Artificial Neural Networks: advanced topics

Filippo Gatti

Université Paris-Saclay
CentraleSupélec, ENS Paris-Saclay, CNRS
LMPS Laboratoire de Mécanique Paris Saclay UMR 9026

*This chapter is a follow-up of the basic introduction provided in the chapter Artificial Neural Networks: layer architectures, optimizers and automatic differentiation, which prepares the readership to the following theoretical insights. In particular, the chapter rephrases the machine learning problem according to an information theory paradigm, that highlights the deep entanglement between the data science perspectives: the probabilistic and the deterministic one. Moreover, the chapter describes the fundamental theoretical result that paved the way to modern machine learning: the universal approximation theorem for a 1-hidden-layer perceptron. This section is followed by a continuum mechanics interpretation of convolutional neural networks, proving why convolutional layers are fundamental in image classification. Finally, further insights on the optimization of a neural network are provided, focusing on the convergence of first-order gradient descent methods. Finally, the automatic differentiation is explained in analogy with tensor algebra, along with some advanced strategies to avoid vanishing gradients in back-propagation algorithms. Some subsections are tagged as **[RECAP]**, since they are meant to refresh the readership's basics on optimization and signal processing fundamentals. The chapter is largely inspired, among others, by Stéphane Mallat's Data Science lecture notes at Collège de France, as well as by different lecture notes of CentraleSupélec's engineering curriculum.*

1 Information theory

The *theory of information* represents a fundamental chapter of the long way to modern machine learning. The inception of such a theory is due to Ronald Fisher and to Claude Shannon, who laid the theoretical framework in their famous works *On the mathematical foundations of theoretical statistics* [Fis22]

and *A Mathematical Theory of Communication* [Sha48]. In the following, the two perspectives will be briefly introduced.

1.1 Basics of Measure Theory [RECAP]

This section is a summary of the *Statistics and Learning* class of Centrale-Supélec [Cou20].
In the following, consider a multi-variate random variable defined over the probability space $\boldsymbol{X} : (\Omega, \mathcal{E}, \mathbb{P}) \to (\mathcal{X}, \Xi)$, with a σ-algebra $\mathcal{E} \in \mathcal{B}(\mathbb{R})$ on Ω ($\mathcal{B}(\mathbb{R})$ are the Borel's sets) and an unknown probability law $\mathbb{P}$ such that $\forall A \in \mathcal{E}_{\mathcal{X}}$, $P_X(A) = \mathbb{P}(\boldsymbol{X}^{-1}(A)) = \mathbb{P}(\boldsymbol{X} \in A)$. If the random variable is continuous, $\mathcal{X} \subset \mathbb{R}^{d_X}$ and $\Xi = \mathcal{B}(\mathcal{X})$; if instead $\mathcal{X}$ is of finite cardinality or countable, $\Xi = \mathcal{P}(\mathcal{X})$, the power set of $\mathcal{X}$.
Moreover, we assume that P_X is σ-finite measure dominated by a σ-finite measure μ, i.e. if $\forall \varepsilon > 0, \exists \delta(\varepsilon)$ such that $P_X(A) < \varepsilon, \forall A \in \Xi$ such that $\mu(A) < \delta(\varepsilon)$. In other words, the negligible sets for μ are negligible for P_X too. In this framework, P_X is absolutely continuous with the respect to μ and famous the Radon-Nikodym theorem holds (see also [Bil95; Cou20]):

Theorem 1. ***Radon-Nikodym theorem***
For two σ-finite measures on a sigma algebra Ξ, namely μ and P_X such that P_X is absolutely continuous with the respect to μ, $\exists p_X \in L^1(\mu)$ such that

$$P_X(A) = \int_A p_X(\boldsymbol{x}) \cdot \mu(d\boldsymbol{x}), \qquad \forall A \in \Xi$$

p_x is called probability density of P_X and it corresponds to the Radon-Nikodym derivative $p_X = \frac{dP_X}{d\mu}$.

In the following, we will consider only Lebesgue reference measures μ for continuous random variables, i.e.:

$$P_X(A) = \int_A p_X(\boldsymbol{x}) \cdot \mu(d\boldsymbol{x}), \qquad \forall A \in \Xi = \mathcal{B}(\mathbb{R}^{d_X}) \tag{1}$$

and countable reference measures for discrete random variables. In this latter case, the probability distribution is discrete and it corresponds to the point mass function corresponding to a counting measure on a subset the power set $\Xi = \mathcal{P}(\mathcal{X})$, such that:

$$\begin{aligned} P_X(A) = \int_A p_X(\boldsymbol{x}) \cdot \mu(d\boldsymbol{x}) = \int_A p(\boldsymbol{x}) \cdot \sum_{\boldsymbol{x}_i \in A} \delta_{\boldsymbol{x}_i}(d\boldsymbol{x}) = \\ = \sum_{\boldsymbol{x}_i \in A} P_X(\boldsymbol{X} = \boldsymbol{x}_i), \forall A \in \Xi = \mathcal{P}(\mathcal{X}) \end{aligned} \tag{2}$$

with $\sum_{\boldsymbol{x}_i \in \mathcal{X}} P_X(\boldsymbol{X} = \boldsymbol{x}_i) = 1$.

1.2 Reminders of basic statistics [RECAP]

Theorem 2. ***Strong law of large numbers (Kolmogorov)***
Given a set of random variables $\boldsymbol{X} = (X_i)_{i=1}^N$ *independent identically distributed (i.i.d.) and Lebesgue integrable, with expected value* $\mu_X < +\infty$*, and a sample average* $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$*, then :*

$$\mathbb{P}\left[\lim_{N\to+\infty} \bar{X}_n - \mu_X\right] = 1$$

which means that the sample average $\bar{X}_N$ *converges almost surely to the expected value* μ*.*

Theorem 3. ***Weak law of large numbers (Khintchine)***
Given a set of random variables $\boldsymbol{X} = (X_i)_{i=1}^N$ *independent identically distributed (i.i.d.) and Lebesgue integrable, with expected value* $\mu_X < +\infty$*, and a sample average* $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$*, then :*

$$\forall \varepsilon > 0, \quad \lim_{N\to+\infty} \mathbb{P}\left[|\bar{X}_n - \mu_X| \leq \varepsilon\right] = 1$$

which means that the sample average $\bar{X}_N$ *converges in probability to* μ

Proof. The theorem proof is based on the Chebyshev's inequality, that states that for a wide class of probability distributions, only a certain fraction of values at a certain distance from the expected value a probability to occur larger than the distance itself:

$$\mathbb{P}\left[|X - \mu| \geq \varepsilon\right] \leq \frac{\sigma_X^2}{\varepsilon^2} \tag{3}$$

Applying Equation (3) to the estimator $\bar{X}_N$ (which is a linear combination of random variables and has a variance $\sigma^2_{\bar{X}_N} = \frac{\sigma_X^2}{N}$), one obtains:

$$\forall \varepsilon > 0 \quad 1 \geq \lim_{N\to+\infty} \mathbb{P}\left[|\bar{X}_n - \mu_X| \leq \varepsilon\right] =$$

$$= 1 - \lim_{N\to+\infty} \mathbb{P}\left[|\bar{X}_n - \mu_X| > \varepsilon\right] \geq 1 - \lim_{N\to+\infty} \frac{\sigma_X^2}{N\varepsilon^2}$$

which proves the convergence in probability. The estimator $\bar{X}_N$ of μ_X is consistent. □

Definition 4. Consistent estimator
An estimator $\boldsymbol{\theta}_N = \boldsymbol{\theta}\left(X_1, \ldots, X_N\right)$ is said to be *consistent* if it converges in probability to its limit $\boldsymbol{\theta}$.

