$, then:

$$D(FG)(\mathbf{x})(\mathbf{y}) = DF(\mathbf{x})(\mathbf{y})G(\mathbf{x}) + F(\mathbf{x})DG(\mathbf{x})(\mathbf{y}), \qquad \text{for } \mathbf{y} \in V.$$

Proof.
Define $T : V \to \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p}$ and $M : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times p}$ by:

$$\begin{aligned} T(\mathbf{x}) &= (F(\mathbf{x}), G(\mathbf{x})), && \text{for } \mathbf{x} \in V, \\ M(\mathbf{A}, \mathbf{B}) &= \mathbf{AB}, && \text{for } \mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}, \\ FG &= M \circ T. \end{aligned}$$

We check the differentiability of M. For $\mathbf{H} \in \mathbb{R}^{m\times n}$ and $\mathbf{K} \in \mathbb{R}^{n\times p}$, since M is bilinear,

$$\begin{aligned} M(\mathbf{A} + \mathbf{H}, \mathbf{B} + \mathbf{K}) &= \mathbf{AB} + \mathbf{HB} + \mathbf{AK} + \mathbf{HK} \\ &= M(\mathbf{A}, \mathbf{B}) + \mathbf{HB} + \mathbf{AK} + \mathbf{HK}. \end{aligned}$$

We have that $\|\mathbf{HK}\| \le \|\mathbf{H}\| \|\mathbf{K}\|$ and as $(\mathbf{H}, \mathbf{K}) \to (\mathbf{O}, \mathbf{O})$,

$$\begin{aligned} \frac{\|\mathbf{H}\| \|\mathbf{K}\|}{\sqrt{\|\mathbf{H}\|^2 + \|\mathbf{K}\|^2}} &\le \frac{\|\mathbf{H}\| \|\mathbf{K}\|}{\|\mathbf{H}\|} = \|\mathbf{K}\| \to 0, \\ \frac{\|\mathbf{H}\| \|\mathbf{K}\|}{\sqrt{\|\mathbf{H}\|^2 + \|\mathbf{K}\|^2}} &\le \frac{\|\mathbf{H}\| \|\mathbf{K}\|}{\|\mathbf{K}\|} = \|\mathbf{H}\| \to 0, \\ \implies \frac{\|\mathbf{H}\| \|\mathbf{K}\|}{\sqrt{\|\mathbf{H}\|^2 + \|\mathbf{K}\|^2}} &\le \min(\|\mathbf{H}\|, \|\mathbf{K}\|) \to 0. \end{aligned}$$

Thus, $\|\mathbf{HK}\| = o(\sqrt{\|\mathbf{H}\|^2 + \|\mathbf{K}\|^2})$ and:

$$DM(\mathbf{A}, \mathbf{B})(\mathbf{H}, \mathbf{K}) = \mathbf{HB} + \mathbf{AK}.$$

By the chain rule,

$$\begin{aligned} D(FG)(\mathbf{x})(\mathbf{y}) &= DM(T(\mathbf{x}))(DT(\mathbf{x})(\mathbf{y})) \\ &= DM(F(\mathbf{x}), G(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y}), DG(\mathbf{x})(\mathbf{y})) \\ &= DF(\mathbf{x})(\mathbf{y})G(\mathbf{x}) + F(\mathbf{x})DG(\mathbf{x})(\mathbf{y}). \end{aligned}$$

□

Remark 5.21.1. The matrix multiplication rule can also be applied to $\mathbb{R}$ and $\mathbb{R}^n$.

Theorem 5.22. Let H_1, H_2 be Hilbert spaces. Let $F, G : H_1 \to H_2$ and define:

$$\langle F, G\rangle(\mathbf{x}) = \langle F(\mathbf{x}), G(\mathbf{x})\rangle_{H_2}, \qquad \text{for } \mathbf{x} \in H_1.$$

If F and G are differentiable at $\mathbf{x} \in H_1$, then:

$$\nabla \langle F, G\rangle(\mathbf{x})(\mathbf{y}) = (DF(\mathbf{x}))^* G(\mathbf{x}) + (DG(\mathbf{x}))^* F(\mathbf{x}).$$

Proof.
Define $T : H_1 \to H_2 \times H_2$ and $M : H_2 \times H_2 \to \mathbb{R}$ by:

$$\begin{aligned} T(\mathbf{x}) &= (F(\mathbf{x}), G(\mathbf{x})), && \text{for } \mathbf{x} \in H_1, \\ M(\mathbf{u}, \mathbf{v}) &= \langle \mathbf{u}, \mathbf{v}\rangle_{H_2}, && \text{for } \mathbf{u}, \mathbf{v} \in H_2, \\ \langle F, G\rangle &= M \circ T. \end{aligned}$$

We check the differentiability of M. For $\mathbf{h}, \mathbf{k} \in H_2$, since M is bilinear,

$$M(\mathbf{u} + \mathbf{h}, \mathbf{v} + \mathbf{k}) = M(\mathbf{u}, \mathbf{v}) + \langle \mathbf{u}, \mathbf{k}\rangle_{H_2} + \langle \mathbf{h}, \mathbf{v}\rangle_{H_2} + \langle \mathbf{h}, \mathbf{k}\rangle_{H_2}.$$

By the Cauchy-Schwarz inequality, $|\langle \mathbf{h}, \mathbf{k}\rangle|_{H_2} \le \|\mathbf{h}\|_{H_2} \|\mathbf{k}\|_{H_2}$. As $(\mathbf{h}, \mathbf{k}) \to (\mathbf{0}, \mathbf{0})$,

$$\begin{aligned} \frac{\|\mathbf{h}\|_{H_2} \|\mathbf{k}\|_{H_2}}{\sqrt{\|\mathbf{h}\|_{H_2}^2 + \|\mathbf{k}\|_{H_2}^2}} &\le \frac{\|\mathbf{h}\|_{H_2} \|\mathbf{k}\|_{H_2}}{\|\mathbf{h}\|_{H_2}} = \|\mathbf{k}\|_{H_2} \to 0, \\ \frac{\|\mathbf{h}\|_{H_2} \|\mathbf{k}\|_{H_2}}{\sqrt{\|\mathbf{h}\|_{H_2}^2 + \|\mathbf{k}\|_{H_2}^2}} &\le \|\mathbf{h}\|_{H_2} \to 0, \\ \implies \frac{\|\mathbf{h}\|_{H_2} \|\mathbf{k}\|_{H_2}}{\sqrt{\|\mathbf{h}\|_{H_2}^2 + \|\mathbf{k}\|_{H_2}}} &\le \min(\|\mathbf{h}\|_{H_2}, \|\mathbf{k}\|_{H_2}) \to 0. \end{aligned}$$

As $(\mathbf{h}, \mathbf{k}) \to (\mathbf{0}, \mathbf{0})$, $\min(\|\mathbf{h}\|_{H_2}, \|\mathbf{k}\|_{H_2}) \to 0$. Thus, $|\langle \mathbf{h}, \mathbf{k}\rangle|_{H_2} = o(\sqrt{\|\mathbf{h}\|_{H_2}^2 + \|\mathbf{k}\|_{H_2}})$ and:

$$DM(\mathbf{u}, \mathbf{v})(\mathbf{h}, \mathbf{k}) = \langle \mathbf{u}, \mathbf{k}\rangle_{H_2} + \langle \mathbf{h}, \mathbf{v}\rangle_{H_2}.$$

By the chain rule,

$$\begin{aligned} D\langle F, G\rangle(\mathbf{x})(\mathbf{y}) &= DM(F(\mathbf{x}), G(\mathbf{x}))(DF(\mathbf{x})(\mathbf{y}), DG(\mathbf{x})(\mathbf{y})) \\ &= \langle F(\mathbf{x}), DG(\mathbf{x})(\mathbf{y})\rangle_{H_2} + \langle DF(\mathbf{x})(\mathbf{y}), G(\mathbf{x})\rangle_{H_2} \\ &= \langle (DG(\mathbf{x}))^* F(\mathbf{x}) + (DF(\mathbf{x}))^* G(\mathbf{x}), \mathbf{y}\rangle_{H_1}. \end{aligned}$$

Therefore, $\nabla \langle F, G\rangle(\mathbf{x}) = (DG(\mathbf{x}))^* F(\mathbf{x}) + (DF(\mathbf{x}))^* G(\mathbf{x})$. □

Remark 5.22.1. There are other bilinear mappings that satisfy a similar differentiation rule, e.g., convolution.

5.3 Hessian of functions

Similar to standard derivatives, what are the second-order derivatives of the Fréchet differentiation? Let H be a Hilbert space, and let $f : H \to \mathbb{R}$. We can consider ∇f as a mapping $\nabla f : H \to H$.

Definition 5.23. Let H be a Hilbert space, and $f : H \to \mathbb{R}$. The **Hessian** of f is defined as:

$$\nabla^2 f(\mathbf{x}) = D(\nabla f)(\mathbf{x}), \qquad \text{for } \mathbf{x} \in H.$$

Remark 5.23.1. $\nabla^2 f \in \mathscr{L}(H, H)$.

Remark 5.23.2. By Theorem 4.15 and 5.18, for $f : \mathbb{R}^n \to \mathbb{R}$,

$$\nabla^2 f(\mathbf{x}) = D(\nabla f)(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_i \, \partial x_j} \right)_{i=1,j=1}^{n,n} \in \mathbb{R}^{n \times n}.$$

Example 5.21. Let $\mathbf{A} \in \mathscr{L}(H_1, H_2)$, where H_1, H_2 are Hilbert spaces. Let $f : H_1 \to \mathbb{R}$ be defined as:

$$f(\mathbf{x}) = \frac{1}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|_{H_2}^2, \qquad \text{for } \mathbf{x} \in H_1,$$

where $\mathbf{b} \in H_2$. We have found that $\nabla f(\mathbf{x}) = \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{A}^*\mathbf{A}\mathbf{x} - \mathbf{A}^*\mathbf{b}$. Then,

$$\nabla^2 f(\mathbf{x}) = D(\nabla f)(\mathbf{x}) = \mathbf{A}^*\mathbf{A}.$$

Example 5.22. Let $\mathbf{A} \in \mathscr{L}(H_1, H_2)$, where H_1, H_2 are Hilbert spaces. Let $f : H_1 \to \mathbb{R}$ and $F : H_2 \to \mathbb{R}$ be defined as:

$$f(\mathbf{x}) = F(\mathbf{A}\mathbf{x}), \text{for } \mathbf{x} \in H_1.$$

By the chain rule,

$$\nabla f(\mathbf{x}) = \mathbf{A}^* \nabla F(\mathbf{A}\mathbf{x}).$$

We have that for $\mathbf{y} \in H_1$,

$$\begin{aligned}
D\mathbf{A}^*(\nabla F(\mathbf{A}\mathbf{x})) &= \mathbf{A}^*, \\
D(\nabla F)(\mathbf{A}\mathbf{x}) &= \nabla^2 F(\mathbf{A}\mathbf{x}), \\
D\mathbf{A}(\mathbf{x})(\mathbf{y}) &= \mathbf{A}\mathbf{y}, \\
\nabla^2 f(\mathbf{x})(\mathbf{y}) = D(\nabla f)(\mathbf{x})(\mathbf{y}) &= D(\mathbf{A}^* \nabla F(\mathbf{A}\mathbf{x}))(\mathbf{y}) \\
&= D\mathbf{A}^*(\nabla F(\mathbf{A}\mathbf{x}))(D(\nabla F)(\mathbf{A}\mathbf{x})(D\mathbf{A}(\mathbf{x})(\mathbf{y}))) \\
&= \mathbf{A}^* \nabla^2 F(\mathbf{A}\mathbf{x}) \mathbf{A}\mathbf{y}.
\end{aligned}$$

Example 5.23. Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined as:

$$f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\langle \mathbf{a}_i, \mathbf{x} \rangle),$$

where $\mathbf{a}_i \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^n$, and $f_i : \mathbb{R} \to \mathbb{R}$. How do we find $\nabla^2 f(\mathbf{x})$? Let:

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T \end{pmatrix}, \qquad F(\mathbf{y}) = \sum_{i=1}^{m} f_i(y_i), \qquad \text{for } \mathbf{y} \in \mathbb{R}^m.$$

Then we have $f(\mathbf{x}) = F(\mathbf{A}\mathbf{x})$. Therefore,

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^T \nabla F(\mathbf{A}\mathbf{x}) = \sum_{i=1}^{m} f_i'(\mathbf{a}_i^T \mathbf{x}) \mathbf{a}_i, \qquad \nabla^2 f(\mathbf{x}) = \mathbf{A}^T \nabla^2 F(\mathbf{A}\mathbf{x}) \mathbf{A} = \sum_{i=1}^{m} f_i''(\mathbf{a}_i^T \mathbf{x}) \mathbf{a}_i \mathbf{a}_i^T.$$

Example 5.24. Let H be a Hilbert space, and $f(\mathbf{x}) = \|\mathbf{x}\|$, where $\mathbf{x} \in H$. What is $\nabla^2 f(\mathbf{x})$? We know that when $\mathbf{x} \neq \mathbf{0}$,

$$\nabla f(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

Let $g : V \to \mathbb{R}$ and $I : V \to V$ be defined as:

$$g(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|}, \qquad I(\mathbf{x}) = \mathbf{x}, \qquad \nabla f = gI.$$

Let $g_1(t) = \frac{1}{\sqrt{t}}$ and $g_2(\mathbf{x}) = \|\mathbf{x}\|^2$. Then $g = g_1 \circ g_2$. We have:

$$g_1'(t) = -\frac{1}{2}t^{-\frac{3}{2}}, \qquad \nabla g_2(\mathbf{x}) = 2\mathbf{x}.$$

Therefore, by the chain rule, since g_1 is continuous for $t > 0$ and $\mathbf{x} \neq \mathbf{0}$,

$$\nabla g(\mathbf{x}) = g_1'(g_2(\mathbf{x})) \cdot \nabla g_2(\mathbf{x}) = -\frac{\mathbf{x}}{\|\mathbf{x}\|^3}.$$

By the scalar multiplication rule,

$$\begin{aligned} \nabla^2 f(\mathbf{x})(\mathbf{y}) &= D(\nabla f)(\mathbf{x})(\mathbf{y}) \\ &= Dg(\mathbf{x})(\mathbf{y}) \cdot I(\mathbf{x}) + g(\mathbf{x}) \cdot DI(\mathbf{x})(\mathbf{y}) \\ &= \left\langle -\frac{1}{\|\mathbf{x}\|^3}\mathbf{x}, \mathbf{y} \right\rangle \mathbf{x} + \frac{\mathbf{y}}{\|\mathbf{x}\|}. \end{aligned}$$

If $H = \mathbb{R}^n$, then:

$$\nabla^2 f(\mathbf{x})(\mathbf{y}) = -\frac{\mathbf{x}\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|^3} + \frac{\mathbf{y}}{\|\mathbf{x}\|}, \qquad \nabla^2 f(\mathbf{x}) = -\frac{1}{\|\mathbf{x}\|^3}\mathbf{x}\mathbf{x}^T + \frac{1}{\|\mathbf{x}\|}\mathbf{I}.$$

5.4 Function Expansion

Let $f : H \to \mathbb{R}$ be a differentiable function, where H is a Hilbert space. From the definition of the gradient at $\mathbf{x}$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|).$$

We aim to derive the expansion up to the second-order derivative. Consider:

$$g(t) = f(\mathbf{x} + t\mathbf{u}),$$

where $\mathbf{x}, \mathbf{u} \in H$ are given, and $t \in \mathbb{R}$. What is the directional derivative?

Theorem 5.24. (**Theorem 4.17**) Let H be a Hilbert space. Let $f : H \to \mathbb{R}$ and $\mathbf{x} \in H$. For any $\mathbf{u} \in H$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u}) = \langle \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \rangle ,$$

if f is differentiable at $\mathbf{x} + t\mathbf{u}$.

