Mathematical Foundation of Statistical Machine Learning *

Parley Ruogu Yang†

This version: Sunday 24th July, 2022
Course date: 18 — 29 July 2022

Abstract

This note is designed to be delivered in a summer school for undergraduate students, with a total teaching duration of 18 hours, for them to gain basic understanding on analysis, optimisation, statistics, and leading to some basic use of Machine Learning. Further topics in analysis are outlined with extension to Machine Learning.

*Latest version: `https://parleyyang.github.io`

†Faculty of Mathematics, University of Cambridge

1 Syllabus

Note: due to time constraint, we need to choose either §1.4 or §1.5 to be taught: if §1.1 takes less time, i.e. that the students are comfortable with the basics of analysis, then §1.5 could be introduced without much of a difficulty, and the extra hour would be taken from §1.2 due to assumed level of familiarity on Euclidean space operations. Otherwise it is recommended to go through §1.4 instead of §1.5.

In 2022, students unanimously voted for §1.5 for their interests in infinity and their strength in basic analysis.

1.1 Real Analysis and Optimisation (4-6 hours)

Objective: we would like to re-visit the basic $\mathbb{R}^n \to \mathbb{R}$ analysis followed by a rigorous treatment on convex optimisation. This leads us to gradient methods and eventually end on a proof of convergence of GD on convex functions.

1. Convergence, divergence, limits, derivatives in univariate and multivariate settings
2. Convex optimisation
3. Gradient Methods and Gradient Descent

1.2 Statistics and motivation of Machine Learning (4-5 hours)

Objective: we start by analysing the Gaussian distribution and linear models to give a proper overview of the basic linear modelling — this then extends to the likelihood method and motivates the use of Machine Learning to solve various problems. We end on the design of loss function and followed by computing experiments.

1. Random variable, Gaussian distribution, moments
2. Linear models, likelihood methods
3. Classification problem vs regression problem: loss function design
4. Neural Networks and algorithms

1.3 Computing experiments and basic Machine Learning (4-5 hours)

Objective: we apply what we have learnt into computer programmes: we code to implement certain model and optimisation, and observe the results. Further remarks are made about the optimisation parts of Machine Learning.

1. Basics of programming
2. Implementing a model
3. Optimisations in action

4. The principle of Machine Learning

5. Penalisation

1.4 Modern Machine Learning (2-3 hours)

Objective: we explore various topics in modern machine learning — this could lead to courses such as ST456 at LSE or various Part II and Part III courses at Cambridge.

1. Neural Networks and towards Deep Learning

2. Modern topics in AI: image classification, time series forecasting, and high-dimensional statistics

1.5 More on infinity and towards Modern Machine Learning (3-5 hours)

Objective: we explore the analysis of "infinite dimensional spaces". Indeed, what is infinite dimensional? Analysis starts with the notion of infinity, but what is the ultimate infinity when we have the dimension of a space to be infinite? Machine Learning related extensions are then brought in.

1. From Euclid to Hilbert: order and disorder in infinite dimensional spaces

2. Reproducible Kernel Hilbert Space

2 Real Analysis and Optimisation

2.1 Sequential Analysis

Definition 2.1. A real sequence $\{x_n\}_{n\in\mathbb{N}}$ is convergent to $C \in \mathbb{R}$ if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N} \quad \text{such that } |x_n - C| < \varepsilon \ \forall n \geq N \tag{1}$$

We note $\lim_{n\to\infty} x_n = C$ or equivalently $x_n \underset{n\to\infty}{\longrightarrow} C$ in this case.

Example 2.2. $x_n = n^{-1}$ is convergent.

Exercise 2.3. Prove

- $x_n = n^a$ is convergent for all $a < 0$.
- $x_n = n$ is not convergent.
- $x_n = \log(n)$ is not convergent.

Definition 2.4. A real sequence $\{x_n\}_{n\in\mathbb{N}}$ is divergent to ∞ if

$$\forall C \in \mathbb{R}, \exists N \in \mathbb{N}, \quad \text{such that } x_n > C \quad \forall n > N \tag{2}$$

We note $\lim_{n\to\infty} x_n = \infty$ in this case.

Exercise 2.5. Prove

- $x_n = n$ is divergent.
- $x_n = \log(n)$ is divergent.

State the definition of $\lim_{n\to\infty} x_n = -\infty$

2.2 Analysis of real functions

Definition 2.6. Let $f : \mathbb{R} \to \mathbb{R}$. Let $y \in \mathbb{R}$. Then we write $C = \lim_{x\to y} f(x)$ if

$$\forall \varepsilon > 0, \exists \delta > 0 \quad \text{such that} \quad \forall x \in \mathbb{R}, \ |x - y| < \delta \implies |f(x) - C| < \varepsilon \tag{3}$$

Example 2.7. Let $f(x) = x$, then $\lim_{x\to y} f(x) = y \ \ \forall y \in \mathbb{R}$

Definition 2.8. Let $f : \mathbb{R} \to \mathbb{R}$. Let $C \in \mathbb{R}$. Then we write $C = \lim_{x\to\infty} f(x)$ if

$$\forall \varepsilon > 0, \exists \delta > 0 \quad \text{such that} \quad \forall x \in \mathbb{R}, \ x > \delta \implies |f(x) - C| < \varepsilon \tag{4}$$

Example 2.9. Let $f(x) = x^{-1}$, then $\lim_{x\to\infty} f(x) = 0$

Exercise 2.10.

- Let $a \in \mathbb{R}$ be a parameter. Define[1] $f(x; a) = x^a$. Prove that $\lim_{x\to\infty} f(x; a) \in \mathbb{R}$ if and only if $a \leq 0$,
- With reference to the definition of sequential divergence, state the definition of $\lim_{x\to\infty} f(x) = \infty$

[1] Remark about the notation on variable and parameter: here, x is a variable and a is a parameter, separated by the semicolon.