Theorem 5. *Given a series of random variables $(X_i)_{i=1}^N$ that converges in probability to X, then any continuous function $f : \mathcal{X} \to \mathbb{R}$ converges in probability too.*

Proof. The continuity of f is expressed by the following expression:

$$\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0 | \quad |f(X_N) - f(X)| \leq \varepsilon, \forall X_N \in [X - \delta; X + \delta]$$

Since X_N converges to X in probability, the following inequality holds:

$$1 \geq \mathbb{P}\left[|f(X_N) - f(X)| \leq \varepsilon\right] \geq \mathbb{P}\left[|X_N) - X| \leq \delta\right]$$

Taking the limit $N \to +\infty$ of the last inequality:

$$1 \geq \lim_{N \to +\infty} \mathbb{P}\left[|f(X_N) - f(X)| \leq \varepsilon\right] \geq 1$$

which proves the statement. □

Remark 6. Since $\bar{X}_N$ (defined in Theorem 3) is a consistent estimator of μ_X, then thanks to Theorem 5, $\bar{X}_N^2$ converges in probability to μ_X and the estimator $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^N X_i^2$ converges in probability to $\mathbb{E}\left[X_i^2\right]$. Therefore, the variance estimator of each X_i s_N^2 that reads:

$$s_N^2 = \frac{1}{N}\sum_{i=1}^N \left(X_i - \bar{X}_N\right)^2 \tag{4}$$

converges in probability to the variance of X_i:

$$\begin{aligned}
&\lim_{N \to +\infty} \mathbb{P}\left(|s_N^2 - \sigma_{X_i}^2| \leq \varepsilon\right) = \\
&= \lim_{N \to +\infty} \mathbb{P}\left[|\frac{1}{N}\sum_{i=1}^N \left(X_i^2 + \bar{X}_N^2 - 2X_i\bar{X}_N\right) - \sigma_{X_i}^2| \leq \varepsilon\right] = \\
&= \lim_{N \to +\infty} \mathbb{P}\left[|\bar{Y}_N - X_N^2 - \sigma_{X_i}^2| \leq \varepsilon\right] = \\
&= \lim_{N \to +\infty} \mathbb{P}\left[|\bar{Y}_N - X_N^2 - \mathbb{E}\left[X_i^2\right] + \mu_X| \leq \varepsilon\right] = 1
\end{aligned} \tag{5}$$

Definition 7. Convergence in distribution
A series of random variables $(X_i)_{i=1}^N$ sampled from a probability distribution $p_N\left((X_i)_{i=1}^N\right)$ converges to the probability distribution $p(X)$ for $N \to +\infty$ if

$$\forall a | P_X(A = \{X \leq a\}) \text{ is continuous and } \lim_{N \to +\infty} P_N\left(A = \{X \leq a\}\right) = P_X\left(a\right)$$

with $P_N\left(A = \{X \leq a\}\right) = \int_A p_N\left(x\right)\mu(dx)$

Theorem 8. ***Central limit theorem***
Given a set of i.i.d. random variables $\mathcal{D}_X = \{X_i\}_{i=1}^N$ *with expected value* $\mathbb{E}_{X_i} = \mu$ *and variance* $\mathbb{V}[X_i] = \sigma^2 < +\infty$ *and a random variable* $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$*, the random variable*

$$Z_N = \sqrt{N}\frac{\bar{X}_N - \mu}{\sigma}$$

has zero expected value, unit variance and its probability distribution $p(Z_N)$ *converges in distribution to a standard normal distribution* $\mathcal{N}(0,1)$

Proof. From Theorem 3, the weak law of large numbers grants that $\bar{X}_N$ converges in probability to μ and its variance $\mathbb{V}\left[\bar{X}_N\right] = \frac{\sigma^2}{N}$. The *characteristic function* of the function $Y_j = \frac{X_j - \mu}{\sigma}$ has a Taylor expansion around its zero expected value that reads:

$$\varphi_{Y_j}(t) = \mathbb{E}\left[e^{itY_j}\right] = 1 - \frac{t^2}{2} + o(t^2)$$

Therefore, the characteristic function of Z_N reads:

$$\varphi_{Z_N}(t) = \mathbb{E}\left[e^{itZ_N}\right] = \prod_{j=1}^{N}\mathbb{E}\left[e^{it\frac{Y_j}{\sqrt{N}}}\right] = \left(\varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right)\right)^N = \left(1 - \frac{t^2}{2N} + o\left(\frac{t^2}{N}\right)\right)^N$$

For $N \to +\infty$ the following results holds:

$$\lim_{N\to+\infty}\varphi_{Z_N}(t) = \lim_{N\to+\infty}\left(1 - \frac{t^2}{2N} + o(\frac{t^2}{N})\right)^N =$$

$$= e^{-\frac{t^2}{2}} = \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[e^{it\xi}\right] = \varphi_\xi(t), \quad t \in \mathbb{R}$$

In other words, the characteristic function of Z_N converges to the one of a random variable with standard normal probability distribution. Thanks to Lévy theorem, the probability distribution $p(Z_N)$ converges in distribution to $\mathcal{N}(0,1)$.

□

1.3 The Fisher's approach

Fisher [Fis22] firstly stated the concept of *information* for the *inference* problem stated in the following. Provided a family of parametric probability law, called statistical model $\mathcal{H}_{\boldsymbol{\theta}} := \left\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\right\}$, if $P \in \mathcal{H}_{\boldsymbol{\theta}}$, there exist $\boldsymbol{\theta}^*$ such that $P_{\boldsymbol{\theta}^*} = P$. $\boldsymbol{\theta}^*$ is unique if and only if $\mathcal{H}_{\boldsymbol{\theta}}$ is identifiable, i.e., if the map $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ is injective. In this sense, the value of $\boldsymbol{\theta}$ is an index to identify any probability distribution in $\mathcal{H}_{\boldsymbol{\theta}}$. Fisher proposed to discover P by finding the corresponding value of the parameter set $\hat{\boldsymbol{\theta}}$ that make $P_{\hat{\boldsymbol{\theta}}}$ the closest possible to $P = P_{\boldsymbol{\theta}^*}$. Fisher proposed to infer the "best" estimator $\hat{\boldsymbol{\theta}}$ from N realizations of observed

variable at stake $\boldsymbol{X}$, collected in the data set $\mathcal{D}_X = \{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1}^N$. Therefore, the $\hat{\boldsymbol{\theta}}$ depends on $\mathcal{D}_X$. In the remainder of the chapter, $\mathcal{X}$ is assumed to be isomorphic to $\mathbb{R}^{d_X}$, Θ a Borel set of $\mathbb{R}^p$ and for the sake of simplicity we assume $\mathcal{H}_{\boldsymbol{\theta}}$ as dominated by a Lebesgue or discrete σ-finite measure μ and represented by the probability density functions $p_{\boldsymbol{\theta}}$. What is the *best* estimator $\hat{\boldsymbol{\theta}}(\mathcal{D}_X)$ from which one can infer $p(\mathcal{E}_{\mathcal{X}})$? In other words, what is the most parsimonious choice of $\boldsymbol{\theta}$ in order to span the whole $\mathcal{E}_{\mathcal{X}}$? What is the information about $\mathcal{E}_{\mathcal{X}}$ contained in $\mathcal{D}_X$ and how we can extract it?