Alternative Proof.
Let $G : \mathbb{R} \to H$ be defined as:

$$G(t) = \mathbf{x} + t\mathbf{u}, \qquad f(\mathbf{x} + t\mathbf{u}) = f(G(t)).$$

Since $DG(t)(s) = s\mathbf{u}$ for $s \in \mathbb{R}$ and $Df(\mathbf{z})(\mathbf{y}) = \langle \nabla f(\mathbf{z}), \mathbf{y} \rangle$ for $\mathbf{y} \in H$, by the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u})(s) = Df(G(t))(DG(t)(s)) = \langle \nabla f(\mathbf{x} + t\mathbf{u}), s\mathbf{u} \rangle = s \langle \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \rangle .$$

Therefore, $\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x} + t\mathbf{u}) = \langle \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \rangle$. □

Corollary 5.25. Let H be a Hilbert space, and let $f : H \to \mathbb{R}$. Let f and ∇f be differentiable at $\mathbf{x} \in H$. For any $\mathbf{u} \in H$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathbf{x} + t\mathbf{u})\bigg|_{t=0} = \left\langle \nabla^2 f(\mathbf{x})(\mathbf{u}), \mathbf{u} \right\rangle .$$

Proof.
By Theorem 5.24,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathbf{x} + t\mathbf{u}) = \frac{\mathrm{d}}{\mathrm{d}t} \langle \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \rangle = \left\langle \frac{\mathrm{d}}{\mathrm{d}t} \nabla f(\mathbf{x} + t\mathbf{u}), \mathbf{u} \right\rangle = \left\langle \nabla^2 f(\mathbf{x} + t\mathbf{u})(\mathbf{u}), \mathbf{u} \right\rangle .$$

Setting $t = 0$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathbf{x} + t\mathbf{u})\bigg|_{t=0} = \left\langle \nabla^2 f(\mathbf{x})(\mathbf{u}), \mathbf{u} \right\rangle .$$

□

Theorem 5.26. Let H be a Hilbert space, and let $f : H \to \mathbb{R}$. Let f and ∇f be differentiable at $\mathbf{x} \in H$. For any $\mathbf{u}, \mathbf{v} \in H$,

$$\frac{\partial^2}{\partial s\, \partial t} f(\mathbf{x} + s\mathbf{u} + t\mathbf{v})\bigg|_{s=t=0} = \left\langle \nabla^2 f(\mathbf{x})(\mathbf{v}), \mathbf{u} \right\rangle = \left\langle \nabla^2 f(\mathbf{x})(\mathbf{u}), \mathbf{v} \right\rangle.$$

Proof.
By Theorem 5.24,

$$\frac{\partial^2}{\partial s\, \partial t} f(\mathbf{x} + s\mathbf{u} + t\mathbf{v}) = \frac{\partial}{\partial t} \left\langle \nabla f(\mathbf{x} + s\mathbf{u} + t\mathbf{v}), \mathbf{u} \right\rangle = \left\langle \frac{\partial}{\partial t} \nabla f(\mathbf{x} + s\mathbf{u} + t\mathbf{v}), \mathbf{u} \right\rangle = \left\langle \nabla^2 f(\mathbf{x} + s\mathbf{u} + t\mathbf{v})(\mathbf{v}), \mathbf{u} \right\rangle.$$

Similarly,

$$\frac{\partial^2}{\partial s\, \partial t} f(\mathbf{x} + s\mathbf{u} + t\mathbf{v}) = \frac{\partial}{\partial s} \left\langle \nabla f(\mathbf{x} + s\mathbf{u} + t\mathbf{v}), \mathbf{v} \right\rangle = \left\langle \frac{\partial}{\partial s} \nabla f(\mathbf{x} + s\mathbf{u} + t\mathbf{v}), \mathbf{v} \right\rangle = \left\langle \nabla^2 f(\mathbf{x} + s\mathbf{u} + t\mathbf{v})(\mathbf{u}), \mathbf{v} \right\rangle.$$

Setting $s = t = 0$,

$$\frac{\partial^2}{\partial s\, \partial t} f(\mathbf{x} + s\mathbf{u} + t\mathbf{v})\bigg|_{s=t=0} = \left\langle \nabla^2 f(\mathbf{x})(\mathbf{v}), \mathbf{u} \right\rangle = \left\langle \nabla^2 f(\mathbf{x})(\mathbf{u}), \mathbf{v} \right\rangle.$$

□

Remark 5.26.1. This shows that $(\nabla^2 f(\mathbf{x}))^* = \nabla^2 f(\mathbf{x})$ (self-adjoint).

We can present the second-order expansion of $f : H \to \mathbb{R}$.

Theorem 5.27. Let H be a Hilbert space, and let $f : H \to \mathbb{R}$. Assume that f and ∇f are differentiable at $\mathbf{x} \in H$. The **second-order Taylor expansion** of f is:

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \left\langle \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle + o(\|\mathbf{y} - \mathbf{x}\|^2), \qquad \text{for } \mathbf{y} \in H.$$

Proof.
Let $g(t) = f(\mathbf{x} + t\mathbf{u})$, where $\mathbf{x}, \mathbf{u} \in H$, $t \in \mathbb{R}$. By the Taylor expansion on $g(t)$,

$$g(t) = g(t_0) + g'(t_0)(t - t_0) + \frac{1}{2} g''(t_0)(t - t_0)^2 + o(|t - t_0|^2).$$

By Theorem 5.24 and Corollary 5.25,

$$f(\mathbf{x} + t\mathbf{u}) = f(\mathbf{x} + t_0\mathbf{u}) + \langle \nabla f(\mathbf{x} + t_0\mathbf{u}), \mathbf{u} \rangle (t - t_0) + \frac{1}{2} \left\langle \nabla^2 f(\mathbf{x} + t_0\mathbf{u})(\mathbf{u}), \mathbf{u} \right\rangle (t - t_0)^2 + o(|t - t_0|)^2.$$

For all $\mathbf{x} \in H$, we can choose:

$$\mathbf{u} = \frac{\mathbf{y} - \mathbf{x}}{\|\mathbf{y} - \mathbf{x}\|}, \qquad t = \|\mathbf{y} - \mathbf{x}\|, \qquad t_0 = 0, \qquad \mathbf{x} + t\mathbf{u} = \mathbf{x} + \|\mathbf{y} - \mathbf{x}\| \left(\frac{\mathbf{y} - \mathbf{x}}{\|\mathbf{y} - \mathbf{x}\|} \right) = \mathbf{y}.$$

Therefore, the second-order Taylor expansion is:

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \left\langle \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle + o(\|\mathbf{y} - \mathbf{x}\|^2).$$

□

5.5 Matrix Differentiation

It is straightforward to compute the derivatives of vectors in Euclidean space. However, in many machine learning tasks, matrix differentiation is also required for operations such as backpropagation. Assume we have a Hilbert space $\mathbb{R}^{m\times n}$ with the inner product defined as:

$$\langle \mathbf{X}, \mathbf{Y}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}y_{ij} = \operatorname{Tr}(\mathbf{X}^T\mathbf{Y}) = \operatorname{Tr}(\mathbf{Y}^T\mathbf{X}), \qquad \text{for } \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}.$$

The induced norm is defined as:

$$\|\mathbf{X}\|_F = \sqrt{\langle \mathbf{X}, \mathbf{X}\rangle} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2}.$$

Example 5.25. (**Linear matrix multiplications**) Let $\mathbf{A} \in \mathbb{R}^{p\times m}$ and $\mathbf{B} \in \mathbb{R}^{n\times q}$. Consider the mapping $F : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ defined as:

$$F(\mathbf{X}) = \mathbf{AXB}, \qquad \text{for } \mathbf{X} \in \mathbb{R}^{m\times n}.$$

Since F is linear, by Theorem 5.12, for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$,

$$DF(\mathbf{X})(\mathbf{Y}) = \mathbf{AYB}.$$

Example 5.26. (**Quadratic matrix multiplications**) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be given. Consider the mapping $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as:

$$F(\mathbf{X}) = \mathbf{XAX}, \qquad \text{for } \mathbf{X} \in \mathbb{R}^{n\times n}.$$

By the matrix multiplication rule, for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} DF(\mathbf{X})(\mathbf{Y}) &= D(\mathbf{X})(\mathbf{Y})\mathbf{AX} + \mathbf{X}D(\mathbf{AX})(\mathbf{Y}) \\ &= \mathbf{YAX} + \mathbf{XAY}. \end{aligned}$$

Example 5.27. (**Transpose**) Consider the mapping $F : \mathbb{R}^{m\times n} \to \mathbb{R}^{n\times m}$ defined as:

$$F(\mathbf{X}) = \mathbf{X}^T.$$

Since F is linear, by Theorem 5.12, for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$,

$$DF(\mathbf{X})(\mathbf{Y}) = \mathbf{Y}^T.$$

Example 5.28. (**Inversion**) Consider the mapping $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as:

$$F(\mathbf{X}) = \mathbf{X}^{-1}, \qquad \text{for } \mathbf{X} \in \mathbb{R}^{n\times n} \text{ and invertible.}$$

Note that $\mathbf{XX}^{-1} = \mathbf{I}$. By the matrix multiplication rule, for invertible $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \mathbf{O} = D(\mathbf{I})(\mathbf{Y}) &= D(\mathbf{X})(\mathbf{Y})\mathbf{X}^{-1} + \mathbf{X}D(\mathbf{X}^{-1})(\mathbf{Y}) \\ \mathbf{O} &= \mathbf{YX}^{-1} + \mathbf{X}D(\mathbf{X}^{-1})(\mathbf{Y}) \\ -\mathbf{X}^{-1}\mathbf{YX}^{-1} &= D(\mathbf{X}^{-1})(\mathbf{Y}). \end{aligned}$$

Therefore, $DF(\mathbf{X})(\mathbf{Y}) = -\mathbf{X}^{-1}\mathbf{YX}^{-1}$ for invertible $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$.

Example 5.29. (**Trace**) Consider the mapping $\operatorname{Tr} : \mathbb{R}^{n\times n} \to \mathbb{R}$ defined as:

$$\operatorname{Tr}(\mathbf{X}) = \sum_{i=1}^{n} \lambda_i(\mathbf{X}), \qquad \text{for } \mathbf{X} \in \mathbb{R}^{n\times n},$$

where $\lambda_i(\mathbf{X})$ is the i-th eigenvalue of $\mathbf{X}$. Since $\operatorname{Tr}(\mathbf{X}) = \langle \mathbf{X}, \mathbf{I}\rangle$, it is linear. By Theorem 5.12, for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$,

$$D\operatorname{Tr}(\mathbf{X})(\mathbf{Y}) = \langle \mathbf{Y}, \mathbf{I}\rangle.$$

Thus, $\nabla \operatorname{Tr}(\mathbf{X}) = \mathbf{I}$.

Theorem 5.28. For any $\mathbf{Y} \in \mathbb{R}^{n\times n}$,
$$DF(\mathbf{I})(\mathbf{Y}) = \operatorname{Tr}(\mathbf{Y}).$$

Proof.
We may find that:
$$\det(\mathbf{X}) = \prod_{i=1}^{n} \lambda_i(\mathbf{X}), \qquad \text{for } \mathbf{X} \in \mathbb{R}^{n\times n},$$
where $\lambda_i(\mathbf{X})$ is the i-th eigenvalue of $\mathbf{X}$. For $\mathbf{Y} \in \mathbb{R}^{n\times n}$,
$$\begin{aligned}
\det(\mathbf{I}+\mathbf{Y}) &= \prod_{i=1}^{n}(1+\lambda_i(\mathbf{Y})) \\
&= \prod_{i=1}^{n} 1 + \sum_{i=1}^{n}\lambda_i(\mathbf{Y}) + \sum_{i<j}\lambda_i(\mathbf{Y})\lambda_j(\mathbf{Y}) + \sum_{i<j<k}\lambda_i(\mathbf{Y})\lambda_j(\mathbf{Y})\lambda_k(\mathbf{Y}) + \cdots \\
&= \det(\mathbf{I}) + \operatorname{Tr}(\mathbf{Y}) + \sum_{i<j}\lambda_i(\mathbf{Y})\lambda_j(\mathbf{Y}) + \sum_{i<j<k}\lambda_i(\mathbf{Y})\lambda_j(\mathbf{Y})\lambda_k(\mathbf{Y}) + \cdots .
\end{aligned}$$

Therefore, by using the fact that $\max_i |\lambda_i(\mathbf{Y})| \le \|\mathbf{Y}\|_2 \le \|\mathbf{Y}\|_F$,
$$\begin{aligned}
0 &\le \lim_{\|\mathbf{Y}\|_F\to 0} \frac{|\det(\mathbf{I}+\mathbf{Y}) - (\det(\mathbf{I}) + \operatorname{Tr}(\mathbf{Y}))|}{\|\mathbf{Y}\|_F} \\
&\le \lim_{\|\mathbf{Y}\|_F\to 0} \frac{\sum_{i<j}|\lambda_i(\mathbf{Y})\lambda_j(\mathbf{Y})| + \sum_{i<j<k}|\lambda_i(\mathbf{Y})\lambda_j(\mathbf{Y})\lambda_k(\mathbf{Y})| + \cdots}{\|\mathbf{Y}\|_F} \\
&\le \lim_{\|\mathbf{Y}\|_F\to 0} \frac{\sum_{i<j}\|\mathbf{Y}\|_F^2 + \sum_{i<j<k}\|\mathbf{Y}\|_F^3 + \cdots}{\|\mathbf{Y}\|_F} = 0.
\end{aligned}$$

Therefore, by the Squeeze Theorem, $D\det(\mathbf{I})(\mathbf{Y}) = \operatorname{Tr}(\mathbf{Y})$. □

Example 5.30. (**Determinant**) Consider the mapping $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ be defined as:
$$\det(\mathbf{X}) = \prod_{i=1}^{n} \lambda_i(\mathbf{X}), \qquad \text{for } \mathbf{X} \in \mathbb{R}^{n\times n},$$
where $\lambda_i(\mathbf{X})$ is the i-th eigenvalue of $\mathbf{X}$. For any invertible $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n\times n}$,
$$\begin{aligned}
D\det(\mathbf{X})(\mathbf{Z}) &= \frac{\mathrm{d}}{\mathrm{d}t}\det(\mathbf{X}+t\mathbf{Z})\Big|_{t=0} && \text{(Theorem 5.24)} \\
&= \det(\mathbf{X})\,\frac{\mathrm{d}}{\mathrm{d}t}\det(\mathbf{I}+t\mathbf{X}^{-1}\mathbf{Z})\Big|_{t=0} \\
&= \det(\mathbf{X}) D\det(\mathbf{I})(\mathbf{X}^{-1}\mathbf{Z}) \\
&= \det(\mathbf{X})\operatorname{Tr}(\mathbf{X}^{-1}\mathbf{Z}) && \text{(Theorem 5.28)} \\
&= \left\langle \det(\mathbf{X})(\mathbf{X}^{-1})^T, \mathbf{Z}\right\rangle .
\end{aligned}$$
Therefore, $D\det(\mathbf{X})(\mathbf{Y}) = \det(\mathbf{X})\operatorname{Tr}(\mathbf{X}^{-1}\mathbf{Y})$ and $\nabla\det(\mathbf{X}) = \det(\mathbf{X})(\mathbf{X}^{-1})^T$.

Theorem 5.29. (**Jacobi's Formula**) Let $f : \mathbb{R} \to \mathbb{R}$ be defined as:
$$f(t) = \det(\mathbf{A}(t)), \qquad \text{for } t \in \mathbb{R},$$
where $\mathbf{A} : \mathbb{R} \to \mathbb{R}^{n\times n}$ is differentiable and invertible. We have:
$$\frac{\mathrm{d}}{\mathrm{d}t}f(t) = \det(\mathbf{A}(t))\operatorname{Tr}\left(\mathbf{A}^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{A}(t)\right)\right).$$

Proof.
By the chain rule,
$$\frac{\mathrm{d}}{\mathrm{d}t}\det(\mathbf{A}(t)) = D\det(\mathbf{A}(t))\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{A}(t)\right) = \det(\mathbf{A}(t))\operatorname{Tr}\left(\mathbf{A}^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{A}(t)\right)\right).$$

□

Example 5.31. Let $F : \mathbb{R}^{n\times n} \to \mathbb{R}$ be defined by:

$$F(\mathbf{X}) = \ln(\det(\mathbf{X})), \qquad \text{for symmetric positive definite } \mathbf{X} \in \mathbb{R}^{n\times n}.$$

By the chain rule, for any symmetric positive definite $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} DF(\mathbf{X})(\mathbf{Y}) &= D\ln(\det(\mathbf{X}))(D\det(\mathbf{X})(\mathbf{Y})) \\ &= \frac{\det(\mathbf{X})\operatorname{Tr}(\mathbf{X}^{-1}\mathbf{Y})}{\det(\mathbf{X})} \\ &= \operatorname{Tr}(\mathbf{X}^{-1}\mathbf{Y}) = \left\langle (\mathbf{X}^{-1})^T, \mathbf{Y} \right\rangle. \end{aligned} \qquad \text{(Example 5.30)}$$

Therefore, $\nabla F(\mathbf{X}) = (\mathbf{X}^{-1})^T$.