Definition 2.11. Let $f : \mathbb{R} \to \mathbb{R}$. Let $y \in \mathbb{R}$. We say f is differentiable at y and that $f'(y) = C$ if

$$\lim_{\delta \to 0} \frac{f(y+\delta) - f(y)}{\delta} = C \tag{5}$$

Let A be a set. If f is differentiable on all $y \in A$, then f is said to be differentiable in A. If there exists $y \in A$ such that f is not differentiable on y, then f is said to be not differentiable in A.

We define $f^{(2)}$ to be the derivative of f', and $f^{(k)}$ to be the derivative of $f^{(k-1)}$ for all $k \in \mathbb{N}_{\geq 2}$ and so on. When $f^{(k)}$ is differentiable in A, we say f is k-th differentiable.

Example 2.12. Let $f(x) = x^a$, $a > 0$, then $f'(x) = x^{a-1}$

Exercise 2.13. Let $f(x) = |x|$. Show f is not differentiable in $\mathbb{R}$.

Exercise 2.14. Fix an arbitrary $k \in \mathbb{N}$. Construct a function $f : \mathbb{R} \to \mathbb{R}$ such that $f^{(k)}$ is differentiable in $\mathbb{R}$ but $f^{(k+1)}$ is not.

Definition 2.15. Let $k \in \mathbb{R}$. $\mathbb{1}$ denotes an indicator function, that is

$$\mathbb{1}[x > k] = \begin{cases} 1 & \text{if } x > k \\ 0 & \text{else} \end{cases} \tag{6}$$

Exercise 2.16. Prove that, for any $k \in \mathbb{R}$, $\mathbb{1}[x > k]$ is not differentiable in $\mathbb{R}$.

Theorem 2.1. *Consider differentiable functions $f, g : \mathbb{R} \to \mathbb{R}$ and let $h : \mathbb{R} \to \mathbb{R}$ be defined by $x \mapsto g(f(x))$. Then*

$$h'(x) = f'(x)g'(f(x)) \tag{7}$$

Exercise 2.17. Consider[2]

$$f(x; \mu, \sigma) = (\sigma\sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{8}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are parameters.

- Show that $f'(x) = \left(\frac{\mu - x}{\sigma^2}\right) f(x)$
- Hence show that $f^{(k)}$ is differentiable for all k

Definition 2.18. Let $F : \mathbb{R} \to \mathbb{R}$ be twice differentiable. Then

$$F(x) = \int_{-\infty}^{x} F'(t)dt \tag{9}$$

We write $\int_{-\infty}^{\infty} F'(t)dt$ as $\lim_{x\to\infty} F(x)$. It is also common to write this object as $\int_{\mathbb{R}} F'(t)dt$.

Note: while the above definition is true, there is a more general definition of integration — it can be even more interesting to study that under a historical context! That is, however, not further elaborated as we need to move on to optimisation.

[2]This is a very interesting function — we will come back to that again in Statistics

Exercise 2.19. Let $a \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$. Prove

$$\int_{\mathbb{R}} xf(x)dx \geq a \int_{a}^{\infty} f(x)dx \tag{10}$$

Definition 2.20. Let $f : \mathbb{R}^n \to \mathbb{R}$. Annotate $x := (x_1, ..., x_n) \in \mathbb{R}$, and e_m as the m-th canonical basis vector in $\mathbb{R}^n$, namely $e_m = (0, ..., 0, 1, 0, ..., 0) \in \mathbb{R}^n$ where the m-th position is 1 and other positions are 0. Then, the partial derivative $\partial_m f(x)$ is defined as

$$\partial_m f(x) := \lim_{\delta \to 0} \frac{f(x + \delta e_m) - f(y)}{\delta} \tag{11}$$

The gradient $\nabla f(x)$ is defined as

$$\nabla f(x) := (\partial_1 f(x), ..., \partial_n f(x)) \tag{12}$$

Remark on the notation: the meaning of $\partial_x f(x, y, ...)$ as $\partial_1 f(x, y, ...)$ and various historical contexts on the messiness. [3]

Example 2.21. Let $f(x, y, z) = xyz$, then $\nabla f(x, y, z) = (yz, xz, xy)$

Exercise 2.22. Consider $f : \mathbb{R}^2 \to \mathbb{R}$ where

$$f(\mu, \sigma; x) = (\sigma\sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right) \tag{13}$$

- Compute $\nabla \log(f(\mu, \sigma))$
- Try computing $\nabla f(\mu, \sigma)$ and discuss the value of taking logarithm.

2.3 Introduction to Optimisation

Definition 2.23.

- Let $X \subset \mathbb{R}$, then $\min(X) = c$ if $c \leq x \ \ \forall x \in X$ and that $c \in X$.
- Let $f : A \to \mathbb{R}$, then $\min(f(A)) = c$ if $c \leq f(x) \ \ \forall x \in A$ and that $\exists y \in A$ such that $f(y) = c$. In addition, $\text{argmin}(f) = \{x \in A : f(x) = c\}$.

Example 2.24. Let $f(x) = x^2$, then $\min(f) = 0$ and $\text{argmin}(f) = \{0\}$

Exercise 2.25.

- State the definition of $\max(X)$ for $X \subset \mathbb{R}$ and state the definition of $\max(f(A)) = c$ and $\text{argmax}(f)$ for $f : A \to \mathbb{R}$
- Prove that
$$\text{argmin}(f) = S \iff \text{argmax}(-f) = S$$
- Let $f : A \to (0, \infty)$. Prove that
$$\text{argmin}(f) = S \iff \text{argmin}(\log(f)) = S$$

Exercise 2.26. Reflect the discussion of Exercise 2.22.