First, some regularity conditions for the statistical model $\mathcal{H}_{\boldsymbol{\theta}}$ must be defined [Bil95; Cou20]:

C1 Θ is an open set and $p_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0 \iff p_{\boldsymbol{\theta}'} > 0, \forall \boldsymbol{x} \in \mathcal{X}$ and $\forall (\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$. This implies that all $p_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}$ have the same support denoted $\square$.

C2 $\forall \boldsymbol{\theta} \in \Theta$, $p_{\boldsymbol{\theta}}$ can be differentiated under the integral[1]:

$$\boldsymbol{\nabla}_\theta \int_\square p_{\boldsymbol{\theta}} \cdot \ \mu(d\boldsymbol{x}) = \int_\square \boldsymbol{\nabla}_\theta p_{\boldsymbol{\theta}} \cdot \ \mu(d\boldsymbol{x}) \tag{7}$$

Moreover, according to Fisher, the estimator $\hat{\boldsymbol{\theta}}$ must be [**Billingsley˙1995Campagne˙2022**; Fis22]:

- unbiased (see Equation (30))
- consistent (i.e. it must converge in probability to $\boldsymbol{\theta}^*$, the parameter corresponding to the data probability distribution P)
- provided that $\mathcal{D}_X$ is *exhaustive*, i.e., sufficient to characterize the real yet unknown probability distribution $P = P_{\boldsymbol{\theta}^*}$

For Fisher, the way to find a consistent estimator $\hat{\boldsymbol{\theta}}(\mathcal{D}_X)$ is to maximize the *likelihood* $p_{\boldsymbol{\theta}}(\mathcal{D}_X) = \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_X)$, in order to *infer* any sample from the "true" probability distribution $\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}$, from the chosen probability family $\mathcal{H}_{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}}(\mathcal{D}_X) = \arg\max_{\boldsymbol{\theta} \in \Theta} p_{\boldsymbol{\theta}}(\boldsymbol{x} | \boldsymbol{x} \in \mathcal{D}_X) \tag{8}$$

When $\mathcal{D}_X$ represents a dataset of independent identically distributed (i.i.d.) variables, the likelihood can be factorized as

$$p_{\boldsymbol{\theta}}(\mathcal{D}_X) = \prod_{i=1}^N p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{9}$$

[1] According to [Bil95], C2 holds if the gradient is locally dominated by an integrable function g, i.e., it exist a neighborhood $\square$ of θ and g such that $\int_\square g(x)\mu(dx) < +\infty$ such that almost everywhere on a neighborhood V of $\boldsymbol{x}$:

$$\left| \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_k} \right| \leq g \tag{6}$$

.

Equation (8) is justified by the following theorem [Cam22]:

Theorem 9. *Given a set of i.i.d. random variables* $\mathcal{D}_X = \{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1}^N$*, then:*

$$\forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \forall \varepsilon > 0 \quad \lim_{N \to +\infty} \mathbb{P}\left[p_{\boldsymbol{\theta}}\left(\mathcal{D}_X\right) - p_{\boldsymbol{\theta}^*}\left(\mathcal{D}_X\right) < \varepsilon\right] = 1$$

Proof. In order to prove the convergence in probability of the event $p_{\boldsymbol{\theta}^*} > p_{\boldsymbol{\theta}}$, one can prove the rephrase the convergence in probability as:

$$\forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \forall \varepsilon > 0 \quad \lim_{N \to +\infty} \mathbb{P}\left[\frac{1}{N} \ln\left(\frac{p_{\boldsymbol{\theta}}\left(\mathcal{D}_X\right)}{p_{\boldsymbol{\theta}^*}\left(\mathcal{D}_X\right)}\right) < \varepsilon\right] = 1 \tag{10}$$

Provided that $\mathcal{D}_X = \{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1}^N$ is a set of i.i.d. variables, the likelihood is defined by Equation (9). It holds that:

$$\frac{1}{N} \ln\left(\frac{p_{\boldsymbol{\theta}}\left(\mathcal{D}_X\right)}{p_{\boldsymbol{\theta}^*}\left(\mathcal{D}_X\right)}\right) = \frac{1}{N} \sum_{i=1}^{N} \ln\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}_i\right)}\right) \tag{11}$$

According to the weak law of large numbers in Theorem 3 (because the factors $\ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{p_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i)}\right)$ are independent identically distributed, the empirical average $\frac{1}{N}\sum_{i=1}^N \ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{p_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i)}\right)$ is a consistent estimator of the expected value $\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{X})}{p_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\right)\right]$, i.e. (see Theorem 5):

$$\forall \varepsilon > 0 \lim_{N \to +\infty} \mathbb{P}\left[\left|\frac{1}{N} \sum_{i=1}^{N} \ln\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}_i\right)}\right) - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{X}\right)}\right)\right]\right| < \varepsilon\right] = 1 \tag{12}$$

The Jensen inequality states that [Cam22]:

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{X}\right)}\right)\right] < 0 \tag{13}$$

since ln function is a concave function and therefore[2]:

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{X}\right)}\right)\right] \leq \ln\left(\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{X}\right)}\right]\right) = 0$$

Provided the consistency of the estimator, expressed in Equation (12), the latter converges in probability to a negative "true" expected value, as proved

[2] If the random variable is discrete, according to Proposition 59, the linear combination of concave (convex) functions with positive coefficients (the point-mass probability distribution evaluate at each value) is concave. In the continuous case, the same argument holds, because $\int \ln p(x) \cdot p(x)$ is a linear combination with positive coefficients $p(x)$.