Example 5.32. (**Eigenvalues and eigenvectors**) Let $\lambda_i : \mathbb{R}^{n\times n} \to \mathbb{R}$ and $\mathbf{v}_i : \mathbb{R}^{n\times n} \to \mathbb{R}$ be defined by:

$$\mathbf{X}\mathbf{v}_i[\mathbf{X}] = (\lambda_i \mathbf{v}_i)[\mathbf{X}], \qquad \text{for symmetric and real } \mathbf{X} \in \mathbb{R}^{n\times n},$$

where $\lambda_i[\mathbf{X}]$ is the i-th eigenvalue of $\mathbf{X}$ with the corresponding unit eigenvector $\mathbf{v}_i[\mathbf{X}]$. By the matrix multiplication rule, for any symmetric and real $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} D(\mathbf{X}\mathbf{v}_i[\mathbf{X}])(\mathbf{Y}) &= D(\lambda_i\mathbf{v}_i)[\mathbf{X}](\mathbf{Y}) \\ D(\mathbf{X})(\mathbf{Y})\mathbf{v}_i[\mathbf{X}] + \mathbf{X}D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= D\lambda_i[\mathbf{X}](\mathbf{Y})\mathbf{v}_i[\mathbf{X}] + \lambda_i[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) \\ \mathbf{Y}\mathbf{v}_i[\mathbf{X}] + \mathbf{X}D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= D\lambda_i[\mathbf{X}](\mathbf{Y})\mathbf{v}_i[\mathbf{X}] + \lambda_i[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}). \end{aligned} \qquad (1)$$

Since $\mathbf{v}_i[\mathbf{X}]$ is a unit vector, by the matrix multiplication rule,

$$(\mathbf{v}_i^T\mathbf{v}_i)[\mathbf{X}] = 1, \qquad (D\mathbf{v}_i[\mathbf{X}])^T\mathbf{v}_i[\mathbf{X}] + \mathbf{v}_i^T[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}] = 2\mathbf{v}_i^T[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}] = 0.$$

Therefore, $\mathbf{v}_i^T[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}] = 0$. Multiplying $\mathbf{v}_i^T[\mathbf{X}]$ to (1),

$$\begin{aligned} \mathbf{v}_i^T[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}] + \mathbf{v}_i^T[\mathbf{X}]\mathbf{X}D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= \mathbf{v}_i^T[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}] + (\lambda_i\mathbf{v}_i^T)[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) \qquad (\mathbf{X} \text{ is symmetric}) \\ &= \mathbf{v}_i^T[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}], \\ D\lambda_i[\mathbf{X}](\mathbf{Y})(\mathbf{v}_i^T\mathbf{v}_i)[\mathbf{X}] + (\lambda_i\mathbf{v}_i^T)[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= D\lambda_i[\mathbf{X}](\mathbf{Y}). \end{aligned}$$

Thus, we can find that:

$$D\lambda_i[\mathbf{X}](\mathbf{Y}) = \mathbf{v}_i^T[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}] = \left\langle (\mathbf{v}_i\mathbf{v}_i^T)[\mathbf{X}], \mathbf{Y} \right\rangle, \qquad \nabla\lambda_i[\mathbf{X}] = (\mathbf{v}_i\mathbf{v}_i^T)[\mathbf{X}].$$

Moreover, substituting $D\lambda_i[\mathbf{X}](\mathbf{Y})$ into (1),

$$\begin{aligned} \mathbf{Y}\mathbf{v}_i[\mathbf{X}] + \mathbf{X}D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= (\mathbf{v}_i\mathbf{v}_i^T)[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}] + \lambda_i[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) \\ (\mathbf{X} - \lambda_i[\mathbf{X}]\mathbf{I})D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= ((\mathbf{v}_i\mathbf{v}_i^T)[\mathbf{X}] - \mathbf{I})\mathbf{Y}\mathbf{v}_i[\mathbf{X}]. \end{aligned} \qquad (2)$$

The matrix $\mathbf{X} - \lambda_i[\mathbf{X}]\mathbf{I}$ is not invertible. Consider when $\mathbf{X}$ is non-singular and $\lambda_i[\mathbf{X}]$ is a simple (unique) eigenvalue. We have:

$$\mathbf{X} - \lambda_i[\mathbf{X}]\mathbf{I} = \sum_{j=1}^{n}(\lambda_j\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}] - \sum_{j=1}^{n}(\lambda_i\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}] = \sum_{j\neq i}((\lambda_j - \lambda_i)\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}].$$

Note that $(\mathbf{v}_i\mathbf{v}_i^T)[\mathbf{X}] - \mathbf{I} = -\sum_{j\neq i}(\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}]$. If we consider each term of the summation by applying to (2),

$$((\lambda_j - \lambda_i)\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) = -(\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}].$$

Therefore,

$$\begin{aligned} D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) = \sum_{j=1}^{n}(\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) = \sum_{j\neq i}(\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}]D\mathbf{v}_i[\mathbf{X}](\mathbf{Y}) &= -\sum_{j\neq i}\frac{(\mathbf{v}_j\mathbf{v}_j^T)[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}]}{(\lambda_j - \lambda_i)[\mathbf{X}]} \\ &= \sum_{j\neq i}\frac{\mathbf{v}_j^T[\mathbf{X}]\mathbf{Y}\mathbf{v}_i[\mathbf{X}]}{(\lambda_i - \lambda_j)[\mathbf{X}]}\mathbf{v}_j. \end{aligned}$$

Read Appendix I (Newton's Method) and J (Deep Neural Network Training in Deep Learning) to see the case studies for Chapter 5.

Appendix A

Clustering, K-means, K-medians

This case study assumes that you have already read Chapter 2.
Suppose we are given N vectors in $\mathbb{R}^n$:

$$\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^n.$$

We want to group them into K different clusters.

Remark A.0.1. $\mathbb{R}^n$ is used for simplicity. In fact, it can be replaced by any vector space.

Before performing clustering on any vector space, we must formulate the problem mathematically.

1. Representation: Starting with N vectors $\{\mathbf{x}_i\}_{i=1}^N$ and K clusters $\{G_j\}_{j=1}^K$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the vectors to be grouped,
   (b) $c_i \in \{1, \cdots, K\}$: the cluster to which $\mathbf{x}_i$ belongs,
   (c) $G_j = \{i : c_i = j\}$: the clusters, which are sets of indices representing the vectors in the group,
   (d) $\mathbf{z}_j \in \mathbb{R}^n$: the representative vector in G_j, not necessarily one of the vectors in $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$.
2. Evaluation: What problem do we want to solve? The vectors in each cluster should be close to each other.
   (a) The distance between the vectors in a cluster and the corresponding representative vector should be minimized. Therefore, we define an optimization function:

   $$d_j = \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

   Our objective for this cluster is to minimize this optimization function.
   (b) Altogether, we get the overall optimization function:

   $$d = \sum_{j=1}^K d_j.$$

   Then, we solve:

   $$\min_{\substack{G_1, \cdots, G_K \\ \mathbf{z}_1, \cdots, \mathbf{z}_K}} d \iff \min_{\substack{G_1, \cdots, G_K \\ \mathbf{z}_1, \cdots, \mathbf{z}_K}} \sum_{j=1}^K \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

3. Optimization: We now have two sets of unknowns $\{G_1, \cdots, G_K\}$ and $\{\mathbf{z}_1, \cdots, \mathbf{z}_K\}$. However, both influence each other. How do we tackle this issue? We use alternating minimization.

Step 0: Initialize $\mathbf{z}_1, \cdots, \mathbf{z}_K$.

Step 1: Fix $\mathbf{z}_1, \cdots, \mathbf{z}_K$ and solve the function with respect to $G_1, \cdots, G_K$:

$$\min_{G_1, \cdots, G_K} \sum_{j=1}^K \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

Step 2: Fix $G_1, \cdots, G_K$ and solve the function with respect to $\mathbf{z}_1, \cdots, \mathbf{z}_K$:

$$\min_{\mathbf{z}_1, \cdots, \mathbf{z}_K} \sum_{j=1}^K \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

Repeat Steps 1 and 2 until convergence is achieved.

How do we solve the functions in the alternating minimization algorithm? To obtain the clusters, we solve the following function:

$$\begin{aligned}\min_{G_1,\cdots,G_K} \sum_{j=1}^{K}\sum_{i\in G_j} \|\mathbf{x}_i-\mathbf{z}_j\|^2 &\iff \min_{c_1,\cdots,c_N} \sum_{i=1}^{N} \|\mathbf{x}_i-\mathbf{z}_{c_i}\|^2 \\ &\iff \min_{c_i\in\{1,\cdots,K\}} \|\mathbf{x}_i-\mathbf{z}_{c_i}\|^2 \\ &\iff c_i = \operatorname*{argmin}_{j\in\{1,\cdots,K\}} \|\mathbf{x}_i-\mathbf{z}_j\|, \qquad \text{for } i=1,\cdots,N.\end{aligned}$$

In simpler terms, finding clusters that minimize the distance between the vectors and the representatives is the same as assigning each vector to the cluster that minimizes the distance to its representative.
Therefore, $\mathbf{x}_i$ is assigned to the cluster whose representative is closest to $\mathbf{x}_i$. The new G_j is then created by:

$$G_j = \{i : c_i = j\}.$$

To obtain the representative vectors, we solve the following function:

$$\min_{\mathbf{z}_1,\cdots,\mathbf{z}_K} \sum_{j=1}^{K}\sum_{i\in G_j} \|\mathbf{x}_i-\mathbf{z}_j\|^2 \iff \min_{\mathbf{z}_j} \sum_{i\in G_j} \|\mathbf{x}_i-\mathbf{z}_j\|^2, \qquad \text{for } j=1,\cdots,K.$$

We only need to consider each cluster separately to find the corresponding representative vector.
At this point, we have not specified which norms are used to construct the clusters. In $\mathbb{R}^n$, one of the most commonly used norms for clustering is the 2-norm. For $j=1,\cdots,K$:

$$\min_{\mathbf{z}_j} \sum_{i\in G_j} \|\mathbf{x}_i-\mathbf{z}_j\|_2^2 \iff \min_{z_{j1},\cdots,z_{jn}} \sum_{i\in G_j}\sum_{k=1}^{n} (x_{ik}-z_{jk})^2.$$

Taking derivatives for each entry of $\mathbf{z}_j$, we have:

$$\begin{aligned}2\sum_{i\in G_j}(z_{jk}-x_{ik}) = 0 \iff z_{jk} &= \frac{1}{|G_j|}\sum_{i\in G_j} x_{ik}, \qquad \text{for } k=1,\cdots,n, \\ \iff \mathbf{z}_j &= \frac{1}{|G_j|}\sum_{i\in G_j}\mathbf{x}_i.\end{aligned}$$

This is called the K-means algorithm.

Definition A.1. The **K-means algorithm** is a method that assigns N vectors into K clusters, where each vector belongs to the cluster with the nearest mean. The steps of the algorithm are as follows:

0: Initialize $\mathbf{z}_1,\cdots,\mathbf{z}_K$.

1: Fix $\mathbf{z}_1,\cdots,\mathbf{z}_K$. For each $\mathbf{x}_i$, assign it to the cluster whose representative is closest in Euclidean distance:

$$c_i = \operatorname*{argmin}_{j\in\{1,\cdots,K\}} \|\mathbf{x}_i-\mathbf{z}_j\|_2^2, \qquad \text{for } i=1,\cdots,N, \qquad G_j=\{i:c_i=j\}, \qquad \text{for } j=1,\cdots,K.$$

2: For each cluster G_j, calculate the new $\mathbf{z}_j$ as the mean of vectors in G_j:

$$\mathbf{z}_j = \frac{1}{|G_j|}\sum_{i\in G_j}\mathbf{x}_i, \qquad \text{for } j=1,\cdots,K.$$

Repeat Step 1-2 until convergence is achieved.

What happens if we switch from the 2-norm to the 1-norm? Derivations are omitted, but we find that it turns into the K-medians algorithm.

Definition A.2. The **K-medians algorithm** is a method that assigns N vectors into K clusters, where each vector belongs to the cluster with the nearest median. The steps of the algorithm are as follows:

0: Initialize $\mathbf{z}_1, \cdots, \mathbf{z}_K$.

1: Fix $\mathbf{z}_1, \cdots, \mathbf{z}_K$. For each $\mathbf{x}_i$, assign it to the cluster whose representative is closest in Manhattan distance:

$$c_i = \underset{j \in \{1, \cdots, K\}}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{z}_j\|_1^2, \qquad \text{for } i = 1, \cdots, N, \qquad G_j = \{i : c_i = j\}, \qquad \text{for } j = 1, \cdots, K.$$

2: For each cluster G_j, calculate the new $\mathbf{z}_j$ as the median of vectors in G_j:

$$\mathbf{z}_j = \text{median}\{\mathbf{x}_i : i \in G_j\}, \qquad \text{for } j = 1, \cdots, K.$$

Repeat Step 1-2 until convergence is achieved.

Remark A.2.1. The K-means algorithm is more sensitive to outliers, while the K-medians algorithm is more robust to outliers.

Appendix B

Kernel K-means/Kernel Trick

This case study assumes that you have already read Chapter 3 and Appendix A.
In the regular K-means algorithm, we assume that the vectors can be separated linearly. However, it fails when the boundary is curved. How do we modify the K-means algorithm to deal with curved data? We can transform the data to a new domain and then apply the K-means algorithm in that transformed domain.

1. Representation: Starting with N vectors $\{\mathbf{x}_i\}_{i=1}^N$ and K clusters $\{G_j\}_{j=1}^K$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the vectors to be grouped.
   (b) $c_i \in \{1, \cdots, K\}$: the cluster to which $\mathbf{x}_i$ belongs.
   (c) $G_j = \{i : c_i = j\}$: the clusters, which are sets of indices representing the vectors in the group.
   (d) H: the feature space that contains the transformed vectors.
   (e) $\phi : \mathbb{R}^n \to H$: the feature map that transforms the vectors.
   (f) $\mathbf{z}_j \in H$: the representative vector in G_j, not necessarily one of the vectors in $\{\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_N)\}$.
2. Evaluation: With the feature map, our objective becomes finding:
$$\min_{\substack{G_1,\cdots,G_K \\ \mathbf{z}_1,\cdots,\mathbf{z}_K}} \sum_{j=1}^{K} \sum_{i \in G_j} \|\phi(\mathbf{x}_i) - \mathbf{z}_j\|^2$$
   In this algorithm, we also need to find a good ϕ that can transform the data well. If we assume that the norm we use has an inner product, our goal is to ensure that:
   (a) If $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ are close ($\|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|_2$ is small), then $\|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2$ is small.
   (b) If $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ are far apart ($\|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|_2$ is large), then $\|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2$ is large.
3. Optimization: Since ϕ depends on the shape of $\mathbf{x}_1, \cdots, \mathbf{x}_N$, it is not easy to find a good ϕ and H explicitly.
   Therefore, we can utilize the **Kernel trick** – define ϕ and H implicitly.
   Moreover, if we only need to know $G_1, \cdots, G_K$, we can eliminate $\mathbf{z}_1, \cdots, \mathbf{z}_K$ in our algorithm.

The K-means algorithm can be modified as follows:

0: Initialize $G_1, \cdots, G_K$.

1: For each $\mathbf{x}_i$, assign it to the cluster whose representative is closest in Euclidean distance:
$$c_i = \operatorname*{argmin}_{j \in \{1,\cdots,K\}} \left\| \phi(\mathbf{x}_i) - \frac{1}{|G_j|} \sum_{k \in G_j} \phi(\mathbf{x}_k) \right\|_2^2, \qquad \text{for } i = 1, \cdots, N, \quad G_j = \{i : c_i = j\}, \qquad \text{for } j = 1, \cdots, K.$$
   Repeat until convergence is achieved.

We can rearrange the equation in the algorithm:

$$\begin{aligned}\left\|\phi(\mathbf{x}_i) - \frac{1}{|G_j|}\sum_{i\in G_j}\phi(\mathbf{x}_i)\right\|_2^2 &= \left\langle \phi(\mathbf{x}_i) - \frac{1}{|G_j|}\sum_{i\in G_j}\phi(\mathbf{x}_i), \phi(\mathbf{x}_i) - \frac{1}{|G_j|}\sum_{i\in G_j}\phi(\mathbf{x}_i)\right\rangle \\ &= \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_i)\rangle - \frac{2}{|G_j|}\left\langle \phi(\mathbf{x}_i), \sum_{i\in G_j}\phi(\mathbf{x}_i)\right\rangle + \frac{1}{|G_j|^2}\left\langle \sum_{i\in G_j}\phi(\mathbf{x}_i), \sum_{i\in G_j}\phi(\mathbf{x}_i)\right\rangle \\ &= \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_i)\rangle - \frac{2}{|G_j|}\sum_{k\in G_j}\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_k)\rangle + \frac{1}{|G_j|^2}\sum_{k\in G_j}\sum_{\ell\in G_j}\langle\phi(\mathbf{x}_k),\phi(\mathbf{x}_\ell)\rangle\,.\end{aligned}$$

We can see that all the inner products are usually in the form of:

$$\langle\phi(\cdot),\phi(\cdot)\rangle\,.$$

By using the Kernel trick, we can define ϕ and H implicitly by defining the kernel function.

Definition B.1. For some ϕ, the **kernel function** is a binary operator $\kappa : (\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\kappa(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle\,.$$

We can then form a generalized K-means algorithm.

Definition B.2. The **Kernel K-means algorithm** is a method that assigns N vectors into K clusters, where each vector is transformed and classified into the cluster with the nearest mean. The steps of the algorithm are as follows:

0: Initialize $G_1, \cdots, G_K$ and define a kernel function κ.

1: For each $\mathbf{x}_i$, assign it to the cluster whose mean is the closest in Euclidean distance after transformation.

$$\begin{aligned} c_i &= \operatorname*{argmin}_{j\in\{1,\cdots,K\}}\left(\kappa(\mathbf{x}_i,\mathbf{x}_i) - \frac{2}{|G_j|}\sum_{k\in G_j}\kappa(\mathbf{x}_i,\mathbf{x}_k) + \frac{1}{|G_j|^2}\sum_{k\in G_j}\sum_{\ell\in G_j}\kappa(\mathbf{x}_k,\mathbf{x}_\ell)\right), && \text{for } i = 1,\cdots,N \\ G_j &= \{i : c_i = j\}, && \text{for } j = 1,\cdots,K.\end{aligned}$$

Repeat Step 1 until convergence is achieved.

Now the problem switches to determining which kernel function we can choose. We have some necessary conditions.