Example 2.27. Let $f(x)$ take the formulation as Equation 8. Then $\text{argmin}(f) = \{\mu\}$ and $\min(f) = (\sigma\sqrt{2\pi})^{-1}$

[3] `https://en.wikipedia.org/wiki/Notation_for_differentiation`

2.4 The $(\mathbb{R}^n, ||\cdot||_p)$ space

Definition 2.28. Let $x \in \mathbb{R}^n$. Write $x = (x_1, x_2, ..., x_n)$. Define $||\cdot||_p : \mathbb{R}^n \to [0, \infty)$ by

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

Exercise 2.29.

- Discuss the possible notion of $||\cdot||_\infty$
- Prove the triangle inequality: $||x + y||_p \leq ||x||_p + ||y||_p \ \ \forall x, y$
- Let $f, g : A \to \mathbb{R}$ where $g(x) := ||f(x)||_p$. Consider the following statement:
$$\text{argmin}(g) = S \iff \text{argmin}(f) = S$$
For what kind of (A, p) would the statement be true?

Lemma 2.2 (Parallelogram Law)**.** *For any $v, w \in \mathbb{R}^n$,*

$$2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2 \tag{14}$$

Definition 2.30.

- A ball in $(\mathbb{R}^n, ||\cdot||_p)$ is defined as $B(x; r) := \{y \in \mathbb{R}^n : ||x - y||_p \leq r\}$
- A local minimiser for $f : \mathbb{R}^n \to \mathbb{R}$ is $x \in \mathbb{R}^n$ if $\exists r > 0$ such that
$$\forall y \in B(x; r), \quad f(y) \geq f(x) \tag{15}$$
- A global minimiser for $f : \mathbb{R}^n \to \mathbb{R}$ is x if $x \in \text{argmin}(f)$

Exercise 2.31.

- State the definition of local maximiser and global maximiser.
- Draw a ball in $(\mathbb{R}^3, ||\cdot||_2)$, $(\mathbb{R}^2, ||\cdot||_1)$, and $(\mathbb{R}^2, ||\cdot||_\infty)$.
- Define a function where all local minimsers are global minimisers.
- Define a function where some but not all local minimsers are global minimsers.
- Define a function where there are no global minimiser.

2.5 Optimisation in $\mathbb{R}^d \to \mathbb{R}$

Definition 2.32. A set A is convex if $\forall x, y \in A, t \in [0, 1]$, we have $tx + (1-t)y \in A$.

A function $f : A \to \mathbb{R}$ is convex if A is a convex set and

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) \ \ \forall x, y \in A \text{ and } \lambda \in [0, 1] \tag{16}$$

Lemma 2.3 (First Order Condition of Convexity)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be once differentiable. Then, f is convex if and only if*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \mathbb{R}^n \tag{17}$$

Theorem 2.4 (Gradient Property of optimisation)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, $x \in \mathbb{R}^d$.*

If $\nabla f(x) = 0$, then x is a global minimiser.

Exercise 2.33.

- Use Lemma 2.3 to prove Theorem 2.4.
- Characterise the gradient property for a maximiser using Theorem 2.4.
- Give an example where Theorem 2.4 applies.
- Give an example where Theorem 2.4 does not apply.
- At the end of Theorem 2.4, could we say x is the unique global minimiser?

2.6 Gradient Methods and Gradient Descent

In the last part, we review the bread-and-butter framework for Machine Learning: Gradient Descent (GD). We take $p = 2$ for the norm, and consider Lipschitz-continuous functions:

Definition 2.34. Let $B > 0$ be a constant. $f : \mathbb{R}^d \to \mathbb{R}$ is B-Lipschitz continuous if

$$|f(x) - f(y)| \leq B||x - y|| \quad \forall x, y \in \mathbb{R}^d$$

Example 2.35.

- $f : [0, +\infty)^d \to \mathbb{R}$ defined by $x \mapsto \sqrt{\prod_{i \in [d]} x_i}$ is not Lipschitz continuous.
- $|| \cdot ||_p$ is Lipschitz continuous for all $p \geq 1$.

Lemma 2.5. *Let f be once differentiable. f is B-Lipschitz if and only if*

$$||\nabla f(x)|| \leq B \quad \forall x \in \mathbb{R}^d$$

For what follows, we consider the problem of finding $\text{argmin}(f)$ where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and once differentiable. We assume the $\text{argmin}(f)$ is a singleton and hence write x^* to be the element in $\text{argmin}(f)$.

Definition 2.36. A Gradient Descent Algorithm is made of an initial point $x_0 \in \mathbb{R}^d$ and a step size $\gamma > 0$. The sequence $(x_n)_{n \in \mathbb{N}}$ follows

$$x_{t+1} = x_t - \gamma \nabla f(x_t) \quad \forall t \in \mathbb{N} \tag{18}$$

Theorem 2.6 (Gradient Descent for Lipschitz convex functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and B-Lipschitz. Suppose $||x_0 - x^*|| \leq R$ and $\gamma := R(B\sqrt{T})^{-1}$, then*

$$T^{-1} \sum_{t \in [T]} f(x_t) - f(x^*) \leq RBT^{-\frac{1}{2}} \tag{19}$$

Exercise 2.37.

- Why did we not say $x_t \underset{t\to\infty}{\longrightarrow} x^*$?
- Consider $f : \mathbb{R}^2 \to \mathbb{R}$ as per defined in Exercise 2.22: is Theorem 2.6 applicable?
- Prove the Theorem 2.6 using Lemma 2.5, Lemma 2.2, and Lemma 2.3.
- Show the upper bound in e Theorem 2.6 with a general $\gamma > 0$, and hence discuss the appropriateness of the chosen γ.