by the Jensen inequality in Equation (13). The objective of the theorem to prove the convergence in probability expressed in Equation (10) is granted by choosing $\varepsilon = \frac{|\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{X})}{p_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\right)\right]|}{2}$ and substituting it into Equation (12), in order to prove that

$$\lim_{N\to+\infty} \mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{p_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i)}\right) - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{X})}{p_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\right)\right]\right| < \frac{|\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{X})}{p_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\right)\right]|}{2}\right] = 1$$

and therefore that:

$$\forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \forall \varepsilon > 0 \quad \lim_{N\to+\infty} \mathbb{P}\left[\frac{1}{N}\ln\left(\frac{p_{\boldsymbol{\theta}}(\mathcal{D}_X)}{p_{\boldsymbol{\theta}^*}(\mathcal{D}_X)}\right) < \varepsilon\right] = 1$$

which proves the statement. □

Theorem 9 proves that the "true" probability distribution of the data, corresponding to $\boldsymbol{\theta}^*$ (if $p_{\boldsymbol{\theta}^*} \in \mathcal{H}_{\boldsymbol{\theta}}$), corresponds to the maximum on $\mathcal{H}_{\boldsymbol{\theta}}$. However, the assumption $p_{\boldsymbol{\theta}^*} \in \mathcal{H}_{\boldsymbol{\theta}}$ is not always satisfied: the "true" probability distribution that generated the dataset $\mathcal{D}_X$ is not known beforehand, which is why Fisher proposed to approximate the it by maximizing the likelihood over a chosen parameter space Θ, that generates $\mathcal{H}_{\boldsymbol{\theta}}$, leading to the "best" parameter estimator $\hat{\boldsymbol{\theta}}$. The "best" parameter $\hat{\boldsymbol{\theta}}$ does not necessarily exist, nor it is unique. For the sake of simplicity, Equation (8) is replaced by the following Maximum Log-Likelihood Estimation:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ \ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ \ln p_{\boldsymbol{\theta}}(\boldsymbol{X}) \tag{14}$$

Since $\mathcal{D}_X$ represents a limited dataset of independent identically distributed (i.i.d.) realizations, the maximization problem in Equation (14) is approximated by finding the estimator of $\hat{\boldsymbol{\theta}}$, noted as $\hat{\boldsymbol{\theta}}(\mathcal{D}_X)$:

$$\hat{\boldsymbol{\theta}}(\mathcal{D}_X) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_X) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ \sum_{i=1}^{N} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{15}$$

$\hat{\boldsymbol{\theta}}$ is a deterministic yet unknown parameter depending on the choice of $\mathcal{D}_X$. The value of $\hat{\boldsymbol{\theta}}$ relies on the optimization problem in Equation (14), with a finite dataset $\mathcal{D}_X$ at stake.
Substituting the likelihood with the log-likelihood is possible because of the following theorem:

Theorem 10. ***Maximum Likelihood of*** $g(\boldsymbol{\theta})$**.** *[Bil95; Cou20]*
Given a set of i.i.d. observations $\mathcal{D}_X = \{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1}^{N}$ *and* $\boldsymbol{\theta}$ *the of parameters that maximizes the likelihood* $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_X)$*. Then* $g\left(\hat{\boldsymbol{\theta}}\right)$ *is the maximum likelihood estimator of* $g(\boldsymbol{\theta})$*.*

Proof. The proof resides on the definition of the following Legendre transform, widely used in the thermodynamic theory of thermoelasticity:

$$\mathcal{L}^*\left(\boldsymbol{\eta};\mathcal{D}_X\right)=\sup_{\boldsymbol{\theta}:g(\boldsymbol{\theta})=\eta}\mathcal{L}\left(g^{-1}\left(\boldsymbol{\eta}\right);\mathcal{D}_X\right) \tag{16}$$

If $g:\Theta\mapsto g\left(\Theta\right)$ is a bijection, Equation (16) reduces to

$$\mathcal{L}^*\left(\boldsymbol{\eta};\mathcal{D}_X\right)=\prod_{i=1}^{N}p(\boldsymbol{x}_i|g^{-1}\left(\boldsymbol{\eta}\right))=\mathcal{L}\left(g^{-1}\left(\boldsymbol{\eta}\right);\mathcal{D}_X\right)$$

since $\boldsymbol{\theta}=g^{-1}\left(\boldsymbol{\eta}\right)$ is unique. Therefore

$$\sup_{\boldsymbol{\eta}}\mathcal{L}\left(\boldsymbol{\eta};\mathcal{D}_X\right)=\sup_{\boldsymbol{\theta}}\mathcal{L}\left(\boldsymbol{\theta};\mathcal{D}_X\right)$$

However, if g is not bijective, $\exists\boldsymbol{\theta}_1\neq\boldsymbol{\theta}_2$ such that $\boldsymbol{\eta}=g\left(\boldsymbol{\theta}_1\right)=g\left(\boldsymbol{\theta}_2\right)$ so that $\mathcal{L}^*\left(\boldsymbol{\eta};\mathcal{D}_X\right)=\prod_{i=1}^{N}p(\boldsymbol{x}_i|g^{-1}\left(\boldsymbol{\eta}\right))$ is not uniquely defined. In this general case, the Legendre transform in Equation (16) must be adopted. □

Remark 11. Theorem 10 and Theorem 9 imply that, if the "true" probability distribution $\boldsymbol{\theta}^*$ that generates the dataset $\mathcal{D}_X$ belongs to the open set Θ, i.e., if $p_{\boldsymbol{\theta}^*}\in\mathcal{H}_{\boldsymbol{\theta}}$, enlarging progressively the size of the realization set N, $\hat{\boldsymbol{\theta}}\left(\mathcal{D}_X\right)=\hat{\boldsymbol{\theta}}_N$ represents a unique sequence converging towards $\boldsymbol{\theta}^*$. As a matter of fact, due to Theorem 9, $\forall\boldsymbol{\theta}\in O_\varepsilon(\boldsymbol{\theta}^*)$, with $O_\varepsilon(\boldsymbol{\theta}^*)$ being an open neighborhood of $\boldsymbol{\theta}^*$, the probability associated to the set of observations S_N defined as:

$$S_N:=\left\{\boldsymbol{x}|\ln p_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}\right)>\sup_{\boldsymbol{\theta}\in\partial O_\varepsilon(\boldsymbol{\theta}^*)}\ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right\}$$

converges to 1:

$$\forall\varepsilon>0\quad\lim_{N\to+\infty}\mathbb{P}\left[|\mathbb{P}\left(S_N\right)-1|<\varepsilon\right]=1$$

which means that all the samples, when the data set of realization is sufficiently large, most probably belong to S_N, with the maximum log-likelihood over the closure $\bar{O}_\varepsilon\left(\boldsymbol{\theta}^*\right)$ corresponding to a point $\boldsymbol{\theta}^*\in\mathrm{Int}\left(\Theta\right)$. Because of the regularity condition C2 on $p_{\boldsymbol{\theta}}$, the latter is continuous and differentiable over $O_\varepsilon(\boldsymbol{\theta}^*)$. Thanks to Rolle's theorem, $\exists\boldsymbol{\theta}_N\in\bar{O}_\varepsilon(\boldsymbol{\theta}^*)$ such that $\boldsymbol{\nabla}_{\boldsymbol{\theta}}\ln p_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_N;\boldsymbol{x}\right)=\mathbf{0}$. This implies that $S_N\subset\tilde{S}_N$, defined as:

$$\tilde{S}_N:=\left\{\boldsymbol{x}|\exists\boldsymbol{\theta}_N\in\bar{O}_\varepsilon(\boldsymbol{\theta}^*)\text{ such that }\boldsymbol{\nabla}_{\boldsymbol{\theta}}\ln p_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_N;\boldsymbol{x}\right)=\mathbf{0}\right\}$$

since a priori $\ln p_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}\right)>\ln p_{\boldsymbol{\theta}_N}\left(\boldsymbol{x}\right)$ is not granted. Therefore, $\mathbb{P}\left(S_N\right)\leq\mathbb{P}\left(\tilde{S}_N\right)\leq 1$. Due to the convergence in probability of $\mathbb{P}\left(S_N\right)$ to 1, then $\mathbb{P}\left(\tilde{S}_N\right)$ converges in probability to 1, which means that for $\forall\varepsilon>0$, $\exists$ an open neighborhood $O_\varepsilon\left(\boldsymbol{\theta}^*\right)$ to which $\boldsymbol{\theta}_N$ most probably belongs, converging most probably towards $\boldsymbol{\theta}^*$, by definition.