Definition B.3. The kernel function κ is a **symmetric kernel** if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{y},\mathbf{x}).$$

With a set of vectors $\mathbf{y}_1, \cdots, \mathbf{y}_m \in \mathbb{R}^n$, we can define a matrix by:

$$\boldsymbol{\kappa} = [\kappa(\mathbf{y}_i,\mathbf{y}_j)]_{i,j}.$$

Assume that we have a new vector that is the linear combination of the m transformed vectors. For any $\mathbf{z} \in \mathbb{R}^m$,

$$0 \le \left\langle \sum_{i=1}^m z_i\phi(\mathbf{y}_i), \sum_{i=1}^m z_i\phi(\mathbf{y}_i)\right\rangle = \sum_{i=1}^m\sum_{j=1}^m z_i z_j\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle = \sum_{i=1}^m\sum_{j=1}^m z_i z_j\kappa(\mathbf{y}_i,\mathbf{y}_j) = \mathbf{z}^T\boldsymbol{\kappa}\mathbf{z}.$$

Definition B.4. The kernel function κ is symmetric positive semi-definite (SPSD) if:

1. $\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{y},\mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
2. For any $m > 0$ and $\mathbf{y}_1, \cdots, \mathbf{y}_m \in \mathbb{R}^n$, the matrix:

$$\boldsymbol{\kappa} = [\kappa(\mathbf{y}_i,\mathbf{y}_j)]_{i,j}$$

is symmetric positive semi-definite.

Therefore, by the following theorem, we have a way to determine which mapping can be used as a kernel function.

Theorem B.5. (**Mercer's Theorem**) If the kernel function κ is continuous, symmetric, and positive semi-definite, then there exists a Hilbert space H and a mapping such that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle .$$

Example B.1. The traditional kernel, which involves no transformation, is defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}, \qquad \phi(\mathbf{x}) = \mathbf{x}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Applying this kernel function gives the normal K-means algorithm.

Remark B.5.1. Most of the time, finding the feature map ϕ is extremely difficult.

Example B.2. The **polynomial kernel** for $\alpha \in \mathbb{Z}$ and $c \in \mathbb{R}$ is defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^\alpha, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

If $\alpha = 2$, then the feature map is an extremely large vector defined by:

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \\ \sqrt{2} x_1 x_2 \\ \vdots \\ \sqrt{2} x_{n-1} x_n \\ \sqrt{2} c x_1 \\ \vdots \\ \sqrt{2} c x_n \\ c \end{pmatrix}, \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

A more general term would require the multinomial theorem.

Example B.3. The **Gaussian kernel** for $\sigma > 0$ is defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{y}\|_2^2} = \exp\left(-\frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{y}\|_2^2 \right), \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Obtaining its feature map would require using the Taylor expansion.

Example B.4. Consider that we use the Gaussian kernel to perform the Kernel K-means algorithm. For the same vectors:

$$\kappa(\mathbf{x}_i, \mathbf{x}_i) = \exp\left(-\frac{1}{\sigma^2} \|\mathbf{x}_i - \mathbf{x}_i\|_2^2 \right) = e^0 = 1, \qquad \text{for } i = 1, \cdots, N.$$

For different vectors, we can normalize the distances between two vectors to lie between 0 and 1 via the transformation:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) \begin{cases} \approx 1, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is small} \\ \approx 0, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is large.} \end{cases}$$

Does the kernel fulfill the goal for what we aimed for with the feature map?

$$\begin{aligned} \|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2^2 &= \|\phi(\mathbf{x}_{i_1})\|_2^2 - 2 \langle \phi(\mathbf{x}_{i_1}), \phi(\mathbf{x}_{i_2}) \rangle + \|\phi(\mathbf{x}_{i_2})\|_2^2 \\ &= \kappa(\mathbf{x}_{i_1}, \mathbf{x}_{i_1}) - 2\kappa(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) + \kappa(\mathbf{x}_{i_2}, \mathbf{x}_{i_2}). \end{aligned}$$

Therefore, the distance in transformed data is given by:

$$\|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2^2 \begin{cases} \approx 0, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is small} \\ \approx 2, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is large.} \end{cases}$$

Appendix C

Metric Learning

This case study assumes that you have already read Chapter 3.
Suppose we are given N vectors in $\mathbb{R}^n$:

$$\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^n,$$

and a set of vector pairs S and D such that:

$$(\mathbf{x}_i, \mathbf{x}_j) \in \begin{cases} S, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are similar,} \\ D, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are dissimilar (different).} \end{cases}$$

How do we find a metric $\|\cdot\|$ such that:

$$\begin{aligned} &\|\mathbf{x}_i - \mathbf{x}_j\| \text{ is small,} && \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in S, \\ &\|\mathbf{x}_i - \mathbf{x}_j\| \text{ is large,} && \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in D. \end{aligned}$$

1. Representation: How do we represent the norm? We can try using the p-norms with $p \geq 1$:

   $$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

   However, the norm set is too small for us to explore. We can use the more generalized norm induced by the weighted inner product with a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$:

   $$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}}.$$

   The set of all SPD matrices is large enough, though it is not closed. Its closure is the set of all SPSD matrices. However, for any SPSD matrices $\mathbf{A}$ that are not SPD matrices:

   $$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}} = 0, \qquad \mathbf{x}^T \mathbf{A} \mathbf{x} \not\Rightarrow \mathbf{x} = \mathbf{0}.$$

   This violates the positive-definiteness property of a norm. It is still a semi-norm.

   Definition C.1. Let V be a vector space. A **semi-norm** on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y} \in V$:

   1. $\|\mathbf{x}\| \geq 0$,
   2. $\|\alpha \mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$ for $\alpha \in \mathbb{R}$,
   3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

   Since we only want to identify similarity, non-negativity is sufficient for our task.

2. Evaluation: Which SPSD matrices are the best?

   Distance should be small for pairs in S, while distance should be large for pairs in D. Since the distance can only approach 0, but extends to ∞, we can start large and minimize the distance for the pairs in S.

   $$\min_{\mathbf{A} \in \mathbb{R}^{n \times n} \text{ is SPSD}} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \text{ such that } \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \geq 1.$$

3. Optimization: We do not perform it here.

Appendix D

Linear Regression

This case study assumes that you have already read Chapter 4.
Suppose that we are given N input-output pairs:

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$$

where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for $i = 1, \cdots, N$. We aim to predict the corresponding output of any given new $\mathbf{x}$.

1. Representations: Starting with N input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the input vectors (independent variable).
   (b) $y_i \in \mathbb{R}$: the corresponding output of $\mathbf{x}_i$ (dependent variable).
   (c) $f : \mathbb{R}^n \to \mathbb{R}$: the function that predicts the output.
2. Evaluation: We aim to find the function f such that:

$$f(\mathbf{x}_i) = y_i, \qquad \text{for } i = 1, \cdots, N.$$

   (a) The set of all functions that map from $\mathbb{R}^n$ to $\mathbb{R}$ is too large because:
      i. We do not have enough data to determine f uniquely.
      ii. There are too many functions that are not suitable for prediction.

   Therefore, we find f in a subset Φ of all functions.
   (b) The choice of Φ is fundamental since too small a subset would result in weak approximation power. We may choose a large enough subset:

$$\Phi = \{\text{all affine functions } f : \mathbb{R}^n \to \mathbb{R}\}.$$

   Our objective becomes finding a linear model f.
   (c) By Theorem 4.6(2), we can rewrite our function as:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b, \qquad \text{for } \mathbf{x} \in \mathbb{R}^n$$

   where $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Therefore, our problem is to solve:

$$\min_{\substack{\mathbf{a}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \sum_{i=1}^N (\langle \mathbf{a}, \mathbf{x}_i \rangle + b - y_i)^2 \iff \min_{\substack{\mathbf{a}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \sum_{i=1}^N (\mathbf{x}_i^T \mathbf{a} + b - y_i)^2$$

   (d) We write that:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_N^T & 1 \end{pmatrix} \in \mathbb{R}^{N\times(n+1)} \qquad \boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$$

   Therefore, our problem now is to solve:

$$\min_{\substack{\mathbf{a}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \sum_{i=1}^N (\mathbf{x}_i^T \mathbf{a} + b - y_i)^2 \iff \min_{\boldsymbol{\beta}\in\mathbb{R}^{n+1}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

3. Optimization: For a limited amount of data, can we always find a unique solution? In order to solve the following equation,
$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$
which has N equations and $n+1$ unknowns, we would need $N \geq n+1$ for a unique solution.

   However, in practice, we may also have cases when $N << n$. For example, in image recognition. Therefore, we need to perform some regularizations to further shrink the set of candidate functions. We will discuss this later. The method on how to find the solution numerically can be found in Appendix H.

The method is called the least squares.

Definition D.1. The **(linear) least squares regression** is a regression method that estimates the model f that best fits the given N input-output pairs, such that:
$$f(\mathbf{x}_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \approx y_i, \qquad \text{for } i = 1, \cdots, N$$
where $\hat{\boldsymbol{\beta}}$ is the solution of:
$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{n+1}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

Theorem D.2. If $\mathbf{X}^T\mathbf{X}$ is invertible, then least squares regression has a unique solution for $\boldsymbol{\beta}$:
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Proof.
Let X_{ij} be the (i,j)-th entry of $\mathbf{X}$ such that $X_{i(n+1)} = 1$. We can rewrite the minimization problem:
$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{n+1}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \Longleftrightarrow \min_{\beta_1, \cdots, \beta_{n+1} \in \mathbb{R}} \sum_{i=1}^{N} (X_{i1}\beta_1 + \cdots + X_{i(n+1)}\beta_{n+1} - y_i)^2$$
Taking the derivatives of β_j for $j = 1, \cdots, n+1$, we have:
$$2\sum_{i=1}^{N} X_{ij}(X_{i1}\beta_1 + \cdots + X_{i(n+1)}\beta_{n+1} - y_i) = 0 \Longleftrightarrow \sum_{i=1}^{N} X_{ij}(X_{i1}\beta_1 + \cdots + X_{i(n+1)}\beta_{n+1}) = \sum_{i=1}^{N} X_{ij}y_i$$
By substituting:
$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \sum_{i=1}^N X_{i1}^2 & \sum_{i=1}^N X_{i1}X_{i2} & \cdots & \sum_{i=1}^N X_{i1}X_{i(n+1)} \\ \sum_{i=1}^N X_{i2}X_{i1} & \sum_{i=1}^N X_{i2}^2 & \cdots & \sum_{i=1}^N X_{i2}X_{i(n+1)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^N X_{i(n+1)}X_{i1} & \sum_{i=1}^N X_{i(n+1)}X_{i2} & \cdots & \sum_{i=1}^N X_{i(n+1)}^2 \end{pmatrix}$$
We can find that:
$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$
Therefore, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ if $\mathbf{X}^T\mathbf{X}$ is invertible. □

In cases where the number of data points is significantly smaller than the dimension of the data, we have to use regularizations, which means that we have to use a subset of affine functions. The subset would be obtained by:
$$\Phi_S = \left\{ f : f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in S \right\}$$
for a given set S. Therefore, our problem now turns into solving:
$$\min_{\boldsymbol{\beta} \in S} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

We can obtain two popular choices of regularizations.

1. Ridge regression: We choose
$$S = \left\{ \boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1} : \|\mathbf{a}\|_2 \leq c \right\}$$
for some $c > 0$. Therefore, the minimization problem is to solve:
$$\min_{\boldsymbol{\beta} \in S} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \iff \min_{\boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1}} \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{a}\|_2^2 \right)$$
where $\lambda > 0$ is a constant depending on c and other factors. The ridge regression is defined as follows:

Definition D.3. The **ridge regression** is a regression method that estimates the model f that best fits the given N input-output pairs, such that:
$$f(\mathbf{x}_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \approx y_i, \qquad \text{for } i = 1, \cdots, N$$
where $\hat{\boldsymbol{\beta}}$ is the solution of:
$$\min_{\boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1}} \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{a}\|_2^2 \right), \qquad \text{for } \lambda > 0$$

Theorem D.4. Ridge regression has a unique solution for $\boldsymbol{\beta}$:
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

This minimization problem can be separated into three parts:

(a) Data-fitting term: $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$.

(b) Regularization term: $\|\mathbf{a}\|_2^2$.

(c) Regularization parameter: λ.
A larger λ means a smaller Φ_S, allowing for looser fitting.
A smaller λ means a larger Φ_S, allowing for better fitting.

2. LASSO regression: We choose
$$S = \left\{ \boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1} : \|\mathbf{a}\|_1 \leq c \right\}$$
for some $c > 0$. Therefore, the minimization problem is to solve:
$$\min_{\boldsymbol{\beta} \in S} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \iff \min_{\boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1}} \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{a}\|_1^2 \right)$$
where $\lambda > 0$ is a constant depending on c and other factors. The LASSO regression is defined as follows:

Definition D.5. The **LASSO regression** is a regression method that estimates the model f that best fits the given N input-output pairs, such that:
$$f(\mathbf{x}_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \approx y_i, \qquad \text{for } i = 1, \cdots, N$$
where $\hat{\boldsymbol{\beta}}$ is the solution of:
$$\min_{\boldsymbol{\beta} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \mathbb{R}^{n+1}} \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{a}\|_1^2 \right), \qquad \text{for } \lambda > 0$$

Similarly, this minimization problem can be separated into three parts:

(a) Data-fitting term: $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$.

(b) Regularization term: $\|\mathbf{a}\|_1^2$.

(c) Regularization parameter: λ.

What are the differences between these two types of regularizations?
In general, LASSO gives a sparse vector $\mathbf{a}$ (more zeros in $\mathbf{a}$). Before we dive in, we first define the support of a vector.

Definition D.6. Let V be a vector space over $\mathbb{R}$ and let $\mathbf{x} \in V$. The **support** of $\mathbf{x}$ is given by:

$$\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}.$$

The model now predicts the output by:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b = \sum_{i=1}^{n} a_i x_i + b = \sum_{i \in \text{supp}(\mathbf{a})} a_i x_i + b$$

Only x_i where $i \in \text{supp}(\mathbf{a})$ contributes to the prediction. If $\mathbf{a}$ is sparse, then:

$$|\text{supp}(\mathbf{a})| << n$$

This allows for more interpretable predictions.

Appendix E

Kernel Ridge Regression

This case study assumes that you have already read Chapter 4 and Appendices B and D.
Similarly, in regular linear regression, we assume that a linear model is sufficient to predict a desirable output. In practice, however, points in datasets often exhibit non-linear relationships. Therefore, similar to the Kernel K-means algorithm, we can also apply the kernel method in regression.

1. Representations: Starting with N input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the input vectors (independent variables).
   (b) $y_i \in \mathbb{R}$: the corresponding outputs of $\mathbf{x}_i$ (dependent variables).
   (c) H: the feature space that contains the transformed input vectors.
   (d) $\phi : \mathbb{R}^n \to H$: the feature map that transforms the input vectors.
   (e) $f : H \to \mathbb{R}$: the function that predicts the outputs.
2. Evaluation: With the feature map, our objective becomes finding an affine function f such that:
$$f(\phi(\mathbf{x}_i)) = y_i, \qquad \text{for } i = 1, \cdots, N.$$
   How do we find a good H, ϕ, and f? Assume that $f : H \to \mathbb{R}$ is affine and bounded. For any $\mathbf{x} \in \mathbb{R}^n$, we may define a new $\widetilde{\phi}$ and $\widetilde{H} = (H, \mathbb{R})$:
$$\widetilde{\phi}(\mathbf{x}) = \begin{pmatrix} \phi(\mathbf{x}) \\ 1 \end{pmatrix} \in \widetilde{H}.$$
   Then for any $\mathbf{x} \in \mathbb{R}^n$,
$$\begin{aligned} f(\phi(\mathbf{x})) &= \langle \phi(\mathbf{x}), \mathbf{a} \rangle + b \\ &= \left\langle \begin{pmatrix} \phi(\mathbf{x}) \\ 1 \end{pmatrix}, \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \right\rangle \\ &= \left\langle \widetilde{\phi}(\mathbf{x}), \widetilde{\mathbf{a}} \right\rangle, \qquad \text{where } \widetilde{\mathbf{a}} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix} \in \widetilde{H}, \end{aligned}$$
   where $\mathbf{a} \in H$ and $b \in \mathbb{R}$. Therefore, $f(\phi(\mathbf{x}))$ is a linear and bounded function on $\widetilde{H}$.
   Thus, we only need to consider linear and bounded functions on H, which means that we want to find $\mathbf{a} \in H$ such that:
$$\langle \mathbf{a}, \phi(\mathbf{x}_i) \rangle \approx y_i, \qquad \text{for } i = 1, \cdots, N.$$
   Therefore, our problem is to solve:
$$\min_{\mathbf{a} \in \mathbb{R}^n} \sum_{i=1}^{N} (\langle \mathbf{a}, \phi(\mathbf{x}_i) \rangle - y_i)^2.$$
   We still need to find explicit ϕ and H.
3. Optimization:
   (a) The feature space H is infinitely dimensional in most cases. It is impossible to find $\mathbf{a}$ with N data points only. Therefore, we need to perform regularization directly.
   (b) We can utilize the Kernel trick—define the kernel function κ implicitly.

How do we perform regularization? Recall that we use a subset of affine functions when performing regularization for regular ridge regression. The subset would now be:

$$\Phi_S = \{f : f(\mathbf{x}) = \langle \mathbf{a}, \phi(\mathbf{x}) \rangle \text{ and } \mathbf{a} \in S\}.$$

Assume that H has a norm $\|\cdot\|_H$. We choose

$$S = \{\mathbf{a} \in \mathbb{R}^n : \|\mathbf{a}\|_H \leq c\},$$

for some $c > 0$. Therefore, the minimization problem is to solve:

$$\min_{\mathbf{a} \in H} \left(\sum_{i=1}^{N} (\langle \mathbf{a}, \phi(\mathbf{x}_i) \rangle - y_i)^2 + \lambda \|\mathbf{a}\|_H^2 \right), \tag{KR}$$

where $\lambda > 0$ is a constant depending on c and other factors.
How do we find the vector $\mathbf{a}$? The following theorem provides the solution.