3 Statistics and motivation of Machine Learning

3.1 Gaussian distribution

Definition 3.1. A random variable $X \in \mathbb{R}$ has a probability density function (pdf) $\phi : \mathbb{R} \to [0, \infty)$ and cumulative density function (cdf) $\Phi : \mathbb{R} \to [0, 1]$ if $\forall x, y \in \mathbb{R}$,

$$\Phi(x) = \mathbb{P}[X \leq x] \tag{20}$$

$$\int_x^y \phi(t)dt = \mathbb{P}[x \leq X \leq y] \tag{21}$$

Definition 3.2. A real-valued random variable X follows Gaussian distribution with mean μ and standard deviation σ if the probability density function of X is defined by

$$f(x; \mu, \sigma) = (\sigma\sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{22}$$

We annotate $X \sim N(\mu, \sigma^2)$.

Definition 3.3. Let $f : \mathbb{R} \to \mathbb{R}$ and X a real-valued random variable, define the expectation

$$\mathbb{E}[f(X)] := \int_{x \in \mathbb{R}} f(x)xdx \tag{23}$$

and variance

$$\mathbb{V}[f(X)] := \mathbb{E}[f(X)^2] - (\mathbb{E}[f(X)])^2 \tag{24}$$

Exercise 3.4. Let $X \sim N(\mu, \sigma^2)$, $k \neq 0$, what is the distribution of kX?

Before we embark on statistical models, we need to understand higher dimensional distributions and the crucial notion of independence.

Definition 3.5. A random variable $X = (X_1, ..., X_n) \in \mathbb{R}^n$ has a probability density function (pdf) $\phi : \mathbb{R}^n \to [0, \infty)$ and cumulative density function (cdf) $\Phi : \mathbb{R}^n \to [0, 1]$ if $\forall x = (x_1, ..., x_n) \in \mathbb{R}^n, A \subset \mathbb{R}^n$,

$$\Phi(x) = \mathbb{P}[X_1 \leq x_1, ..., X_n \leq x_n] \tag{25}$$

$$\int_{t \in A} \phi(t)dt = \mathbb{P}[X \in A] \tag{26}$$

Definition 3.6. A pair of random variables (X, Y) is independent if $\phi_{X,Y}(x, y) = \phi_X(x)\phi_Y(y)$ where $\phi_{X,Y}$ is the pdf of the random variable (X, Y) and ϕ_X, ϕ_Y are the pdf of X and Y respectively.

Definition 3.7. Let (X, Y) be a pair of random variables. The covariance is defined as

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Lemma 3.1. *If (X, Y) are independent, then $Cov(X, Y) = 0$*

Definition 3.8. Let $X \in \mathbb{R}^d$. X follows Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and standard deviation $\Sigma \in \mathbb{R}^{d\times d}$ if the probability density function of X is defined by

$$f(x;\mu,\sigma) = (|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{d}{2}})^{-1} \exp\left(-\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right) \tag{27}$$

We annotate $X \sim N(\mu, \Sigma)$ in this case.

Lemma 3.2. *Using the above notations, and for any $m \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$, we have $m + AX \sim N(m + \mu, A\Sigma A^T)$*

Warning: to prove the above lemma, you will need to use some linear algebra, which we don't have time to cover.

Lemma 3.3. *If X is Gaussian and that $X = (X_1, X_2)$ with $Cov(X_1, X_2) = 0$, then X_1 and X_2 are independent.*

Exercise 3.9. Let $X = (X_1, ..., X_n) \in \mathbb{R}^n$ and let $X_i \sim N(\mu, \sigma) \ \ \forall i$.

- Write down the distribution of X.
- Induce the distribution of $\sum_{i\in[n]} X_i$

Definition 3.10. Consider random variables $X = (X_1, ..., X_n) \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Fix an arbitrary constant $y \in \mathbb{R}^m$. The conditional variable $X|Y = y$ has a probability density function (pdf) $\phi : \mathbb{R}^n \to [0, \infty)$ and cumulative density function (cdf) $\Phi : \mathbb{R}^n \to [0, 1]$ if $\forall x = (x_1, ..., x_n) \in \mathbb{R}^n, A \subset \mathbb{R}^n$,

$$\Phi(x) = \mathbb{P}[X_1 \le x_1, ..., X_n \le x_n | Y = y] \tag{28}$$

$$\int_{t\in A} \phi(t)dt = \mathbb{P}[X \in A | Y = y] \tag{29}$$

Exercise 3.11. Let p_X, p_Y, $p_{X,Y}$, $p_{X|Y=y}$ be the densities of X and Y, joint density of (X, Y) and conditional density of $X|Y = y$ respectively.

- $p_{X,Y}(x, y) = p_Y(y) \times p_{X|Y=y}(x) \ \ \forall x, y$
- If X, Y are independent, then $p_{X|Y=y}(x) = p_X(x) \ \ \forall x, y$
- If $Y = f(X) + \varepsilon$ and $\varepsilon \sim N(0, \sigma^2)$ is independent of X, then $Y|X \sim N(f(X), \sigma^2)$
- If $Y \in \{0, 1\}$, and denote $q := p_Y(0)$, then $p_{Y|X=x}(y) = f(x)(qp_{X|Y=0}(x) + (1-q)p_{X|Y=1}(x))$. Prove this and state what is $f(x)$.

We are now in a good shape to talk about data.

3.2 Linear models

We are interested in analysing data composed of observations paired in the following way: $(x_1, y_1), ..., (x_n, y_n)$ where $x_j \in \mathbb{R}^m$ and $y_j \in \mathbb{R}$ for all $j \in [n]$. In this subsection, we assume n to be large, and that we don't consider small n (for instance, if we require at least $n \ge 3$ somewhere, we will do so without saying).

In linear model, we assume $y_i = x_i^T\beta + \varepsilon_i \ \ \forall i$ and that ε_i (usually known as the noise) is drawn identically and independently (iid) from $N(0, \sigma^2)$ where σ^2 is unknown. This is written as $\varepsilon_i \sim iidN(0, \sigma^2)$

Exercise 3.12. Discuss what would happen if the noise is not drawn iid.