Remark 12. The converging sequence $\boldsymbol{\theta}_N$ in Remark 11 maximizes the log-likelihood in the sense that

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}_N; \mathcal{D}_X \right) = \mathbf{0}$$

This expression has multiple solutions $\boldsymbol{\theta}_N$ but thanks to what exposed in the Remark 11 proves that $\boldsymbol{\theta}_N$ converges in probability towards $\boldsymbol{\theta}^*$ and it this sense one should interpret the quest for the MLE.

The following theorem proves that the sequence $\boldsymbol{\theta}_N$ converges to $\boldsymbol{\theta}^*$ in standard normal probability distribution (see Theorem 17). Before proving this aspect, other regularity conditions of the statistical models must be introduced.
If the regularity condition C1 holds and $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})$ (and $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_X)$) is differentiable almost everywhere in $\square$, one can define the *score* as:

$$\boldsymbol{s} \left(\boldsymbol{\theta}; \boldsymbol{X} \right) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}} (\boldsymbol{\theta}; \boldsymbol{X}) \tag{17}$$

and its estimator over a limited dataset $\boldsymbol{s} \left(\boldsymbol{\theta}; \mathcal{D}_X \right)$. If conditions C1 and C2 hold, from Equation (14) one can prove that the score is a centered random vector in $\boldsymbol{\theta}^*$:

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}} (\boldsymbol{\theta}^*; \boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} s(\boldsymbol{\theta}^*; \boldsymbol{X}) = \mathbf{0} \tag{18}$$

For a finite set of i.i.d. samples $\mathcal{D}_X$[3], Equation (18) is straightforward:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}} (\boldsymbol{\theta}^*; \mathcal{D}_X) &= \boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}} \ln \left(p_{\boldsymbol{\theta}^*} (\boldsymbol{\theta}^*; \mathcal{D}_X) \right) = \\ &= \sum_{i=1}^{N} \frac{p_{\boldsymbol{\theta}^*} (\boldsymbol{x}_i)}{p_{\boldsymbol{\theta}^*} (\boldsymbol{x}_i)} \boldsymbol{\nabla}_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}^*} (\boldsymbol{x}_i) = \frac{1}{N} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \sum_{i=1}^{N} p_{\boldsymbol{\theta}^*} (\boldsymbol{x}_i) = 0 \end{aligned} \tag{19}$$

The score in $\hat{\boldsymbol{\theta}}$ is nihil $\boldsymbol{s} \left(\hat{\boldsymbol{\theta}}; \mathcal{D}_X \right) = \mathbf{0}$ by Remarks 11 and 12.

Example 1. **Likelihood maximization with `PyTorch`**
In the following example, the maximum likelihood criterion is implemented in `PyTorch`, in order to find the parameters of the statistical model

$$\mathcal{H}_{\boldsymbol{\theta}} := \left\{ \frac{1}{(2\pi\theta_\sigma^2)^{\frac{1}{2}}} e^{-\frac{(x-\theta_\mu)^2}{2 \cdot \theta_\sigma^2}} ; \frac{1}{2\theta_\sigma} e^{-\frac{|x-\theta_\mu|}{\theta_\sigma}} \right\}$$

of probability distributions $p_{\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$, $p_{\boldsymbol{\theta}} : x \mapsto p_{\boldsymbol{\theta}}(x)$ and with

$$\Theta := \{\theta_\mu, \theta_\sigma\}, \quad \theta_\mu \in \mathbb{R}, \theta_\sigma \in \mathbb{R}^+$$

[3]The proof holds for any $p_{\boldsymbol{\theta}}$ satisfying the C1 and C2 conditions, both for discrete and continuous random variables. For the latter, an integral over $\mathcal{X}$ replaces the sum and by applying condition C2 in order to derive under the integral and viceversa.

The two probability distributions in $\mathcal{H}_{\boldsymbol{\theta}}$ correspond to Gaussian and Laplace distributions respectively. The "true" probability distribution belongs to $\mathcal{H}_{\boldsymbol{\theta}}$ and it is set by selecting two parameters

$$\bar{\theta}_\mu = \mu_X, \quad \bar{\theta}_\sigma = \sigma_X$$

Therefore, $p_{\bar{\theta}_\mu, \bar{\theta}_\sigma} = p_{\mu_X, \sigma_X} \in \mathcal{H}_{\boldsymbol{\theta}}$, which implies that $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^*$. The random variable X generates the data set of i.i.d. realizations $\mathcal{D}_X = \{x_i\}_{i=1}^N$, with N =10000. For the sake of simplicity, instead of maximizing the empirical average log-likelihood defined in Equation (15), the code minimizes the empirical average *Negative Log-Likelihood* $\mathcal{NLL}$, defined as:

$$\mathcal{NLL}(\boldsymbol{\theta}; \mathcal{D}_X) = \arg\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} \ln \frac{1}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \tag{20}$$

which still satisfies the convergence in probability expressed in Theorem 9 that reads:

$$\forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \forall \varepsilon > 0 \quad \lim_{N \to +\infty} \mathbb{P}\left[\frac{1}{N} \ln p_{\boldsymbol{\theta}}(\mathcal{D}_X) < \varepsilon + \frac{1}{N} p_{\boldsymbol{\theta}^*}(\mathcal{D}_X)\right] = 1 \tag{21}$$

In particular, given the statistical model in this example, the $\mathcal{NLL}$ function for a Gaussian distribution reads:

$$\mathcal{NLL}(\boldsymbol{\theta}; \mathcal{D}_X) = \frac{N}{2} \ln(2\pi) + N \ln \theta_\sigma + \frac{1}{2\theta_\sigma^2} \sum_{i=1}^{N} (x_i - \theta_\mu)^2 \tag{22}$$

and for a Laplace distribution:

$$\mathcal{NLL}(\boldsymbol{\theta}; \mathcal{D}_X) = N(\ln \theta_\sigma + \ln 2) + \frac{1}{\theta_\sigma} \sum_{i=1}^{N} |x_i - \theta_\mu| \tag{23}$$

The minimization is achieved by using the `AdamW` optimizing algorithm (see Section 3.4.5 and https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html), over n_e=100000 epochs.