Theorem E.1. (**Representer Theorem**) The solution of (KR) must be in the form:

$$\mathbf{a} = \sum_{i=1}^{N} c_i \phi(\mathbf{x}_i), \qquad \text{where } \mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix} \in \mathbb{R}^N.$$

Proof.
For any $\mathbf{a} \in H$, we claim that $\mathbf{a}$ can be decomposed as:

$$\mathbf{a} = \mathbf{a}_S + \sum_{i=1}^{N} c_i \phi(\mathbf{x}_i), \qquad \text{where } \mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix} \in \mathbb{R}^N, \langle \mathbf{a}_S, \phi(\mathbf{x}_i) \rangle = 0, \qquad \text{for } i = 1, \cdots, N.$$

We first prove the case when $N = 1$. Let $S = \{\mathbf{v} \in H : \langle \mathbf{v}, \phi(\mathbf{x}_1) \rangle = 0\}$, which is a hyperplane. We may write $\mathbf{a}$ as:

$$\mathbf{a} = P_S \mathbf{a} + (\mathbf{a} - P_S \mathbf{a}).$$

By Theorem 4.11, setting $\mathbf{x} = \mathbf{0}$, we have:

$$\langle \mathbf{a} - P_S \mathbf{a}, P_S \mathbf{a} \rangle = 0.$$

Moreover, by the explicit formula for $P_S \mathbf{a}$,

$$\mathbf{a} - P_S \mathbf{a} = \frac{\langle \phi(\mathbf{x}_1), \mathbf{a} \rangle}{\|\phi(\mathbf{x}_1)\|_H^2} \phi(\mathbf{x}_1).$$

Therefore, by setting:

$$c_i = \frac{\langle \phi(\mathbf{x}_1), \mathbf{a} \rangle}{\|\phi(\mathbf{x}_1)\|_H^2}, \qquad \mathbf{a}_S = P_S \mathbf{a},$$

we obtain:

$$\mathbf{a} = \mathbf{a}_S + c_1 \phi(\mathbf{x}_1), \qquad \text{where } \langle \mathbf{a}_S, \phi(\mathbf{x}_1) \rangle = 0.$$

For general N, we can use projection onto the intersection of the hyperplanes:

$$S = \bigcap_{i=1}^{N} \{\mathbf{v} \in H : \langle \mathbf{v}, \phi(\mathbf{x}_i) \rangle = 0\}.$$

This concludes the claim. The proof continues on the next page.

We may rearrange the objective function:

$$\begin{aligned}\sum_{i=1}^{N}(\langle \mathbf{a}, \phi(\mathbf{x}_i)\rangle - y_i)^2 + \lambda \|\mathbf{a}\|_H^2 &= \sum_{i=1}^{N}\left(\left\langle \mathbf{a}_S + \sum_{j=1}^{N} c_i\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\right\rangle - y_i\right)^2 + \lambda\left\|\mathbf{a}_S + \sum_{i=1}^{N} c_i\phi(\mathbf{x}_i)\right\|_H^2 \\ &= \sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j\langle\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle - y_i\right)^2 + \lambda\left\|\sum_{i=1}^{N} c_i\phi(\mathbf{x}_i)\right\|_H^2 + \lambda\|\mathbf{a}_S\|_H^2 \\ &= \sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j\langle\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle - y_i\right)^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N} c_ic_j\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle + \lambda\|\mathbf{a}_S\|_H^2 .\end{aligned}$$

We can define the kernel function and kernel matrix as follows:

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle\phi(\mathbf{x}), \phi(\mathbf{y})\rangle , \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \qquad\qquad \boldsymbol{\kappa} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}.$$

Therefore, we may rewrite the objective function as:

$$(\text{KR}) \iff \min_{\substack{\mathbf{c}\in\mathbb{R}^N \\ \mathbf{a}_S\in H}} (\|\boldsymbol{\kappa}^T\mathbf{c} - \mathbf{y}\|_2^2 + \lambda\mathbf{c}^T\boldsymbol{\kappa}\mathbf{c} + \lambda\|\mathbf{a}_S\|_H^2), \qquad \text{such that } \langle\mathbf{a}_S, \phi(\mathbf{x}_i)\rangle = 0 \text{ for } i = 1, \cdots, N. \tag{P1}$$

For now, we drop the constraints in (P1):

$$\min_{\substack{\mathbf{c}\in\mathbb{R}^N \\ \mathbf{a}_S\in H}} (\|\boldsymbol{\kappa}^T\mathbf{c} - \mathbf{y}\|_2^2 + \lambda\mathbf{c}^T\boldsymbol{\kappa}\mathbf{c} + \lambda\|\mathbf{a}_S\|_H^2) \iff \min_{\mathbf{c}\in\mathbb{R}^n}(\|\boldsymbol{\kappa}^T\mathbf{c} - \mathbf{y}\|_2^2 + \lambda\mathbf{c}^T\boldsymbol{\kappa}\mathbf{c}) \text{ and } \min_{\mathbf{a}_S\in H} \lambda\|\mathbf{a}_S\|_H^2 . \tag{P2}$$

We set the solution of (P2) to be:

$$\mathbf{c}^* = \operatorname*{argmin}_{\mathbf{c}\in\mathbb{R}^N}(\|\boldsymbol{\kappa}^T\mathbf{c} - \mathbf{y}\|_2^2 + \lambda\mathbf{c}^T\boldsymbol{\kappa}\mathbf{c}), \qquad\qquad \mathbf{a}_S^* = \mathbf{0}.$$

We show that $(\mathbf{c}^*, \mathbf{a}_S^*)$ is also a solution of (P1).
Since (P2) only removed the constraints from (P1), we find that:

$$(\text{P1}) \geq (\text{P2}).$$

Since $\mathbf{a}_S^* = \mathbf{0}$, we find that:

$$\langle\mathbf{a}_S^*, \phi(\mathbf{x}_i)\rangle = 0, \qquad \text{for } i = 1, \cdots, N.$$

Therefore, $(\mathbf{c}^*, \mathbf{a}_S^*)$ satisfies the constraints in (P1), and we have:

$$(\text{P1}) \leq (\text{P2}).$$

Thus, $(\mathbf{c}^*, \mathbf{a}_S^*)$ is a solution of (P1), and the solution of (KR) is given by:

$$\mathbf{a} = \mathbf{a}_S^* + \sum_{i=1}^{N} c_i^*\phi(\mathbf{x}_i) = \sum_{i=1}^{N} c_i^*\phi(\mathbf{x}_i), \qquad \text{where } \mathbf{c}^* = \operatorname*{argmin}_{\mathbf{c}\in\mathbb{R}^N}(\|\boldsymbol{\kappa}^T\mathbf{c} - \mathbf{y}\|_2^2 + \lambda\mathbf{c}^T\boldsymbol{\kappa}\mathbf{c}).$$

□

Similar to the K-Means Algorithm, we may define Kernel ridge regression as follows:

Definition E.2. Kernel ridge regression is a regression method that estimates the model f that best fits the given N input-output pairs, where each input vector is transformed such that:

$$f(\phi(\mathbf{x}_i)) \approx y_i, \qquad \text{for } i = 1, \cdots, N.$$

The steps of the algorithm are as follows:

1: Choose a kernel function κ.

2: Form a kernel matrix $\boldsymbol{\kappa} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{i,j} \in \mathbb{R}^{N \times N}$.

3: Solve $\mathbf{c}^*$ from:

$$\mathbf{c}^* = \underset{\mathbf{c} \in \mathbb{R}^N}{\operatorname{argmin}} (\|\boldsymbol{\kappa}^T \mathbf{c} - \mathbf{y}\|_2^2 + \lambda \mathbf{c}^T \boldsymbol{\kappa} \mathbf{c}).$$

4: Obtain the regression function:

$$f(\phi(\mathbf{x})) = \sum_{i=1}^{N} c_i^* \kappa(\mathbf{x}_i, \mathbf{x}), \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

Appendix F

Linear Classification

This case study assumes that you have already read Chapter 4 and Appendices B and E.
Suppose that we are given N input-output pairs:

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$$

where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for $i = 1, \cdots, N$. We aim to classify any given new $\mathbf{x}$ as either -1 or 1.

1. Representation: Starting with N input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the input vectors.
   (b) $y_i \in \{-1, 1\}$: the corresponding class of $\mathbf{x}_i$.
   (c) $f : \mathbb{R}^n \to \mathbb{R}$: the function that allows for classification based on its output for $\mathbf{x}_i$.
2. Evaluation: We use the function f to define a separation between two classes.
   (a) Given a new $\mathbf{x} \in \mathbb{R}^n$, the predicted label would be:

   $$y = \operatorname{sgn}(f(\mathbf{x}))$$

   Here, we use $\{\mathbf{x} : f(\mathbf{x}) = 0\}$ to separate the two classes. To enhance confidence, we should find a function such that:

   $$\begin{cases} f(\mathbf{x}_i) \geq 1, & \text{if } y_i = 1, \\ f(\mathbf{x}_i) \leq -1, & \text{if } y_i = -1 \end{cases}, \qquad \text{for } i = 1, \cdots, N.$$

   (b) What function class should f belong to? Similar to linear regression, we can use the set of all affine functions. We need to find $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that:

   $$\begin{cases} \langle \mathbf{a}, \mathbf{x}_i \rangle + b \geq 1, & \text{if } y_i = 1, \\ \langle \mathbf{a}, \mathbf{x}_i \rangle + b \leq -1, & \text{if } y_i = -1 \end{cases}, \qquad \text{for } i = 1, \cdots, N.$$

   This means we are using hyperplanes to separate points in different classes.
   (c) Which hyperplane is the best among all the possible ones? The larger the margin, the more stable the classification.

   Definition F.1. The **Support Vector Machine (SVM)** is an algorithm that finds the best hyperplane that maximizes the margin between the two classes. Let:

   $$S_+ = \{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle + b = 1\}, \qquad S_- = \{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle + b = -1\}.$$

   The margin is defined by:

   $$\|\mathbf{x}_+ - \mathbf{x}_-\|_2, \qquad \text{where } \begin{cases} \mathbf{x}_+ = P_{S_+}\mathbf{x}_-, \\ \mathbf{x}_- = P_{S_-}\mathbf{x}_+. \end{cases}$$

   By the projection formula, since $\mathbf{x}_- \in S_-$,

   $$\mathbf{x}_+ = \mathbf{x}_- - \left(\frac{\langle \mathbf{a}, \mathbf{x}_- \rangle - (1 - b)}{\|\mathbf{a}\|_2^2} \right) \mathbf{a} = \mathbf{x}_- - \left(\frac{(-1 - b) - (1 - b)}{\|\mathbf{a}\|_2^2} \right) \mathbf{a} = \mathbf{x}_- + \frac{2}{\|\mathbf{a}\|_2^2} \mathbf{a}.$$

   Therefore, the margin is $\|\mathbf{x}_+ - \mathbf{x}_-\|_2 = \frac{2}{\|\mathbf{a}\|_2}$.

We can prevent data points from falling into the margin. Therefore, our problem is to solve:

$$\max_{\substack{\mathbf{a}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \frac{2}{\|\mathbf{a}\|_2} \text{ such that } \begin{cases} \langle \mathbf{a}, \mathbf{x}_i \rangle + b \geq 1, & \text{for } y_i = 1, \\ \langle \mathbf{a}, \mathbf{x}_i \rangle + b \leq -1, & \text{for } y_i = -1 \end{cases}, \qquad \text{for } i = 1, \cdots, N.$$

The hard-margin SVM is defined as follows:

Definition F.2. The **hard-margin Support Vector Machine** is an algorithm that finds the best hyperplane S that maximizes the margin between the two classes, requiring that all data points be linearly separable and not allowed to be on the other side. The hyperplane is in the form of:

$$S = \{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle + b = 0\},$$

where $(\mathbf{a}, b)$ is the solution of:

$$\min_{\substack{\mathbf{a}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \|\mathbf{a}\|_2^2 \text{ such that } y_i(\langle \mathbf{a}, \mathbf{x}_i \rangle + b) \geq 1, \qquad \text{for } i = 1, \cdots, N. \tag{SVM-1}$$

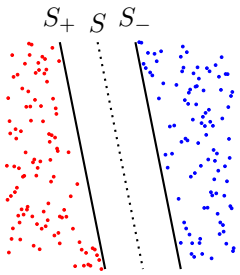


Figure F.1: Hard-margin SVM

Remark F.2.1. (SVM-1) is not robust to noise in training data.

Remark F.2.2. In some cases, the hyperplane cannot separate the two classes in the presence of high noise.

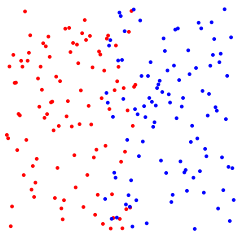


Figure F.2: No solution to (SVM-1)

Is there a way to improve SVM? We need to first reformulate (SVM-1). We define a function $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ by:

$$h(t) = \begin{cases} 0, & \text{if } t \geq 0, \\ +\infty, & \text{if } t < 0. \end{cases}$$

Then we can rewrite (SVM-1) as:

$$\text{(SVM-1)} \iff \min_{\substack{\mathbf{a}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \left(\sum_{i=1}^{N} h(y_i(\langle \mathbf{a}, \mathbf{x}_i \rangle + b) - 1) + \lambda \|\mathbf{a}\|_2^2 \right),$$

where $\lambda > 0$. This still fulfills the constraint that no data points must sit on the margin.

In order to consider the case when some data points cannot be separated with a hard margin, we can soften the function h by defining it as:

$$h(t) = \max(0, -t) = \begin{cases} 0, & \text{if } t \geq 0, \\ |t|, & \text{if } t < 0. \end{cases}$$

This function is called the hinge loss function. The soft-margin SVM is defined as follows:

Definition F.3. The **soft-margin Support Vector Machine** is an algorithm that finds the best hyperplane S that maximizes the margin between the two classes, allowing some data points to be misclassified or fall within the margin. The hyperplane is in the form of:

$$S = \{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle + b = 0\},$$

where $(\mathbf{a}, b)$ is the solution of:

$$\min_{\substack{\mathbf{a} \in \mathbb{R}^n \\ b \in \mathbb{R}}} \left(\sum_{i=1}^{N} \max(0, 1 - y_i(\langle \mathbf{a}, \mathbf{x}_i \rangle + b)) + \lambda \|\mathbf{a}\|_2 \right), \tag{SVM-2}$$

where $\lambda > 0$.

Remark F.3.1. We can replace h with some other functions. For example,

$$h(t) = \log(1 + e^{-t}),$$

which is a smooth function that approximates hinge loss, called logistic loss.

So far, we have only introduced classifiers that are linear. Similar to what we did in regression, we can apply the Kernel trick to create a non-linear classifier. Consider a feature map $\phi : \mathbb{R}^n \to H$. The hyperplane will become:

$$S = \{\mathbf{x} : \langle \mathbf{a}, \phi(\mathbf{x}) \rangle = 0\}.$$

The formulation of SVM becomes:

$$\begin{aligned} \sum_{i=1}^{N} h(y_i \langle \mathbf{a}, \phi(\mathbf{x}_i) \rangle - 1) + \lambda \|\mathbf{a}\|_H^2 &= \sum_{i=1}^{N} h\left(y_i \left\langle \mathbf{a}_S + \sum_{j=1}^{N} c_j \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \right\rangle - 1 \right) + \lambda \left\| \mathbf{a}_S + \sum_{j=1}^{N} c_j \phi(\mathbf{x}_j) \right\|_H^2 \\ &= \sum_{i=1}^{N} h\left(y_i \sum_{j=1}^{N} c_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle - 1 \right) + \lambda \|\mathbf{a}_S\|_H^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle . \end{aligned}$$

We introduce the kernel function $\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and the kernel matrix $\boldsymbol{\kappa} = \{\kappa(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^N \in \mathbb{R}^{N \times N}$. Therefore, we would be solving:

$$\min_{\substack{\mathbf{c} \in \mathbb{R}^N \\ \mathbf{a}_S \in H}} (h(y_i[\boldsymbol{\kappa}^T \mathbf{c}]_i - 1) + \lambda \mathbf{c}^T \boldsymbol{\kappa} \mathbf{c} + \lambda \|\mathbf{a}_S\|_H^2), \qquad \text{such that } \langle \mathbf{a}_S, \phi(\mathbf{x}_i) \rangle = 0 \text{ for } i = 1, \cdots, N. \tag{K-SVM}$$

Similar to Appendix E, by the Representer Theorem, the solution of (K-SVM) is:

$$\mathbf{a} = \sum_{i=1}^{N} c_i^* \phi(\mathbf{x}_i), \qquad \mathbf{c}^* = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^N} \left(\sum_{i=1}^{N} h(y_i[\kappa^T \mathbf{c}]_i - 1) + \lambda \mathbf{c}^T \boldsymbol{\kappa} \mathbf{c} \right), \qquad \mathbf{a}_S = \mathbf{0}.$$

Therefore, we may define the Kernel SVM as follows:

Definition F.4. The **Kernel Support Vector Machine** is an algorithm that finds the best hyperplane that maximizes the margin between the two classes in the feature space, where the input vectors are transformed. The hyperplane is in the form of:

$$S = \{\mathbf{x} : \langle \mathbf{a}, \phi(\mathbf{x}) \rangle = 0\},$$

where $\mathbf{a}$ can be found by the following algorithm:

1: Choose a kernel function κ.