Definition 3.13. In the ordinary least square (OLS) framework, we are interested in minimising the sum of squared residuals from a fitted model, in particular, minimise

$$f(\beta) := \sum_{i\in[n]} (y_i - x_i^T\beta)^2 \tag{30}$$

Exercise 3.14.

- Let $m = 1$ and $x_j = 1\ \forall j$. Find argmin(f).
- Let $m = 2$ and $x_j = (1, z_j)\ \forall j$. Find argmin$(f)$.

To achieve notational ease for larger m, we introduce the design matrix X defined as

$$X := \begin{pmatrix} x_1^T \\ x_2^T \\ ... \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n\times m} \tag{31}$$

Then note $Y := (y_1, ..., y_n) \in \mathbb{R}^n$ and we can re-write Equation 30 as

$$f(\beta) := ||Y - X\beta||_2^2 \tag{32}$$

Theorem 3.4 (OLS solution)**.** *Suppose* $Rank(X) = m \leq n$*, then* argmin$(f) = \{\hat{\beta}^{OLS}\}$ *where*

$$\hat{\beta}^{OLS} = (X^TX)^{-1}X^TY \tag{33}$$

Corollary 3.5. $\hat{\beta}^{OLS}|X \sim N(\beta, \sigma^2(X^TX)^{-1})$

3.3 Likelihood methods

Another statistical estimation method is known as the maximum likelihood estimation (MLE) method. Let p_i be the probability density function of data $y_i|x_i$, which is $N(x_i^T\beta, \sigma^2)$, then the joint density is $\prod_{i\in[n]} p_i$. We denote such an object as the likelihood of our data, noted $l(\beta, \sigma^2; X, Y) = l(\theta; D)$

Definition 3.15. Let $\mathcal{L} : \Theta \to \mathbb{R}$ be the likelihood function. The MLE concerns with maximising $\mathcal{L}(\theta; D)$. We denote $\hat{\theta}^{MLE}$ as the maximiser.[4]

Exercise 3.16. Why are we concerned with $\mathcal{L}(\theta; D)$ instead of $\mathcal{L}(\theta|D)$? Discuss.

Exercise 3.17. Show that for linear model,

$$\log(\mathcal{L}(\theta; D)) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i\in[n]}(y_i - x_i^T\beta)^2$$

And hence prove that $\hat{\theta}^{MLE} = (\hat{\beta}^{MLE}, \hat{\sigma}^{2,MLE})$ where $\hat{\beta}^{MLE} = \hat{\beta}^{OLS}$

[4]It is common that this is the unique maximiser.

3.4 Classification Problem and Motivation of Neural Networks

In subsection 3.2, we investigated the problem wherein we fit a function $y = x^T\beta$ amongst the data to minimise the squared residuals. Here, we consider classification problem where $y \in Y$ and $|Y| = k \in \mathbb{N}_{\geq 2}$.

In a binary classification model, we have data $\{(x_i, y_i)\}_{i\in[n]}$ and $y_i \in \{-1, 1\}$. [5] A loss function takes the predicted value $\hat{y}$ against the true value as observed in data noted y and outputs a real value.

Example 3.18 (Mean Squared Error)**.** $l(y, \hat{y}) = (y - \hat{y})^2$

Example 3.19 (Classification Problem: correct count loss)**.**

$$l : \{-1, 1\} \times \{-1, 1\} \to \{0, 1\}$$

defined by $l(y, \hat{y}) = \max(-y\hat{y}, 0)$

Exercise 3.20. Discuss the intuition behind the correct count loss.

Exercise 3.21. Discuss the problem of linear regression in light of classification problem.

Definition 3.22. A threshold function $\rho : \mathbb{R} \to \mathbb{R}$ is defined as $\rho(x) = \mathbb{1}[x \geq 0]$

Definition 3.23 (XOR Problem)**.** Let there be data $(x_1, y_1), ..., (x_4, y_4)$ where $x_i \in \mathbb{R}^2, y_i \in \mathbb{R}$ for all $i \in [4]$. In particular:

$$\begin{aligned}(x_1, y_1) =&((-1, -1), -1)\\ (x_2, y_2) =&((-1, 1), 1)\\ (x_3, y_3) =&((1, -1), 1)\\ (x_4, y_4) =&((1, 1), -1)\end{aligned}$$

Let f be a prediction function where $f : \mathbb{R}^2 \to \mathbb{R}$, we use a threshold function to harmonise the prediction, in particular, we consider the loss of this prediction function as

$$L(f) = \sum_{i\in[4]} \max(-y_i g(f(x_i)), 0) \tag{34}$$

$$g(f(x_i)) = \mathbb{1}[f(x_i) \geq 0] - \mathbb{1}[f(x_i) < 0] \ \forall i \tag{35}$$

Exercise 3.24. Discuss the rationale behind Equation 34 and Equation 35.

Theorem 3.6 (Failure of linear functions in classification problems)**.** *Let* $\mathcal{F}^{Linear} = \{f(x) = x^T\beta : \beta \in \mathbb{R}^2\}$*, then*

$$\min_{f\in\mathcal{F}^{Linear}} L(f) = 1$$

Theorem 3.7 (Success of NN in classification problems)**.** *Consider a set of two-layer Neural Networks (NN). Let*

$$N(\rho) = \{f(x) = \rho(w_1 x + b_1)^T w_2 + b_2 : w_1 \in \mathbb{R}^{2\times 2}, w_2, b_1 \in \mathbb{R}^2, b_2 \in \mathbb{R}\} \tag{36}$$

Then

$$\min_{f\in N(\rho)} L(f) = 0 \tag{37}$$

[5] It is more common to have $y_i \in \{0, 1\}$. We use -1 here for ease of the loss function construction.