The code below minimizes the $\mathcal{NLL}$ to find the "best" estimators $\hat{\theta}_\mu \approx \bar{\theta}_\mu = \mu_X$ and $\hat{\theta}_\sigma \approx \bar{\theta}_\sigma = \sigma_X$. The result of the iterative minimization is depicted in Figure 1

```
# Source: https://stackoverflow.com/questions/72469496/
#how-to-use-pytorch-for-maximum-likelihood-estimation-with-restrict
#-optimization
import numpy as np
import torch
from matplotlib import pyplot as plt

'''
Gaussian probability distribution
```

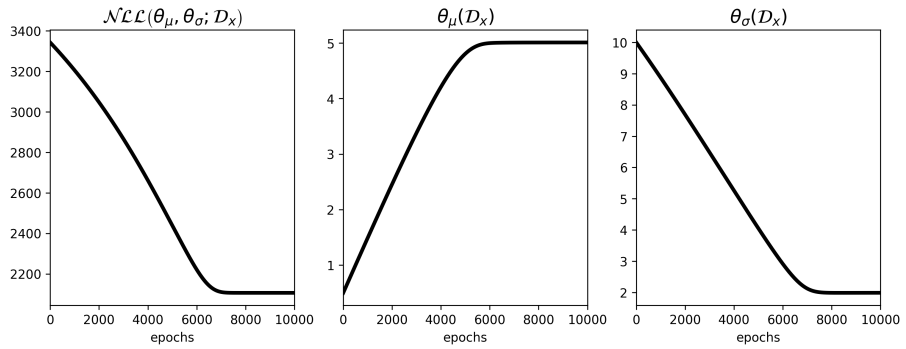

Figure 1: Iterative minimization (with `AdamW` algorithm) of $\mathcal{NLL}(\theta_\mu, \theta_\sigma; \mathcal{D}_X)$. The convergence of the "best" parameters $(\hat{\theta}_\mu, \hat{\theta}_\sigma)$ towards the "true" parameters $(\bar{\theta}_\mu, \bar{\theta}_\sigma)$ is shown, with the respect to the iteration number (epoch). The "true" probability distribution corresponds to $\mathcal{N}(\mu_X, \sigma_X)$, with μ_X=5 and σ_X=2. The figure was generated with code below.

```
'''
# Fix the pseudo-random generator see to grant reproductibility
torch.manual_seed(0)
# Create a dataset of nX samples from normal distribution
N = 10000
# real mean
mu_X = torch.tensor(np.array([5.0]),
                    dtype=torch.float64,
                    requires_grad=False).tile((N,))
# real standard deviation
sigma_X = torch.tensor(np.array([2.0]),
                    dtype=torch.float64,
                    requires_grad=False)

# generate random samples from p=N(mu_X, sigma_X)
D_X = torch.normal(mean=mu_X, std=sigma_X)
D_X.requires_grad = False

# Initialize the values of the estimators
theta_mu = torch.tensor(np.array([0.5]),
                        dtype=torch.float64,
                        requires_grad=True)
theta_sigma = torch.tensor(np.array([10.0]),
                           dtype=torch.float64,
                           requires_grad=True)

# Define the optimizer
learning_rate = 0.0001
optimizer = torch.optim.AdamW([theta_mu, theta_sigma], lr = learning_rate)

n_e = 100000

# Minimize the Negative Log-Likelihood iteratively
track_nll=[]
track_theta_mu = []
track_theta_sigma = []
```

```
for epoch in range(n_e):
    # Compute negative log-likelihood
    nll = N*(0.5*np.log(2.0*np.pi)+theta_sigma.log())
    nll+= ((((D_X-theta_mu)/theta_sigma).pow(2))/2.0).sum()
    optimizer.zero_grad()
    nll.backward()


    if epoch % 100 == 0:
        print("NLL: {}; theta_mu: {}; theta_sigma: {}".format(nll.data.numpy(),
                                                              theta_mu.data.numpy(),
                                                              theta_sigma.data.numpy()))
    optimizer.step()
    track_nll.append(float(nll))
    track_theta_mu.append(float(theta_mu))
    track_theta_sigma.append(float(theta_sigma))

# plot convergence curves
fig, ax = plt.subplots(nrows=1, ncols=3, sharex=True, figsize=(12,4))
ax[0].plot(track_nll,
           color='k',
           linewidth=3,
           label=r"$\mathcal{NLL}$")
ax[1].plot(track_theta_mu,
           color='k',
           linewidth=3,
           label=r"$\theta_\mu$")
ax[2].plot(track_theta_sigma,
           color='k',
           linewidth=3,
           label=r"$\theta_\sigma$")
ax[0].set_xlim(0,n_e)
ax[0].set_xlabel("epochs")
ax[1].set_xlabel("epochs")
ax[2].set_xlabel("epochs")
ax[0].set_title(r"$\mathcal{NLL}\left(\theta_\mu,\theta_\sigma;\mathcal{D}_x\right)$", fontsize=15)
ax[1].set_title(r"$\theta_\mu\left(\mathcal{D}_x\right)$",
                fontsize=15)
ax[2].set_title(r"$\theta_\sigma\left(\mathcal{D}_x\right)$",
                fontsize=15)
fig.savefig("NLL_gauss.png", dpi=300, bbox_inches="tight")

```

```
# Source: https://stackoverflow.com/questions/72469496/
#how-to-use-pytorch-for-maximum-likelihood-estimation-with-restrict
#-optimization
import numpy as np
import torch
from matplotlib import pyplot as plt

'''
Laplace probability distribution
'''
from torch.distributions.laplace import Laplace
# Fix the pseudo-random generator see to grant reproductibility
torch.manual_seed(0)
# Create a dataset of nX samples from normal distribution
N = 10000
# real mean
mu_X = torch.tensor(np.array([5.0]),
                    dtype=torch.float64,
                    requires_grad=False).tile((N,))
# real standard deviation
sigma_X = torch.tensor(np.array([2.0]),
```