2: Form a kernel matrix $\boldsymbol{\kappa} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{i,j} \in \mathbb{R}^{N \times N}$.

3: Solve $\mathbf{c}^*$ from:

$$\underset{\mathbf{c} \in \mathbb{R}^N}{\operatorname{argmin}} \left(\sum_{i=1}^{N} h(y_i [\kappa^T \mathbf{c}]_i - 1) + \lambda \mathbf{c}^T \boldsymbol{\kappa} \mathbf{c} \right).$$

4: Obtain the classifier function:

$$\langle \mathbf{a}, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{N} c_i^* \kappa(\mathbf{x}_i, \mathbf{x}), \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

5: Predict the label of any input:

$$y = \operatorname{sgn} \left(\sum_{i=1}^{N} c_i^* \kappa(\mathbf{x}_i, \mathbf{x}) \right), \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

Appendix G

Solvability and Optimality

This case study assumes that you have already read Chapter 4.
Let H be a Hilbert space. Suppose that we have a function $f : H \to \mathbb{R}$. We aim to find the solution of:

$$\min_{\mathbf{x}\in H} f(\mathbf{x}). \tag{OPT}$$

1. Representations: We define the following variables:
   (a) $\mathbf{x}^* \in H$: the solution of (OPT).
   (b) $f : H \to \mathbb{R}$: the function that we want to minimize.
2. Evaluation: We would mostly focus on this part. More specifically, we want to ask:
   (a) What optimality and solvability conditions does the solution need to satisfy?
   (b) What conditions does the function need to satisfy to guarantee a solution?
3. Optimization: This will be discussed in Appendix H.

From the problem, we can immediately say $\mathbf{x}$ is a solution of (OPT) if:

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in H} f(\mathbf{x}) \iff f(\mathbf{x}^*) \leq f(\mathbf{x}), \qquad \text{for } \mathbf{x} \in H. \tag{0th optimality condition}$$

Definition G.1. If $\mathbf{x}^* \in H$ satisfies the 0th optimality condition, then $\mathbf{x}^*$ is called a **global minimizer** of f.

Remark G.1.1. The existence of $\mathbf{x}^*$ is not guaranteed automatically.

Example G.1. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by:

$$f(x) = x(x-1)(x+1), \qquad \text{for } x \in \mathbb{R}.$$

When $x \to -\infty$, $f(x) \to -\infty$. There is no solution to (OPT).

Example G.2. Let $f : H \to \mathbb{R}$ be defined by:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle, \qquad \text{for } \mathbf{x} \in H,$$

for some non-zero $\mathbf{a} \in H$. Then, setting $\mathbf{x} = c\mathbf{a}$ for some $c \in \mathbb{R}$, we have:

$$f(c\mathbf{a}) = \langle \mathbf{a}, c\mathbf{a} \rangle = c \|\mathbf{a}\|^2.$$

When $c \to -\infty$, $f(\mathbf{x}) \to -\infty$. Therefore, there is no solution to (OPT).

Example G.3. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by:

$$f(x) = \frac{1}{x}, \qquad \text{for } x > 0.$$

Since $f(x)$ is a decreasing function for $x > 0$, for any solution x^*, we can always find a more optimal solution larger than x^*. Therefore, there is no solution to (OPT).

It seems that only the 0th-order optimality condition does not guarantee a solution. Let us investigate what additional conditions need to be satisfied.

Theorem G.2. Let $f : H \to \mathbb{R}$ be differentiable at $\mathbf{x}^* \in H$. Then,

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x} \in H} f(\mathbf{x}) \implies \nabla f(\mathbf{x}^*) = \mathbf{0}.$$

Proof.
By the Taylor expansion, for any $\mathbf{x} \in H$,

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle + o(\|\mathbf{x} - \mathbf{x}^*\|).$$

Suppose that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. We choose $\widetilde{\mathbf{x}} = \mathbf{x}^* - t\nabla f(\mathbf{x}^*)$ with $t > 0$. We have:

$$f(\widetilde{\mathbf{x}}) = f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), -t\nabla f(\mathbf{x}^*) \rangle + o(t \|\nabla f(\mathbf{x}^*)\|) = f(\mathbf{x}^*) - t \|\nabla f(\mathbf{x}^*)\|^2 + o(t \|\nabla f(\mathbf{x}^*)\|).$$

Since we have:

$$\lim_{t \to 0} \frac{o(t \|\nabla f(\mathbf{x}^*)\|)}{t \|\nabla f(\mathbf{x}^*)\|} = 0,$$

for all $c > 0$, there exists $t > 0$ such that:

$$ct \|\nabla f(\mathbf{x}^*)\| > o(t \|\nabla f(\mathbf{x}^*)\|).$$

We choose $c = \|\nabla f(\mathbf{x}^*)\| > 0$. Then,

$$t \|\nabla f(\mathbf{x}^*)\|^2 > o(t \|\nabla f(\mathbf{x}^*)\|).$$

Therefore, $f(\widetilde{\mathbf{x}}) < f(\mathbf{x}^*)$, which contradicts the 0th optimality condition. Thus, $\nabla f(\mathbf{x}^*) = \mathbf{0}$. □

Remark G.2.1. The reverse is not true in general.

If we just look at $\nabla f(\mathbf{x}^*) = 0$, what can it be?

Definition G.3. For any $\mathbf{x}^* \in H$, if there exists $\varepsilon > 0$ such that for all $\mathbf{x} \in H$ such that $\|\mathbf{x} - \mathbf{x}^*\| \leq \varepsilon$,

$$f(\mathbf{x}^*) \leq f(\mathbf{x}),$$

then $\mathbf{x}^*$ is called a **local minimizer**.

Definition G.4. For any $\mathbf{x}^* \in H$, if there exist $\mathbf{u}, \mathbf{v} \in H$ and $\varepsilon > 0$ such that for all $t \in \mathbb{R}$ with $|t| \leq \varepsilon$,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^* + t\mathbf{u}) \text{ and } f(\mathbf{x}^*) \leq f(\mathbf{x}^* + t\mathbf{v}),$$

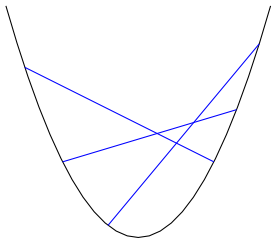
then $\mathbf{x}^*$ is called a **saddle point**.

Remark G.4.1. Saddle points only exist if $\dim(H) \geq 2$.

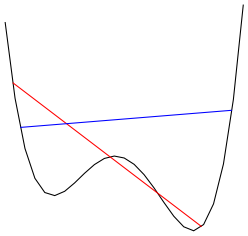
Global minimizers and maximizers, local minimizers and maximizers, and saddle points all satisfy $\nabla f(\mathbf{x}^*) = \mathbf{0}$. If we want only the global minimizer to satisfy this condition, what restrictions should we impose on the function?

Definition G.5. Let V be a vector space. A function $f : V \to \mathbb{R}$ is called **convex** if, for any $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$



(a) Convex



(b) Non-convex

Example G.4. Let V be a vector space with norm $\|\cdot\|$, and let $f : V \to \mathbb{R}$ be defined by:

$$f(\mathbf{x}) = \|\mathbf{x}\|, \qquad \text{for } \mathbf{x} \in V.$$

For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in [0, 1]$, by the triangle inequality,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = \|\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\| \leq \alpha \|\mathbf{x}\| + (1-\alpha) \|\mathbf{y}\| = \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$

Therefore, f is convex.

Remark G.5.1. $f(\mathbf{x}) = \|\mathbf{x}\|$ is convex for any norm on V.

Example G.5. Let V be a vector space with norm $\|\cdot\|$, and let $f : V \to \mathbb{R}$ be defined by:

$$f(\mathbf{x}) = \|\mathbf{x}\|^2, \qquad \text{for } \mathbf{x} \in V.$$

For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in [0, 1]$,

$$\begin{aligned}
f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \|\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\|^2 \\
&= \alpha^2 \|\mathbf{x}\|^2 + (1-\alpha)^2 \|\mathbf{y}\|^2 + 2\alpha(1-\alpha) \langle \mathbf{x}, \mathbf{y} \rangle \\
&= \alpha \|\mathbf{x}\|^2 + (1-\alpha) \|\mathbf{y}\|^2 + 2\alpha(1-\alpha) \langle \mathbf{x}, \mathbf{y} \rangle + (\alpha^2 - \alpha) \|\mathbf{x}\|^2 + (\alpha^2 - \alpha) \|\mathbf{y}\|^2 \\
&= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - \alpha(1-\alpha)(\|\mathbf{x}\|^2 - 2 \langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2) \\
&= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - \alpha(1-\alpha) \|\mathbf{x} - \mathbf{y}\|^2 \\
&\leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).
\end{aligned}$$

Therefore, f is convex.

Example G.6. Let V be a vector space, and let $f : V \to \mathbb{R}$ be defined by:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad \text{for } \mathbf{x} \in V,$$

where $g(\mathbf{x})$ is linear and b is a constant. For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in [0, 1]$,

$$\begin{aligned}
f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) + b \\
&= \alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y}) + (\alpha b + (1-\alpha) b) \\
&= \alpha(g(\mathbf{x}) + b) + (1-\alpha)(g(\mathbf{y}) + b) \\
&= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).
\end{aligned}$$

Therefore, f is convex.

Theorem G.6. Let V be a vector space, and let $f_1, f_2, \cdots, f_n : V \to \mathbb{R}$ be convex functions. Let $f : V \to \mathbb{R}$ be defined by:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i f_i(\mathbf{x}), \qquad \text{for } \mathbf{x} \in V,$$

where $c_i \geq 0$ for $i = 1, \cdots, n$. Then f is convex.

Proof.
For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in [0, 1]$, since $c_i \geq 0$,

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \sum_{i=1}^{n} c_i f_i(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \\ &\leq \sum_{i=1}^{n} c_i(\alpha f_i(\mathbf{x}) + (1-\alpha) f_i(\mathbf{y})) \\ &= \alpha \sum_{i=1}^{n} c_i f_i(\mathbf{x}) + (1-\alpha) \sum_{i=1}^{n} c_i f_i(\mathbf{y}) \\ &= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}). \end{aligned}$$

Therefore, f is convex. □

Remark G.6.1. Both f_1 and f_2 being convex does not necessarily imply that $f_1 - f_2$ is convex.

Example G.7. Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be defined by:

$$f_1(x) = x, \qquad f_2(x) = x^2, \qquad \text{for } x \in \mathbb{R}.$$

If we set $x = 0$, $y = 1$, $\alpha = 0.5$,

$$(f_1 - f_2)(\alpha x + (1-\alpha)y) = 0.25, \qquad \alpha(f_1 - f_2)(x) + (1-\alpha)(f_1 - f_2)(y) = 0 < (f_1 - f_2)(\alpha x + (1-\alpha)y).$$

Therefore, $f_1 - f_2$ is not convex.

Theorem G.7. Let V be a vector space. Let $f : \mathbb{R} \to \mathbb{R}$ be convex, and let $g : V \to \mathbb{R}$ be affine. Then $f \circ g$ is convex.

Proof.
For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in [0, 1]$,

$$(f \circ g)(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = f(\alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y})) \leq \alpha(f \circ g)(\mathbf{x}) + (1-\alpha)(f \circ g)(\mathbf{y}).$$

Therefore, $f \circ g$ is convex. □

Remark G.7.1. f and g being convex does not necessarily imply that $f \circ g$ is convex.

Example G.8. Let $f, g : \mathbb{R} \to \mathbb{R}$ be defined by:

$$f(x) = -\ln(x), \qquad g(x) = \frac{1}{x}, \qquad \text{for } x > 0.$$

We know that both f and g are convex since they are decreasing. However,

$$(f \circ g)(x) = \ln(x),$$

which is an increasing function. Therefore, $f \circ g$ is not convex.

Before proving that f being convex is sufficient to find a solution, we first need the following lemma.

Lemma G.8. Let H be a Hilbert space. If $f : H \to \mathbb{R}$ is differentiable, then:

$$f \text{ is convex} \iff f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \qquad \text{for } \mathbf{x}, \mathbf{y} \in H.$$

Proof.

$\implies$ We first prove the case when $H = \mathbb{R}$. Assume that $f : \mathbb{R} \to \mathbb{R}$ is convex. Then, for any $x, y \in \mathbb{R}$ and $\alpha \in [0, 1)$,

$$\begin{aligned} f(y) &\geq \frac{f(\alpha x + (1-\alpha)y) - \alpha f(x)}{1-\alpha} \\ &\geq f(x) + \frac{f(\alpha x + (1-\alpha)y) - f(x)}{1-\alpha} \\ &\geq f(x) + \frac{f(x + (1-\alpha)(y-x)) - f(x)}{(1-\alpha)(y-x)}(y-x). \end{aligned}$$

As $\alpha \to 1^-$, for all $x, y \in \mathbb{R}$,

$$f(y) \geq f(x) + f'(x)(y-x).$$

Now, we prove the general case. Assume that $f : H \to \mathbb{R}$ is convex. For any $\mathbf{x}, \mathbf{y} \in H$, let:

$$g(t) = f(t\mathbf{x} + (1-t)\mathbf{y}), \qquad \text{for } t \in \mathbb{R}.$$

For any $s, t \in \mathbb{R}$ and $\alpha \in [0, 1]$,

$$\begin{aligned} g(\alpha s + (1-\alpha)t) &= f((\alpha s + (1-\alpha)t)\mathbf{x} + (1 - \alpha s - (1-\alpha)t)\mathbf{y}) \\ &= f(\alpha(s\mathbf{x} + (1-s)\mathbf{y}) + (1-\alpha)(t\mathbf{x}) + (1-t)\mathbf{y}) \\ &\leq \alpha f(s\mathbf{x} + (1-s)\mathbf{y}) + (1-\alpha) f(t\mathbf{x} + (1-t)\mathbf{y}) = \alpha g(s) + (1-\alpha) g(t). \end{aligned}$$

Therefore, g is convex. Moreover, by Theorem 4.17,

$$g'(t) = \langle \nabla f(t\mathbf{x} + (1-t)\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Using our previous result,

$$\begin{aligned} g(0) &\geq g(1) + g'(1)(0-1) \\ f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle. \end{aligned}$$

$\impliedby$ Assume that $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ for all $\mathbf{x}, \mathbf{y} \in H$. For any $\alpha \in [0, 1]$, we define:

$$\mathbf{z} = \alpha \mathbf{x} + (1-\alpha)\mathbf{y}.$$

Therefore, applying the inequality,

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle = f(\mathbf{z}) + (1-\alpha) \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{y} \rangle \\ f(\mathbf{y}) &\geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle = f(\mathbf{z}) + \alpha \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{x} \rangle. \end{aligned}$$

We have:

$$\begin{aligned} \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) &\geq f(\mathbf{z}) + \alpha(1-\alpha) \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{y} \rangle + \alpha(1-\alpha) \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{x} \rangle \\ &\geq f(\mathbf{z}) = f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}). \end{aligned}$$

Therefore, f is convex.

□

Theorem G.9. (**1st order optimality condition**) If $f : H \to \mathbb{R}$ is convex and differentiable at $\mathbf{x}^* \in H$, then:

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x} \in H} f(\mathbf{x}) \iff \nabla f(\mathbf{x}^*) = \mathbf{0}.$$

Proof.

$\implies$ We have proven this in Theorem G.2.

$\impliedby$ We have $\nabla f(\mathbf{x}^*) = \mathbf{0}$ for some $\mathbf{x}^* \in H$. Since f is convex and differentiable at $\mathbf{x}^*$, by Lemma G.8,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle = f(\mathbf{x}^*), \qquad \text{for } \mathbf{x} \in H.$$

Therefore, by the 0th order optimality condition,

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x} \in H} f(\mathbf{x}).$$

□

Appendix H

Gradient Descent

This case study assumes that you have already read Chapter 4 and Appendix G.
Let H be a Hilbert space. Suppose we have a function $f : H \to \mathbb{R}$. We aim to find:

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in H} f(\mathbf{x}).$$

How can we find the solution numerically? Assume we have an estimate $\mathbf{x}^{(k)}$ of $\mathbf{x}^*$. How can we find a better estimate $\mathbf{x}^{(k+1)}$? If f is differentiable at $\mathbf{x}^{(k)}$, we can use an affine approximation:

$$f(\mathbf{x}) \approx f(\mathbf{x}^{(k)}) + \left\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle.$$

This approximation is accurate only if $\|\mathbf{x} - \mathbf{x}^{(k)}\|$ is small. Once $\mathbf{x}^{(k)} = \mathbf{x}^*$, we have $f(\mathbf{x}) \approx f(\mathbf{x}^*)$. Instead of finding $\mathbf{x}^{(k+1)}$ globally, we can focus locally:

$$\mathbf{x}^{(k+1)} = \operatorname*{argmin}_{\mathbf{x}\in H} \left(f(\mathbf{x}^{(k)}) + \left\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle \right) \text{ such that } \|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \alpha_k \|\nabla f(\mathbf{x}^{(k)})\|,$$

where $\alpha_k > 0$ is a small number. The idea is that $\|\nabla f(\mathbf{x}^{(k)})\|$ will slowly approach 0 as $\mathbf{x}^{(k)}$ gets closer to $\mathbf{x}^*$, requiring smaller steps to converge gradually with a sufficiently small α_k. This is equivalent to solving:

$$\min_{\mathbf{x}\in H} \left\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle \text{ such that } \|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \alpha_k \|\nabla f(\mathbf{x}^{(k)})\|.$$

By the Cauchy-Schwarz inequality,

$$\left\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle \geq -\|\nabla f(\mathbf{x}^{(k)})\| \|\mathbf{x} - \mathbf{x}^{(k)}\| \geq -\alpha_k \|\nabla f(\mathbf{x}^{(k)})\|^2.$$

To attain equality, we require:

$$\mathbf{x} - \mathbf{x}^{(k)} = -\alpha_k \nabla f(\mathbf{x}^{(k)}).$$

Therefore, we can rewrite the formula as:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla f(\mathbf{x}^{(k)}).$$

We have the following algorithm:

Definition H.1. The **Gradient Descent algorithm** is a method to minimize a differentiable function $f : H \to \mathbb{R}$. The steps of the algorithm are as follows:

0: Initialize $\mathbf{x}^{(0)}$.