4 Computing experiments and basic Machine Learning

4.1 The basics of programming: an Object-oriented programming (OOP) viewpoint

Object-oriented programming (OOP) is a programming paradigm based on the concept of "objects", which can contain data and code: data in the form of attributes or properties, and code, in the form of methods. The following is an example of OOP

```
class Book:
  def __init__(self, title, quantity, author, price):
    self.title = title
    self.quantity = quantity
    self.author = author
    self.price = price

  def __repr__(self):
     return f"Book: {self.title}, Quantity: {self.quantity}, \
      Author: {self.author}, Price: {self.price}"
```

Exercise 4.1. What would be the printed output if we have code

```
book1 = Book('Book 1', 12, 'Author 1', 120)
print(book1)
```

Exercise 4.2. Conceptualise an OLS estimation procedure (Equation 33) using OOP in Python. You can use standard packages such as Pandas, or just conceptualise using pseudo code.

In what follows, we proceed with Python programming on Google Colab and / or equivalent notebooks to code for exercises. H stands for computation by hand and C stands for coding exercise.

Exercise 4.3.

- (H) Consider data $(1,2),(2,4),(3,4),(4,5),(5,6)$ where the first entry is the value of x and the second entry is the value of y. Design the OLS regression with constant and compute the OLS estimator.
- (C) Input the data, and then write a function to compute the OLS estimator.
- (C) Use `statsmodels.regression` package to estimate the OLS estimator. [6]

[6] `https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html`

4.2 Optimisation in action

Exercise 4.4.

- (H) Recall Definition 2.36 and various exercises before that. Design a GD method for OLS estimation with large n and $m = 2$.
- (C) Code your function. Consider thoroughly how you would like to stop the for loop.
- (C) Use the data we had in Exercise 4.3 and run your function. Discuss the result.

4.3 The principle of Machine Learning

We continue the set up from subsection 3.2 and subsection 3.4. A more generalised notion of the purpose of Machine Learning for Statistical purposes is to reduce Risks.

Definition 4.5. Consider an input space $\mathbb{X}$ and an output space $\mathbb{Y}$, often $\mathbb{Y} = \mathbb{R}$. We denote the random variables X and Y as the input and output of a decision function[7] $h : \mathbb{X} \to \mathbb{Y}$. With a loss function $l : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$, we define the risks of a decision function as

$$R(h) = \mathbb{E}[l(h(X), Y)] \tag{38}$$

Let H be a set of functions mapping $\mathbb{X} \to \mathbb{Y}$, the principle of Machine Learning is to minimise the risks, that is, to find $\text{argmin}_{h \in H} R(h)$

Exercise 4.6. Let $\mathbb{Y} = \mathbb{R}$.

- When $l(y_0, y_1) = (y_0 - y_1)^2$, show that $\text{argmin}_{h \in H} R(h) = \text{argmin}_{h \in H} \mathbb{E}[(\mathbb{E}[Y|X] - g(X))^2]$ and hence if the function $h^{reg}(x) = \mathbb{E}[Y|X = x] \in H$, then $\text{argmin}_{h \in H} R(h) = h^{reg}$
- When $l(y_0, y_1) = |y_0 - y_1|$, what would $\text{argmin}_{h \in H} R(h)$ be reduced to?

Definition 4.7. In empirical risk minimisation, we are given a dataset $(x_1, y_1), ..., (x_n, y_n)$, and the empirical risk of a decision function h is given by

$$\hat{R}(h) = n^{-1} \sum_{i \in [n]} l(h(x_i), y_i) \tag{39}$$

From now on, we assume $\mathbb{Y} = \mathbb{R}$ and $l(y_0, y_1) = (y_0 - y_1)^2$. This is also known as the ordinary least squared settings.

Exercise 4.8. In subsection 3.2, we assumed n to be sufficiently large. Fix n. What would happen when $m > n$? What would happen when $m \to \infty$?

[7] Some may say this as a hypothesis function, but that may be confused with the notion of hypothesis testing. In regression setting, this could be noted as regression function.

4.4 Penalisation

Definition 4.9. Let $\lambda > 0$. A ridge regression with parameter λ is a minimisation problem where we minimise

$$f(\beta) := \sum_{i \in [n]} (y_i - x_i^T \beta)^2 + \lambda ||\beta||_2^2 \tag{40}$$

instead of Equation 30.

Exercise 4.10.

- (H) Construct an experiment where $m = 80$ and $n = 30$ with the data simulated by $y_i = 3x_{1,i} + 2x_{5,i} + 2x_{29,i} + \varepsilon_i \quad \forall i \in [n]$
- (C) Code this experiment.
- (C) Use `sklearn.linear_model` package [8] to estimate the ridge regression. Try different λ and comment on the result.
- (H) Analyse the behaviour of $\min_{\beta \in \mathbb{R}^m} f(\beta)$ as $\lambda \to 0$ and $\lambda \to \infty$.

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`

5 More on infinity and towards Modern Machine Learning

5.1 Normed vector space

Definition 5.1. A vector space V over $\mathbb{R}$ is a set V along with notions of $+$, $\cdot$ such that

1. $x + y \in V \quad \forall x, y \in V$
2. Annotate $\lambda x := \lambda \cdot x$, then $\lambda x \in V \quad \forall x \in V,\ \lambda \in \mathbb{R}$
3. $\exists 0 \in V$ such that $x + 0 = x \quad \forall x \in V$ and $1 \in \mathbb{R}$ remains a scalar identity operation: $1x = x1 = x \quad \forall x \in V$
4. $x + y = y + x$ and $x + (y + z) = (x + y) + z \ \forall x, y, z \in V$
5. $\alpha(\beta x + \theta y) = (\alpha\beta)x + (\alpha\theta)y \quad \forall x, y \in V$ and $\alpha \in \mathbb{R}$

Definition 5.2. A subspace of U in V is a set $U \subset V$ such that $\forall x, y \in U$ and $\lambda \in \mathbb{R}$, we have $\lambda(x + y) \in U$

Example 5.3. Fix an arbitrary $y \in \mathbb{R}^n$. The set

$$X = \{x \in \mathbb{R}^n : x^T y = 0\}$$

is a subspace of $\mathbb{R}^n$. Visualise this set in the case of $n = 2$ and $n = 3$.