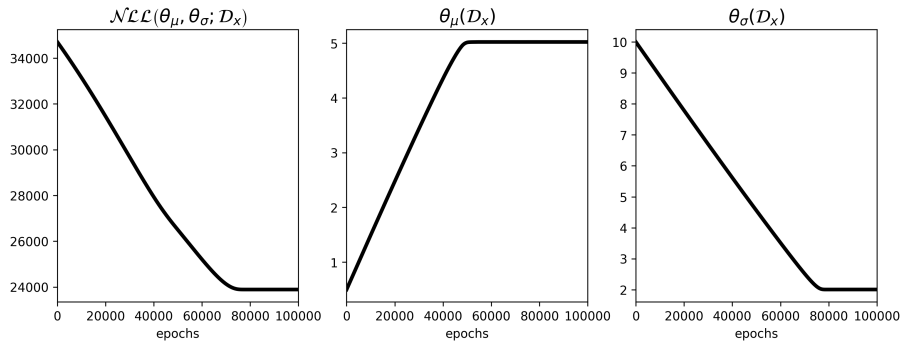

Figure 2: Iterative minimization (with `AdamW` algorithm) of $\mathcal{NLL}(\theta_\mu, \theta_\sigma; \mathcal{D}_X)$. The convergence of the "best" parameters $(\hat{\theta}_\mu, \hat{\theta}_\sigma)$ towards the "true" parameters $(\bar{\theta}_\mu, \bar{\theta}_\sigma)$ is shown, with the respect to the iteration number (epoch). The "true" probability distribution corresponds to Laplace(μ_X, σ_X), with μ_X=5 and σ_X=2. The figure was generated with code below.

```
                    dtype=torch.float64,
                    requires_grad=False)

# generate random samples from p=N(mu_X, sigma_X)
generator = Laplace(mu_X, sigma_X)
D_X = generator.sample()
D_X.requires_grad = False

# Initialize the values of the estimators
theta_mu = torch.tensor(np.array([0.5]),
                        dtype=torch.float64,
                        requires_grad=True)
theta_sigma = torch.tensor(np.array([10.0]),
                           dtype=torch.float64,
                           requires_grad=True)

# Define the optimizer
learning_rate = 0.0001
optimizer = torch.optim.AdamW([theta_mu, theta_sigma], lr = learning_rate)

n_e = 100000

# Minimize the Negative Log-Likelihood iteratively
track_nll=[]
track_theta_mu = []
track_theta_sigma = []
for epoch in range(n_e):
  # Compute negative log-likelihood
  nll = N*(np.log(2.0)+theta_sigma.log())
  nll+= (D_X-theta_mu).abs().sum()/theta_sigma.abs()
  optimizer.zero_grad()
  nll.backward()


  if epoch % 100 == 0:
    print("NLL: {}; theta_mu: {}; theta_sigma: {}".format(nll.data.numpy(),
                                                          theta_mu.data.numpy(),
```

```
                                                    theta_sigma.data.numpy()))
    optimizer.step()
    track_nll.append(float(nll))
    track_theta_mu.append(float(theta_mu))
    track_theta_sigma.append(float(theta_sigma))

# plot convergence curves
fig, ax = plt.subplots(nrows=1, ncols=3, sharex=True, figsize=(12,4))
ax[0].plot(track_nll,
           color='k',
           linewidth=3,
           label=r"$\mathcal{NLL}$")
ax[1].plot(track_theta_mu,
           color='k',
           linewidth=3,
           label=r"$\theta_\mu$")
ax[2].plot(track_theta_sigma,
           color='k',
           linewidth=3,
           label=r"$\theta_\sigma$")
ax[0].set_xlim(0,n_e)
ax[0].set_xlabel("epochs")
ax[1].set_xlabel("epochs")
ax[2].set_xlabel("epochs")
ax[0].set_title(r"$\mathcal{NLL}\left(\theta_\mu,\theta_\sigma;\mathcal{D}_x\right)$",
                fontsize=15)
ax[1].set_title(r"$\theta_\mu\left(\mathcal{D}_x\right)$",
                fontsize=15)
ax[2].set_title(r"$\theta_\sigma\left(\mathcal{D}_x\right)$",
                fontsize=15)
fig.savefig("NLL_laplace.png", dpi=300, bbox_inches="tight")
```

Based on Equation (19) and assuming an unbiased estimator (see Equation (30)), i.e., $\mathbb{E}_{\mathcal{D}_X \subset \mathcal{X}}\left(\hat{\boldsymbol{\theta}}\right) = 0$, in his seminal work, Fisher stated the concept of *information* associated to the parameters $\mathbb{I}(\boldsymbol{\theta}^*)$ as the variance of $\boldsymbol{\nabla}_\theta \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{X})$:

$$\mathbb{I}_F(\boldsymbol{\theta}^*; \boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}}\left[\boldsymbol{\nabla}_\theta \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{X}) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}; \boldsymbol{X})\right] \geq 0 \tag{24}$$

$\mathbb{I}_F\left(\boldsymbol{\theta}^*; \boldsymbol{X}\right)$ in Equation (24) is called Fisher Information Matrix (FIM), and it is a $N \times N$ positive semidefinite matrix. The FIM corresponds to the variance of the score $\mathbb{V}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}}\left[\boldsymbol{s}\left(\boldsymbol{\theta}^*; \boldsymbol{X}\right)\right]$ If $\boldsymbol{X}$ is composed by i.i.d. variables, the FIM reads:

$$\begin{aligned}\mathbb{I}_F(\boldsymbol{\theta}^*; \mathcal{D}_X) = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_i) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_i)\right] - \\ - \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_i) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_j)\right] = N \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)\end{aligned} \tag{25}$$

with the term $\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_i) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_j)\right] = \mathbf{0}$ because of the fact that the score function is centered in $\boldsymbol{\theta}^*$ [Bil95; Cou20; Cam22]. Equation (25) is another way to express the fact that the variance of

a sum of i.i.d. variable is the sum of the variances. If $\mathbb{I}_F(\boldsymbol{\theta}^*)$ is positive definite, then it defines a Riemann metric on the N-dimensional parameter space. A set of i.i.d. random variables, with high Fisher information has a score with large variance, i.e., it is *informative*. The "narrower" (in average on $\mathcal{D}_X$) is minimum of the log-likelihood, the more informative is $\boldsymbol{\theta}^*$ and therefore the higher is the variance of the scire (the largest the span of data around the minimum) [Cam22; Cou20]. Under the two following extra regularity conditions:

C3 For almost every $\boldsymbol{x} \in \square$, it exists a continuous mapping $h : \Theta \to p_{\boldsymbol{\theta}}$ such that $h : \theta \times \square \mapsto h_{\boldsymbol{\theta}}(\boldsymbol{x})$ with $h \in C^2(\Theta)$

C4 $\forall \boldsymbol{\theta} \in \Theta$, $\forall i, j$, $1 \leq i, j \leq p$ the transport theorem can be applied as follows[4]:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \int_{\square} \boldsymbol{\nabla}_{\theta} p_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{X}) \; \mu(d\boldsymbol{x}) = \int_{\square} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{X}) \; \mu(d\boldsymbol{x}) \tag{26}$$

Under C1, C2, C3 and C4 regularity conditions and for $\boldsymbol{\nabla}_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{X}) \in C^2(\Theta)$, the Fisher's information can be also expressed, following Equation (25), as follows:

$$\mathbb{I}_F(\boldsymbol{\theta}^*; \mathcal{D}_X) = -\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}} \left[\mathbf{H}_{\ell}(\boldsymbol{\theta}^*, \mathcal{D}_X) \right] \tag{27}$$

with $\mathbf{H}_{\ell}$ being the Hessian matrix of $\ell(\boldsymbol{\theta}; \mathcal{D}_X) = \ln p_{\boldsymbol{\theta}}(\mathcal{D}_X)$. For i.i.d. samples in $\mathcal{D}_X$, this is proven by considering Equation (24) and conditions C1, C2, C3 and C4:

$$\begin{aligned}
\mathbb{I}_F(\boldsymbol{\theta}^*; \mathcal{D}_X) =& \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}} \left[\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_i) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; x_i) \right] = \\
=& \sum_{i=1}^{N} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{x}_i) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{x}_i) = \\
=& \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \left(\sum_{i=1}^{N} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}} \cdot p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{x}_i) \right) - \\
& - \sum_{i=1}^{N} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \left(\ln p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{x}_i) \cdot p_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*; \boldsymbol{x}_i) \right) = -\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}} \left[\mathbf{H}_{\ell}(\boldsymbol{\theta}^*, \mathcal{D}_X) \right]
\end{aligned} \tag{28}$$

In order to determine the accuracy of the iterative gradient descent methods employed in machine learning to approximate $\boldsymbol{\theta}^*$ with $\hat{\boldsymbol{\theta}}$ (see Section 4.2), the Cramér-Rao bound (CRB) can be helpful since it represents an accuracy limit to $\boldsymbol{\theta}^*$.