1: For $k = 0, 1, \cdots$,

1. Choose the step size α_k.
2. Update $\mathbf{x}^{(k+1)}$ by:
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla f(\mathbf{x}^{(k)}).$$

Repeat until convergence is achieved.

If it converges, then $\nabla f(\mathbf{x}^{(\infty)}) = \mathbf{0}$. From Appendix G,

1. If f is convex, then $\mathbf{x}^{(\infty)}$ is the solution.
2. If f is non-convex, then $\mathbf{x}^{(\infty)}$ may not be a global minimizer.

We can focus on a prominent example: optimization in least squares:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2, \tag{LS}$$

where $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\mathbf{b}\in\mathbb{R}^m$ are given. This problem often arises in regressions (Appendix D), such as:

1. Linear regression: For given $\mathbf{X}\in\mathbb{R}^{N\times(n+1)}$ and $\mathbf{y}\in\mathbb{R}^N$,
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{n+1}} \|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|_2^2.$$
2. Ridge regression: For given $\mathbf{X}\in\mathbb{R}^{N\times(n+1)}$, $\mathbf{y}\in\mathbb{R}^N$, and $\lambda>0$,
$$\min_{\boldsymbol{\beta}=\begin{pmatrix}\mathbf{a}\\ b\end{pmatrix}\in\mathbb{R}^{n+1}} \left(\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|_2^2+\lambda\|\mathbf{a}\|_2^2\right).$$

Let $f(\mathbf{x})=\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ for $\mathbf{x}\in\mathbb{R}^n$. For any $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ and $\alpha\in[0,1]$,

$$\begin{aligned} f(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) &= \frac{1}{2}\|\mathbf{A}(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})-\mathbf{b}\|_2^2 \\ &= \frac{1}{2}\|\alpha(\mathbf{A}\mathbf{x}-\mathbf{b})+(1-\alpha)(\mathbf{A}\mathbf{y}-\mathbf{b})\|_2^2 \\ &\le \frac{\alpha}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2+\frac{1-\alpha}{2}\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2^2 = \alpha f(\mathbf{x})+(1-\alpha)f(\mathbf{y}). \end{aligned}$$

Therefore, f is convex. Before proving that f is differentiable, we must first prove the following theorem.

Theorem H.2. Let $g:\mathbb{R}^m\to\mathbb{R}$ be a differentiable function, and let $\mathbf{A}\in\mathbb{R}^{m\times n}$. Define $h:\mathbb{R}^n\to\mathbb{R}$ by:

$$h(\mathbf{x})=g(\mathbf{A}\mathbf{x}), \qquad \text{for } \mathbf{x}\in\mathbb{R}^n.$$

Then, for any $\mathbf{x}\in\mathbb{R}^n$:

$$\nabla h(\mathbf{x})=\mathbf{A}^T\nabla g(\mathbf{A}\mathbf{x}).$$

Proof.
We have $\|\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{x}\|_2\le\|\mathbf{A}\|_2\|\mathbf{y}-\mathbf{x}\|_2$. As $\|\mathbf{y}-\mathbf{x}\|_2\to 0$, by the Squeeze Theorem, $\|\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{x}\|_2\to 0$. Therefore,

$$\begin{aligned} 0 &\le \lim_{\|\mathbf{y}-\mathbf{x}\|_2\to 0}\frac{\left|h(\mathbf{y})-(h(\mathbf{x})+\langle\mathbf{A}^T\nabla g(\mathbf{A}\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle)\right|}{\|\mathbf{y}-\mathbf{x}\|_2} \\ &\le \lim_{\|\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{x}\|_2\to 0}\frac{\left|g(\mathbf{A}\mathbf{y})-(g(\mathbf{A}\mathbf{x})+\langle\mathbf{A}^T\nabla g(\mathbf{A}\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle)\right|}{\|\mathbf{y}-\mathbf{x}\|_2} \\ &\le \|\mathbf{A}\|_2\lim_{\|\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{x}\|_2\to 0}\frac{\left|g(\mathbf{A}\mathbf{y})-(g(\mathbf{A}\mathbf{x})+\langle\mathbf{A}^T\nabla g(\mathbf{A}\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle)\right|}{\|\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{x}\|_2}=0. \end{aligned}$$

Thus, by definition, $\nabla h(\mathbf{x})=\mathbf{A}^T\nabla g(\mathbf{A}\mathbf{x})$. □

Based on this theorem, let $g(\mathbf{y})=\frac{1}{2}\|\mathbf{y}-\mathbf{b}\|_2^2$ for any $\mathbf{y}\in\mathbb{R}^m$. We have:

$$\nabla f(\mathbf{x})=\mathbf{A}^T\nabla g(\mathbf{A}\mathbf{x})=\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b}).$$

Thus, f is differentiable. By the 1st order optimality condition,

$$\mathbf{x}^*=\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n}\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \iff \nabla f(\mathbf{x}^*)=\mathbf{0} \iff \mathbf{A}^T\mathbf{A}\mathbf{x}=\mathbf{A}^T\mathbf{b}.$$

If we apply the gradient descent algorithm to least squares, for $k = 0, 1, 2, \cdots$,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}).$$

How do we choose the step sizes α_k? By substituting the gradient descent formula into f, we get:

$$\alpha_k = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{A}(\mathbf{x}^{(k)} - \alpha \mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})) - \mathbf{b}\|_2^2,$$

which finds the optimal step size such that f is minimized in the next step of the gradient descent. Denote:

$$\begin{aligned}
g(\alpha) &= \frac{1}{2}\|\mathbf{A}(\mathbf{x}^{(k)} - \alpha \mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})) - \mathbf{b}\|_2^2 \\
&= \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}\|_2^2 - \alpha(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})^T\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}) + \frac{\alpha^2}{2}\|\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2 \\
&= \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}\|_2^2 - \alpha\|\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2 + \frac{\alpha^2}{2}\|\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2, \\
g'(\alpha) &= -\|\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2 + \alpha\|\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2.
\end{aligned}$$

By substituting $g'(\alpha_k) = 0$,

$$\begin{aligned}
0 = g'(\alpha_k) &= -\|\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2 + \alpha_k\|\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2, \\
\alpha_k &= \frac{\|\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2}{\|\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})\|_2^2}.
\end{aligned}$$

We have the following algorithm:

Definition H.3. The **Steepest Descent algorithm** is a method to find the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ for given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. The steps of the algorithm are as follows:

0: Initialize $\mathbf{x}^{(0)}$.

1: For $k = 0, 1, \cdots$,

1. Compute $\mathbf{g}^{(k)}$ by:
$$\mathbf{g}^{(k)} = \mathbf{A}^T(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}).$$
2. Compute α_k by:
$$\alpha_k = \frac{\left\|\mathbf{g}^{(k)}\right\|_2^2}{\left\|\mathbf{A}\mathbf{g}^{(k)}\right\|_2^2}.$$
3. Update $\mathbf{x}^{(k+1)}$ by:
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \mathbf{g}^{(k)}.$$

Repeat until convergence is achieved.

Appendix I

Newton's Method

This case study assumes that you have already read Chapter 5 and Appendix H.
Let H be a Hilbert space. Suppose we have a function $f : H \to \mathbb{R}$. We aim to find the solution of:

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in H} f(\mathbf{x}).$$

Given an estimate $\mathbf{x}^{(k)}$ of $\mathbf{x}^*$, instead of using an affine approximation as in Appendix H, we can use the second-order Taylor expansion:

$$f(\mathbf{x}) \approx f(\mathbf{x}^{(k)}) + \left\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle$$

if f and ∇f are differentiable at $\mathbf{x} \in H$ and $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|$ is small. We may denote $F_k : H \to \mathbb{R}$ as:

$$F_k(\mathbf{x}) = f(\mathbf{x}^{(k)}) + \left\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle.$$

Therefore, by the first-order optimality condition, our problem can be rewritten as:

$$\mathbf{x}^{(k+1)} = \operatorname*{argmin}_{\mathbf{x}\in H} F_k(\mathbf{x}) \iff \nabla F_k(\mathbf{x}^{(k+1)}) = \mathbf{0}.$$

Since $\nabla^2 f(\mathbf{x})$ is linear and self-adjoint,

$$\begin{aligned}
\nabla F_k(\mathbf{x}) &= \nabla f(\mathbf{x}^{(k)}) + \frac{1}{2}((\nabla^2 f(\mathbf{x}^{(k)}))^*(\mathbf{x} - \mathbf{x}^{(k)}) + \nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)})) \\
&= \nabla f(\mathbf{x}^{(k)}) + \nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}), \\
\mathbf{0} = \nabla F_k(\mathbf{x}^{(k+1)}) &= \nabla f(\mathbf{x}^{(k)}) + \nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}), \\
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - (\nabla^2 f(\mathbf{x}^{(k)}))^{-1}\nabla f(\mathbf{x}^{(k)}).
\end{aligned}$$

Therefore, we have the following method:

Definition I.1. The **Newton's method** is a technique to minimize a twice-differentiable function $f : H \to \mathbb{R}$. The steps of the algorithm are as follows:

0: Initialize $\mathbf{x}^{(0)}$.

1: For $k = 0, 1, \cdots$, update $\mathbf{x}^{(k+1)}$ by:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (\nabla^2 f(\mathbf{x}^{(k)}))^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Repeat until convergence is achieved.

Remark I.1.1. Newton's method usually requires fewer iterations than gradient descent if it converges, because the second-order approximation F_k is more accurate than the first-order approximation used in gradient descent.

There are some main concerns in Newton's method:

1. It converges to $\mathbf{x}^*$ only if $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|$ is sufficiently small. We can add a step size to the algorithm, similar to gradient descent.

Definition I.2. The **Newton's method with a step size** is a technique to minimize a twice-differentiable function $f : H \to \mathbb{R}$. The steps of the algorithm are as follows:

0: Initialize $\mathbf{x}^{(0)}$.

1: For $k = 0, 1, \cdots$,

1. Choose the step size α_k.
2. Update $\mathbf{x}^{(k+1)}$ by:
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k(\nabla^2 f(\mathbf{x}^{(k)}))^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Repeat until convergence is achieved.

We can also modify the problem as follows:

$$\mathbf{x}^{(k+1)} = \operatorname*{argmin}_{\mathbf{x}\in H} F_k(\mathbf{x}) \text{ such that } \|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \alpha_k \|\nabla f(\mathbf{x}^{(k)})\|,$$

where $\alpha_k > 0$ is the step size. The derivation is omitted, but we obtain the following algorithm:

Definition I.3. The **damped Newton's method** is a technique to minimize a twice-differentiable function $f : H \to \mathbb{R}$. The steps of the algorithm are as follows:

0: Initialize $\mathbf{x}^{(0)}$.

1: For $k = 0, 1, \cdots$,

1. Choose the regularization parameter λ_k.
2. Update $\mathbf{x}^{(k+1)}$ by:
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (\nabla^2 f(\mathbf{x}^{(k)}) + \lambda_k \mathbf{I})^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Repeat until convergence is achieved.

There are many variations of this damped Newton's method that adaptively adjust the regularization parameter, such as the trust-region algorithm and the Levenberg-Marquardt algorithm.

2. The computation of the inversion of the Hessian is too expensive. We need to approximate $\nabla^2 f(\mathbf{x}^{(k)})$ by some simple linear operators $\mathbf{B}_k$ so that finding $\mathbf{B}_k^{-1}$ is much easier than finding $(\nabla^2 f(\mathbf{x}^{(k)}))^{-1}$. Therefore,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{B}_k^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Definition I.4. The **quasi-Newton method** is a technique to minimize a twice-differentiable function $f : H \to \mathbb{R}$ without the need to find the inverse of the Hessian. The steps of the algorithm are as follows:

0: Initialize $\mathbf{x}^{(0)}$.

1: For $k = 0, 1, \cdots$,

1. Find the approximation $\mathbf{B}_k$ of $\nabla^2 f(\mathbf{x}^{(k)})$.
2. Update $\mathbf{x}^{(k+1)}$ by:
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{B}_k^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Repeat until convergence is achieved.

Remark I.4.1. If $\mathbf{B}_k$ is diagonal, then it is a scaled gradient descent.

There are many methods to find $\mathbf{B}_k$, such as:

1. Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS),
2. Davidon-Fletcher-Powell formula (DFP).

The inverse is calculated using the Sherman-Morrison formula.

Appendix J

Deep Neural Network Training in Deep Learning

This case study assumes that you have already read Chapter 5 and Appendices E and H.
Suppose that we are given N input-output pairs:

$$(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(N)}, y_N)$$

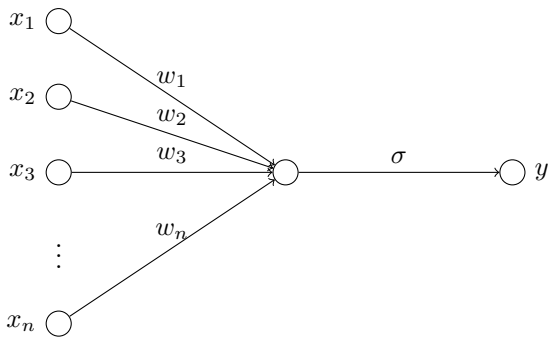
where $\mathbf{x}^{(i)} \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for $i = 1, \ldots, N$. Previously, we used different methods to find the function f that predicts the outputs:

1. Linear regression: f is an affine function.
2. Kernel ridge regression: f is a linear function in the feature domain.

In deep learning, we choose f to be a function generated by deep neural networks. For simplicity, we only consider a fully-connected neural network (FCNN). The simplest FCNN is:

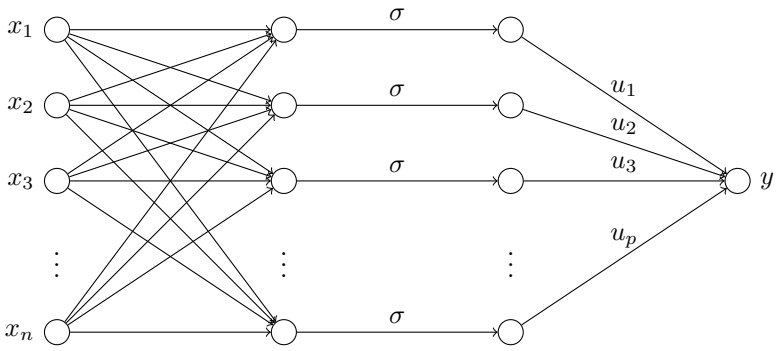
$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i\right) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$$

where $\mathbf{w} \in \mathbb{R}^n$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function.



We can have multiple sums with different sets of weights:

$$y = \sum_{i=1}^{p} u_i \sigma(\langle \mathbf{w}_i, \mathbf{x} \rangle), \qquad \text{where } \mathbf{u} \in \mathbb{R}^p.$$



This is the single-layer neural network.

We should formalize some notations. Define:

$$\mathbf{W} = \begin{pmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_p^T \end{pmatrix} \in \mathbb{R}^{p\times n}, \qquad \mathbf{u} \in \mathbb{R}^p, \qquad \sigma(\mathbf{y}) = \begin{pmatrix} \sigma(y_1) \\ \vdots \\ \sigma(y_p) \end{pmatrix}.$$

We aim to find $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{W} \in \mathbb{R}^{p\times n}$ such that:

$$f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) \approx y_i, \qquad \text{for } i = 1,\dots,N, \qquad f_{\mathbf{W},\mathbf{u}}(\mathbf{x}) = \langle \mathbf{u}, \sigma(\mathbf{W}\mathbf{x})\rangle, \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

We define $F_i(\mathbf{W},\mathbf{u}) = (f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)^2$ and $F(\mathbf{W},\mathbf{u}) = \sum_{i=1}^N F_i(\mathbf{W},\mathbf{u})$. Our problem now is to solve:

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times n}\\ \mathbf{u}\in\mathbb{R}^p}} \sum_{i=1}^N (f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)^2 \iff \min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times n}\\ \mathbf{u}\in\mathbb{R}^p}} F(\mathbf{W},\mathbf{u}) \iff \min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times n}\\ \mathbf{u}\in\mathbb{R}^p}} \sum_{i=1}^N F_i(\mathbf{W},\mathbf{u}).$$

Remark J.0.1. $F(\mathbf{W},\mathbf{u})$ is differentiable if σ is differentiable.