Definition 5.4. A norm on a vector space V is a map $||\cdot|| : V \to [0, +\infty)$ such that

1. $||x|| = 0 \iff x = 0 \ \forall x \in V$
2. $||\lambda x|| = |\lambda| ||x|| \quad \forall x \in V$ and $\lambda \in \mathbb{R}$
3. $||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in V$

A normed vector space is then a pair $(V, ||\cdot||)$

Definition 5.5. Let V be a vector space and $E \subset V$ with $|E| = k \in \mathbb{N}$.

- The span of E is defined as

$$span(E) = \{\sum_{j \in [n]} \alpha_j e_j : n \in \mathbb{N}, e_j \in E, \alpha_j \in \mathbb{R} \, \forall j\}$$

- E is linearly independent if for any $n \in \mathbb{N}$, $\alpha_j \in \mathbb{R}$ and $e_j \in E$,

$$\sum_{j \in [n]} \alpha_j e_j = 0 \quad \implies \alpha_j = 0 \ \forall j \in [n]$$

- E is a Hamel basis for V if it is linearly independent and $span(E) = V$.We say V to be finite dimensional and write $dim(V) = k$ in this case. If V does not have a Hamel basis, we say V to be infinite dimensional.

Example 5.6. $(\mathbb{R}^n, ||\cdot||_p)$ is a finite dimensional normed vector space

Definition 5.7. Let $(V, ||\cdot||)$ be a normed vector space. A ball in this space is defined as

$$B(x, r) = \{y \in V : ||x - y|| < r\}$$

Definition 5.8. Let $f : (V, ||\cdot||) \to \mathbb{R}$. f is continuous at x if

$$\forall \varepsilon > 0,\ \exists \delta > 0 \text{ such that } \forall y \in B(x, \delta), |f(y) - f(x)| < \varepsilon \tag{41}$$

We say f is continuous on $A \subset V$ if f is continuous at all $x \in A$.

Definition 5.9. Let $V = \mathbb{R}$ and $a < b$, $p \geq 1$. Denote $C([a, b])$ to be the set of all continuous functions $f : [a, b] \to \mathbb{R}$. Denote $||\cdot||_{L^p} : C([a, b]) \to [0, +\infty)$ to be a function

$$||f||_{L^p} := \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}} \tag{42}$$

Exercise 5.10. $(C([0, 1]), ||\cdot||_{L^1})$ is an infinite dimensional normed vector space.

To do so, you need to first show $C([0, 1])$ to be a vector space, then show that it is infinite dimensional — this can be done by proving by contradiction. Then, it remains to show $||\cdot||_{L^1}$ to be a norm on such space.

Example 5.11. The set

$$X = \{f : C([0, 1]) : \int_0^1 f(x)dx = 0\}$$

is a subspace of $C([0, 1])$.

Definition 5.12. Let $p \geq 1$,

$$\ell^p := \{x = (x_j)_{j=1}^{\infty} : x_j \in \mathbb{R}, \sum_{j=1}^{\infty} |x_j|^p < \infty\}$$

Define $||\cdot||_{\ell^p} : \ell^p \to [0, +\infty)$ to be a function

$$||x||_{\ell^p} := \left(\sum_{j=1}^{\infty} |x_j|^p \right)^{\frac{1}{p}} \tag{43}$$

Example 5.13. $(\ell^p, ||\cdot||_{\ell^p})$ is an infinite dimensional normed vector space.

5.2 Disorder in infinity: lack of completion

Throughout this subsection, we consider a normed vector space $(V, ||\cdot||)$.

Definition 5.14.

- A sequence $(x_n)_{n\in\mathbb{N}}$ converges to $x \in V$ if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ such that } x_n \in B(x, \varepsilon)\ \forall n \geq N \tag{44}$$

- A sequence $(x_n)_{n\in\mathbb{N}}$ is Cauchy if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ such that } ||x_n - x_m|| < \varepsilon\ \forall n, m \geq N \tag{45}$$

- $(V, ||\cdot||)$ is complete if every Cauchy sequence converges. Otherwise $(V, ||\cdot||)$ is incomplete.

Example 5.15. $(\mathbb{R}, |\cdot|)$ is complete. This is due to Bolzano–Weierstrass theorem.

Exercise 5.16. Show that $(\mathbb{R}^d, ||\cdot||_2)$ is complete.

Exercise 5.17. Show that $(C([0,1]), ||\cdot||_{L^1})$ is incomplete.

Exercise 5.18. Show that $(\ell^p, ||\cdot||_{\ell^p})$ is complete.

5.3 Order in infinity: Hilbert space, separability, and ℓ^2

By the end of this subsection, we would like to appreciate the following theorem, which is a very strong characterisation to relate Hilbert space into a well-established space.

Theorem 5.1. *Any infinite-dimensional separable Hilbert space H is isometrically isomorphic to $(\ell^2, ||\cdot||_{\ell^2})$*

Definition 5.19. An inner product is a map on vector space V, written as $\langle , \rangle : V \times V \to \mathbb{R}$ such that $\forall x, y, z \in V$ and $\forall \alpha \in \mathbb{R}$,

1. $\langle x, x\rangle \geq 0$ with equality if and only if $x = 0$
2. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
3. $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$
4. $\langle x, y\rangle = \langle y, x\rangle$

Example 5.20. Let $V = \ell^2$, consider $\langle , \rangle : V \times V \to \mathbb{R}$ defined by

$$\langle x, y\rangle = \sum_{j=1}^{\infty} x_j y_j \tag{46}$$

Then we can show this is an inner product.

Exercise 5.21. Show that when we define $||\cdot|| : V \to \infty$ by $||x|| := \langle x, x\rangle^{\frac{1}{2}}$ Then $(V, ||\cdot||)$ becomes a normed vector space. We call this norm the induced norm.

Definition 5.22. A Hilbert space H is an inner-product space for which it is complete with respect to the induced norm.

Definition 5.23.

- A function $f : X \to Y$ is an injection if

$$\forall a, b \in X, \quad f(a) = f(b) \implies a = b$$

- A function $f : X \to Y$ is a surjection if $\forall y \in Y$, $\exists x \in X$ such that $f(x) = y$.
- A function $f : X \to Y$ is a bijection if it is both injective and bijective.

- A function $f : X \to Y$ is linear if

$$\forall \alpha, \beta \in \mathbb{R}, x, y \in X, \quad f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$$

- A function $f : (X, ||\cdot||_X) \to (Y, ||\cdot||_Y)$ is an isometry if

$$\forall x \in X, \quad ||Tx||_Y = ||x||_X$$

- Two normed vector spaces $(X, ||\cdot||_X), (Y, ||\cdot||_Y)$ are isometrically isomorphic if there exists a bijective linear isometry $f : X \to Y$.

Exercise 5.24. Show that $\mathbb{C}$ is isometrically isomorphic to $\mathbb{R}^2$. Emphasise how you would define the norm and the isometry.

Lemma 5.2. *Let V be a finite-dimensional vector space. Then there exists a norm $||\cdot||$ such that $(V, ||\cdot||)$ is isometrically isomorphic to $(\mathbb{R}^n, ||\cdot||_2)$ where $n = dim(V)$.*

In relation to appreciating Theorem 5.1, we need to introduce further notations and concepts to assist infinite dimensional analysis.

Definition 5.25.

- A non-empty set A is countable if there exists a surjection from $\mathbb{N}$ to A.
- A set $A \subset (V, ||\cdot||)$ is closed if whenever $\{x_n\}_{n\in\mathbb{N}} \subset A$ with $x_n \underset{n\to\infty}{\longrightarrow} x$, it follows that $x \in A$. The closure of A, noted $\bar{A}$, is the intersection of all closed subsets of V that contain A.
- A set $A \subset (V, ||\cdot||)$ is dense in V if $\bar{A} = V$.
- A normed vector space[9] is separable if it contains a countable dense subset.

Example 5.26. The set of Euclidean basis vector $\{e_j\}_{j\in\mathbb{N}}$ is countable.

Exercise 5.27. $(\ell^p, ||\cdot||_{\ell^p})$ is separable. Prove.

We are now in a position to appreciate Theorem 5.1.

5.4 Hilbert space analysis

Further to the characterisation of separable Hilbert space, we return to a general Hilbert space and study one more property before applying to Machine Learning.

Definition 5.28. Let H be a Hilbert space. Let $X \subset H$. Define the orthogonal complement of X as

$$X^\perp = \{u \in H : \langle u, x\rangle = 0 \ \forall x \in X\}$$

Theorem 5.3 (Orthogonal projection in Hilbert space)**.** *Let U be a closed linear subspace of a Hilbert space H, then for any $x \in H$, there exists uniquely $u \in U, v \in U^\perp$ such that $x = u + v$*

Corollary 5.4. *Using the above notation, $||x||^2 = ||u||^2 + ||v||^2$.*

[9]There is a more general definition related to metric space, which we are not getting into as we intend to simplify the notations.

5.5 Applications to Machine Learning

Definition 5.29. Let $\mathbb{X}$ be the input space. A function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space H and a map $\phi : \mathbb{X} \to H$ such that

$$\forall x, x' \in X, \quad k(x, x') = \langle \phi(x), \phi(x') \rangle \tag{47}$$

Example 5.30. Let $\mathbb{X} = \mathbb{R}$ with $\phi(x) = x$, then $k(x, y) = x^T y$ is the kernel.

Definition 5.31. Let H be a Hilbert space of functions $f : \mathbb{X} \to \mathbb{R}$. A function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called a reproducing kernel of H if

$$\forall x \in \mathbb{X}, \quad k_x = k(\cdot, x) \in H \tag{48}$$

and

$$\forall x \in X, \forall f \in H, \langle f, k(\cdot, x) \rangle = f(x) \tag{49}$$

If H has a reproducing kernel, it is called a reproducing kernel Hilbert space (RKHS).

Exercise 5.32. A reproducing kernel is a kernel. Prove.

Recall the setting in Definition 4.7: we are now in a position to state and prove the Representer theorem in kernel machine learning. H below always stands for a RKHS.

Theorem 5.5 (Representer Theorem)**.** *Let $\Omega : [0, \infty) \to \mathbb{R}$ be increasing* [10]*. If*

$$\mathcal{H} = \operatorname*{argmin}_{f \in H} \hat{R}(f) + \Omega(||f||^2)$$

is not empty, then $\exists f^* \in \mathcal{H}$ such that $f^* = \sum_{i \in [n]} \alpha_i k(\cdot, x_i)$ for some constants α_i. Furthermore, if $\Omega : [0, \infty) \to \mathbb{R}$ is strictly increasing [11]*, then $\forall f \in \mathcal{H}$, $\exists$ constants α_i such that $f = \sum_{i \in [n]} \alpha_i k(\cdot, x_i)$.*

Hint of the proof: construct a projection of an arbitrary member of $\mathcal{H}$ onto the span of some basis we are after.

[10] That is, $\forall x, y \in [0, \infty)$ such that $x > y$, we have $\Omega(x) \geq \Omega(y)$

[11] That is, $\forall x, y \in [0, \infty)$ such that $x > y$, we have $\Omega(x) > \Omega(y)$