Remark J.0.2. $F(\mathbf{W},\mathbf{u})$ is not convex in general.

We can apply gradient descent to train the neural network, which requires $\nabla F(\mathbf{W},\mathbf{u})$. To find that, we need $\nabla_{\mathbf{W}} F(\mathbf{W},\mathbf{u})$ and $\nabla_{\mathbf{u}} F(\mathbf{W},\mathbf{u})$. We have:

$$\nabla_{\mathbf{W}} F = \sum_{i=1}^N \nabla_{\mathbf{W}} F_i, \qquad \nabla_{\mathbf{u}} F = \sum_{i=1}^N \nabla_{\mathbf{u}} F_i.$$

Remark J.0.3. $\nabla_{\mathbf{W}} F(\mathbf{W},\mathbf{u})$ means finding $\nabla F(\mathbf{W},\mathbf{u})$ with respect to $\mathbf{W}$ while keeping $\mathbf{u}$ fixed.

To find $\nabla_{\mathbf{W}} F_i(\mathbf{W},\mathbf{u})$, let:

$$G_1(t) = (t-y_i)^2, \qquad \text{for } t \in \mathbb{R}, \quad G_2(\mathbf{W}) = f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}), \qquad \text{for } \mathbf{W} \in \mathbb{R}^{p\times n}, \quad F_i(\mathbf{W},\mathbf{u}) = (G_1 \circ G_2)(\mathbf{W}).$$

By the chain rule,

$$\nabla_{\mathbf{W}} F_i(\mathbf{W},\mathbf{u}) = G_1'(G_2(\mathbf{W})) \cdot \nabla G_2(\mathbf{W}) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \cdot \nabla_{\mathbf{W}} f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}).$$

Again, let:

$$H_1(\mathbf{z}) = \langle \mathbf{u}, \mathbf{z}\rangle, \qquad \text{for } \mathbf{z} \in \mathbb{R}^p, \quad H_2(\mathbf{W}) = \sigma(\mathbf{W}\mathbf{x}^{(i)}), \qquad \text{for } \mathbf{W} \in \mathbb{R}^{p\times n}, \quad f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) = (H_1 \circ H_2)(\mathbf{W}).$$

By the chain rule,

$$\nabla_{\mathbf{W}} f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) = (D(\sigma(\mathbf{W}\mathbf{x}^{(i)})))^* \mathbf{u}.$$

It remains to find $D(\sigma(\mathbf{W}\mathbf{x}^{(i)}))$. Again, by the chain rule, for any $\mathbf{W},\mathbf{V} \in \mathbb{R}^{p\times n}$,

$$D(\sigma(\mathbf{W}\mathbf{x}^{(i)}))(\mathbf{V}) = D\sigma(\mathbf{W}\mathbf{x}^{(i)})(D(\mathbf{W}\mathbf{x}^{(i)})(\mathbf{V})) = D\sigma(\mathbf{W}\mathbf{x}^{(i)})(\mathbf{V}\mathbf{x}^{(i)}).$$

By Theorem 5.18, for any $\mathbf{z} \in \mathbb{R}^p$,

$$D\sigma(\mathbf{z}) = \begin{pmatrix} \sigma'(z_1) & 0 & \dots & 0 \\ 0 & \sigma'(z_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma'(z_p) \end{pmatrix} = \operatorname{diag}(\sigma'(\mathbf{z})), \qquad D(\sigma(\mathbf{W}\mathbf{x}^{(i)}))(\mathbf{V}) = \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^{(i)}))V\mathbf{x}^{(i)}.$$

By the definition of the adjoint operator, since $\langle \mathbf{u},\mathbf{v}\rangle = \mathbf{v}^T\mathbf{u} = \operatorname{Tr}(\mathbf{u}\mathbf{v}^T)$ for $\mathbf{u},\mathbf{v} \in \mathbb{R}^n$,

$$\begin{aligned} \left\langle D(\sigma(\mathbf{W}\mathbf{x}^{(i)}))(\mathbf{V}), \mathbf{z}\right\rangle &= \left\langle \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^*))\mathbf{V}\mathbf{x}^{(i)}, \mathbf{z}\right\rangle \\ &= \mathbf{z}^T \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^*))\mathbf{V}\mathbf{x}^{(i)} \\ &= \operatorname{Tr}(\mathbf{x}^{(i)}\mathbf{z}^T \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^*))\mathbf{V}) = \left\langle \mathbf{V}, \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^*))\mathbf{z}(\mathbf{x}^{(i)})^T\right\rangle. \end{aligned}$$

Therefore, $(D(\sigma(\mathbf{W}\mathbf{x}^{(i)}))(\mathbf{V}))^*\mathbf{u} = \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^*))\mathbf{u}(\mathbf{x}^{(i)})^T$ and thus,

$$\nabla_{\mathbf{W}} F_i(\mathbf{W},\mathbf{u}) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \cdot \nabla_{\mathbf{W}} f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) \operatorname{diag}(\sigma'(\mathbf{W}\mathbf{x}^*))\mathbf{u}(\mathbf{x}^{(i)})^T.$$

To find $\nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u})$, we can again let:

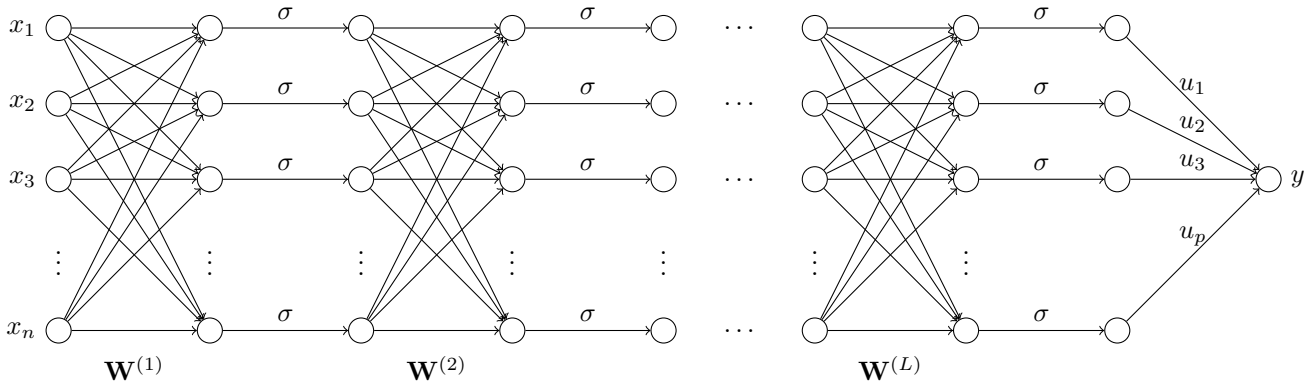
$$G_1(t) = (t - y_i)^2, \qquad \text{for } t \in \mathbb{R}, \qquad G_2(\mathbf{u}) = \left\langle \mathbf{u}, \sigma(\mathbf{W}\mathbf{x}^{(i)}) \right\rangle, \qquad \text{for } \mathbf{u} \in \mathbb{R}^p, \qquad F_i(\mathbf{W}, \mathbf{u}) = (G_1 \circ G_2)(\mathbf{u}).$$

By the chain rule,

$$\nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u}) = G_1'(G_2(\mathbf{u})) \cdot \nabla G_2(\mathbf{W}) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \cdot \nabla_{\mathbf{u}} \left(\left\langle \mathbf{u}, \sigma(\mathbf{W}\mathbf{x}^{(i)}) \right\rangle \right) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)\sigma(\mathbf{W}\mathbf{x}^{(i)})$$

We can extend the neural network into multi-layer neural network. Assume that there are L hidden layers.

$$f_{\mathbf{W},\mathbf{u}}(\mathbf{x}) = \left\langle \mathbf{u}, \sigma(\mathbf{W}^{(L)} \cdots \sigma(\mathbf{W}^{(2)} \sigma(\mathbf{W}^{(1)}))) \right\rangle = \left\langle \mathbf{u}, (\sigma \circ \mathbf{W}^{(L)} \circ \cdots \circ \sigma \circ \mathbf{W}^{(2)} \circ \sigma \circ \mathbf{W}^{(1)})(\mathbf{x}) \right\rangle$$



Let $\mathbf{W} = [\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(L)}] \in \mathbb{R}^{n_1 \times n} \times \mathbb{R}^{n_2 \times n_1} \times \cdots \times \mathbb{R}^{n_L \times n_{L-1}}$. We aim to find $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{W}$ such that:

$$f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) \approx y_i, \qquad \text{for } i = 1, \cdots, N$$

Similarly, we define $F_i(\mathbf{W}, \mathbf{u}) = (f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)^2$ and $F(\mathbf{W}, \mathbf{u}) = \sum_{i=1}^N F_i(\mathbf{W}, \mathbf{u})$. Our problem now is to solve:

$$\min_{\substack{\mathbf{W} \in \mathbb{R}^{p \times n} \\ \mathbf{u} \in \mathbb{R}^p}} \sum_{i=1}^N (f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)^2 \iff \min_{\substack{\mathbf{W} \in \mathbb{R}^{p \times n} \\ \mathbf{u} \in \mathbb{R}^p}} F(\mathbf{W}, \mathbf{u}) \iff \min_{\substack{\mathbf{W} \in \mathbb{R}^{p \times n} \\ \mathbf{u} \in \mathbb{R}^p}} \sum_{i=1}^N F_i(\mathbf{W}, \mathbf{u}).$$

Now, we need $\nabla_{\mathbf{W}^{(\ell)}} F_i(\mathbf{W}, \mathbf{u})$ for $\ell = 1, \cdots, L$ and $\nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u})$. Define:

$$\begin{aligned}
&\mathbf{v}^{(0)} = \mathbf{x}^{(i)} \qquad && \mathbf{v}^{(i)} = \sigma(\mathbf{W}^{(i)}\mathbf{v}^{(i-1)}), && \text{for } i = 1, \cdots, L \\
&\mathbf{s}^{(L)} = \sigma \qquad && \mathbf{s}^{(j)} = \sigma \circ \mathbf{W}^{(L)} \circ \cdots \circ \mathbf{W}^{(j+1)} \circ \sigma = \mathbf{s}^{(j+1)} \circ \mathbf{W}^{(j+1)} \circ \sigma, && \text{for } j = 0, \cdots, L-1
\end{aligned}$$

Therefore, we can rewrite $f_{\mathbf{W},\mathbf{u}}$ into:

$$f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) = \left\langle \mathbf{u}, \mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}) \right\rangle, \qquad \text{for } \ell = 1, \cdots, L.$$

It is similar to single-layer neural network for $\nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u})$. We have:

$$\nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u}) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \cdot \nabla_{\mathbf{u}} \left(\left\langle \mathbf{u}, \mathbf{v}^{(L)} \right\rangle \right) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)\mathbf{v}^{(L)}$$

However, for $\ell = 1, \cdots, L$, for $\mathbf{V}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$,

$$\begin{aligned}
D_{\mathbf{W}^{(\ell)}} F_i(\mathbf{W}, \mathbf{u})(\mathbf{V}^{(\ell)}) &= 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \cdot D_{\mathbf{W}^{(\ell)}} \left(\left\langle \mathbf{u}, \mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}) \right\rangle \right) (\mathbf{V}^{(\ell)}) \\
&= 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \left\langle \mathbf{u}, D(\mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}))(\mathbf{V}^{(\ell)}) \right\rangle \\
&= 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i) \left\langle \mathbf{u}, D\mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)})\mathbf{V}^{(\ell)}\mathbf{v}^{(\ell-1)} \right\rangle \\
&= \left\langle 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)(D\mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}))^T \mathbf{u}(\mathbf{v}^{(\ell-1)})^T, \mathbf{V}^{(\ell)} \right\rangle
\end{aligned}$$

Denote $\mathbf{M}^{(\ell)} = D\mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)})$. Then,

$$\nabla_{\mathbf{W}^{(\ell)}} F_i(\mathbf{W}, \mathbf{u}) = 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)(\mathbf{M}^{(\ell)})^T \mathbf{u}(\mathbf{v}^{(\ell-1)})^T$$

How do we find $\mathbf{M}^{(\ell)}$? We can use recursion. By the chain rule,

$$\begin{aligned} D\mathbf{s}^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}) &= D\mathbf{s}^{(\ell+1)}(\mathbf{W}^{(\ell+1)}\sigma(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)})) \circ D\mathbf{W}^{(\ell+1)}(\sigma(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)})) \circ D\sigma(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}) \\ &= D\mathbf{s}^{(\ell+1)}(\mathbf{W}^{(\ell+1)}\mathbf{v}^{(\ell)}) \circ D\mathbf{W}^{(\ell+1)}(\mathbf{v}^{(\ell)}) \circ D\sigma(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}), \\ \mathbf{M}^{(\ell)} &= \mathbf{M}^{(\ell+1)}\mathbf{W}^{(\ell+1)} \operatorname{diag}(\sigma'(\mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)})). \end{aligned}$$

For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, we define the element-wise multiplication as:

$$\mathbf{a} \odot \mathbf{b} = \begin{pmatrix} a_1 b_1 \\ \vdots \\ a_n b_n \end{pmatrix} = \operatorname{diag}(\mathbf{a})\mathbf{b}.$$

For simplicity, we combine some of the matrices as follows:

$$\mathbf{z}^{(\ell)} = (\mathbf{M}^{(\ell)})^T\mathbf{u}, \qquad \mathbf{p}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}.$$

Expanding $\mathbf{z}^{(\ell)}$, we have:

$$\begin{aligned} \mathbf{z}^{(\ell)} &= (\mathbf{M}^{(\ell)})^T\mathbf{u} \\ &= \operatorname{diag}(\sigma'(\mathbf{p}^{(\ell)}))(\mathbf{W}^{(\ell+1)})^T(\mathbf{M}^{(\ell+1)})^T\mathbf{u} \\ &= \sigma'(\mathbf{p}^{(\ell)}) \odot (\mathbf{W}^{(\ell+1)})^T\mathbf{z}^{(\ell+1)}, \\ \nabla_{\mathbf{W}^{(\ell)}} F_i(\mathbf{W}, \mathbf{u}) &= 2(f_{\mathbf{W},\mathbf{u}}(\mathbf{x}^{(i)}) - y_i)\mathbf{z}^{(\ell)}(\mathbf{v}^{(\ell-1)})^T. \end{aligned}$$

We now have the following gradient descent algorithm:

Algorithm 1 Gradient Descent for Deep Neural Network

Require: $\alpha_k > 0$
  $\nabla_{\mathbf{u}} F(\mathbf{W}, \mathbf{u}) \leftarrow \mathbf{0}$
  for $\ell = L-1, \ldots, 1$ **do**
    $\nabla_{\mathbf{W}^{(\ell)}} F(\mathbf{W}, \mathbf{u}) \leftarrow \mathbf{O}$
  end for
  for $i = 1, \ldots, N$ **do**
    $\mathbf{v}^{(0)} \leftarrow \mathbf{x}^{(i)}$
    for $\ell = 1, 2, \ldots, L$ **do**
      $\mathbf{p}^{(\ell)} \leftarrow \mathbf{W}^{(\ell)}\mathbf{v}^{(\ell-1)}$
      $\mathbf{v}^{(\ell)} \leftarrow \sigma(\mathbf{p}^{(\ell)})$
    end for
    $a \leftarrow \left\langle \mathbf{u}, \mathbf{v}^{(L)} \right\rangle$
    $\nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u}) \leftarrow 2(a - y_i)\mathbf{v}^{(L)}$
    $\nabla_{\mathbf{u}} F(\mathbf{W}, \mathbf{u}) \leftarrow \nabla_{\mathbf{u}} F(\mathbf{W}, \mathbf{u}) + \nabla_{\mathbf{u}} F_i(\mathbf{W}, \mathbf{u})$
    $\mathbf{z}^{(L)} \leftarrow \sigma'(\mathbf{p}^{(L)}) \odot \mathbf{u}$
    $\nabla_{\mathbf{W}^{(L)}} F_i(\mathbf{W}, \mathbf{u}) \leftarrow 2(a - y_i)\mathbf{z}^{(L)}(\mathbf{v}^{(\ell-1)})^T$
    $\nabla_{\mathbf{W}^{(L)}} F(\mathbf{W}, \mathbf{u}) \leftarrow \nabla_{\mathbf{W}^{(L)}} F(\mathbf{W}, \mathbf{u}) + \nabla_{\mathbf{W}^{(L)}} F_i(\mathbf{W}, \mathbf{u})$
    for $\ell = L-1, \ldots, 1$ **do**
      $\mathbf{z}^{(\ell)} \leftarrow \sigma'(\mathbf{p}^{(\ell)}) \odot (\mathbf{W}^{(\ell+1)})^T\mathbf{z}^{(\ell+1)}$
      $\nabla_{\mathbf{W}^{(\ell)}} F_i(\mathbf{W}, \mathbf{u}) \leftarrow 2(a - y_i)\mathbf{z}^{(\ell)}(\mathbf{v}^{(\ell-1)})^T$
      $\nabla_{\mathbf{W}^{(\ell)}} F(\mathbf{W}, \mathbf{u}) \leftarrow \nabla_{\mathbf{W}^{(\ell)}} F(\mathbf{W}, \mathbf{u}) + \nabla_{\mathbf{W}^{(\ell)}} F_i(\mathbf{W}, \mathbf{u})$
    end for
  end for
  Update the weights by gradient descent.