Theorem 13. ***Cramér-Rao bound***
Given a set of i.i.d. variables $\mathcal{D}_X = \{\boldsymbol{x}_i\}_{i=1}^N$, sampled from a probability distribution $p_{\boldsymbol{\theta}^*} \in \mathcal{H}_{\boldsymbol{\theta}}$, with $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}) = \prod_{i=1}^N p_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i)$, and a statistical model (or

[4] Under the same conditions that led to Equation (6)

algorithm) to estimate $\boldsymbol{\theta}^*$*, that reads* $\boldsymbol{\theta} = \mathrm{A}(\boldsymbol{x})$ *then the mean estimator reads:*

$$\bar{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X})\right]$$

and the variance of the estimator is bounded by below as follows:

$$\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(\mathrm{A}(\boldsymbol{X})\right) \geq \frac{\|\boldsymbol{\nabla}_{\boldsymbol{\theta}^*}\bar{\boldsymbol{\theta}}\|^2}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)} \tag{29}$$

A *is unbiased if:*

$$\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(\mathrm{A}(\boldsymbol{X})\right) \geq \frac{1}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)} \tag{30}$$

Proof. The gradient $\boldsymbol{\nabla}_{\boldsymbol{\theta}^*}\mathrm{A}$ can be computed by differentiating under the integral, as:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}^*}\mathrm{A} = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X}) \cdot \boldsymbol{\nabla}_{\boldsymbol{\theta}^*} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right] =$$

$$= \mathbb{C}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X}) \cdot \boldsymbol{\nabla}_{\boldsymbol{\theta}^*} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right] +$$

$$+\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X})\right] \cdot \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^*} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right]$$

Since $\boldsymbol{\theta}^*$ generates the i.i.d. dataset the score is centered (see Equation (18)), which implies that the following expression holds

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}^*}\mathrm{A} = \mathbb{C}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X}) \cdot \boldsymbol{\nabla}_{\boldsymbol{\theta}^*} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right]$$

and the square norm is bounded by above as follows (Cauchy-Schwartz's inequality):

$$\|\boldsymbol{\nabla}_{\boldsymbol{\theta}^*}\mathrm{A}\|^2 = \left\|\mathbb{C}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X}) \cdot \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right]\right\|^2 \leq$$

$$\leq \mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}(\boldsymbol{X})\right] + \mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^*} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right]$$

The term $\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^*} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right]$ corresponds to $\mathbb{I}_F(\boldsymbol{\theta}^*; \boldsymbol{X}) = N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)$ which proves the CRB in the general case.
A is unbiased if $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ which implies that:

$$\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(\mathrm{A}(\boldsymbol{X})\right) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^2\right] \geq \frac{1}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)} \tag{31}$$

□

Remark 14. Choosing an unbiased Equation (31) CRB allow to assess the accuracy of the estimator $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_X)$, obtained, for instance, via gradient descent algorithm $\mathrm{A}(\boldsymbol{X})$.
The CRB states that the higher is the information of an estimator $\hat{\boldsymbol{\theta}}$, the higher is the variance of the score and, because of the CRB, the lower is its variance (the narrower is the log-likelihood maximum). The FIM and the CRB steer the convergence rate of any algorithm that attempts at maximize the log-likelihood of any statistical model adopted to infer from the available dataset. However, from a practical standpoint, the FIM matrix (dense) can easily reach a unbearable computational cost, which is why, in practice, it is rarely explicitly adopted. First order gradient descent methods are adopted instead (see Equation (63)).

Remark 15. Provided the CRB for an unbiased estimator, its "efficiency" can be measure as the ratio between the variance lower bound expressed by the CRB and stated in Equation (31) and the variance of the estimator:

$$\eta_{\boldsymbol{\theta}} = \frac{1}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1) \cdot \mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(\mathrm{A}(\boldsymbol{X})\right)} \leq 1 \tag{32}$$

Remark 16. One can notice that, maximizing the log-likelihood leads to minimize the so called Kullback-Leibler distance $\mathbb{D}_{KL}\left(p \| p_{\boldsymbol{\theta}}\right) = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln \frac{p}{p_{\boldsymbol{\theta}}}\right] < +\infty$[5] between $p_{\boldsymbol{\theta}}$ and the true probability distribution $p(x)$ (unknown). As a matter of fact:

$$\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln p_{\boldsymbol{\theta}}\right] = \max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln \frac{p_{\boldsymbol{\theta}}}{p}\right] + \max_{\boldsymbol{\theta}\in\Theta} \underbrace{\mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln p\right]}_{\leq 0} \leq \min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln \frac{p}{p_{\boldsymbol{\theta}}}\right] \tag{33}$$

$\mathbb{D}_{KL}\left(p \| p_{\boldsymbol{\theta}}\right)$ measures a *distance* between the real probability distribution and parametric probability density $p_{\boldsymbol{\theta}}$. Note that $\mathbb{D}_{KL}\left(p \| p_{\boldsymbol{\theta}}\right) \neq \mathbb{D}_{KL}\left(p_{\boldsymbol{\theta}} \| p\right)$. It is worth noticing that, in practice, $\mathbb{D}_{KL}\left(p \| p_{\boldsymbol{\theta}}\right) > \mathbb{D}_{KL}\left(p \| \mathcal{H}_{\boldsymbol{\theta}}\right) = \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{D}_{KL}\left(p \| \mathcal{H}_{\boldsymbol{\theta}}\right) \neq 0$ if $p \notin \mathcal{H}_{\boldsymbol{\theta}}$. This means that our estimator will have poor chances to discover P [Cou20; Cam22].

Example 2. **Compute $\mathbb{D}_{KL}\left(p_{\boldsymbol{\theta}} \| p\right)$ with `PyTorch`**[6]
To avoid underflow issues when computing this quantity, this loss expects the argument input in the log-space. As all the other losses in `PyTorch`, this function expects the first argument, input, to be the output of the model (e.g. the neural network) and the second, target, to be the observations in the dataset.

[5] As a convention $0 \cdot \ln 0 = 0 \cdot \ln \frac{0}{0} = 0$
[6] https://pytorch.org/docs/stable/generated/torch.nn.KLDivLoss.html

```python
import torch.nn.functional as F
# loss = loss_pointwise.sum() / input.size(0)
kl_loss = nn.KLDivLoss(reduction="batchmean")
# input should be a distribution in the log space
ptheta = F.log_softmax(torch.randn(3, 5, requires_grad=True), dim=1)
# Sample a batch of distributions. Usually this would come from the dataset
p = F.softmax(torch.rand(3, 5), dim=1)
output = kl_loss(ptheta, p)

kl_loss = nn.KLDivLoss(reduction="batchmean", log_target=True)
log_target = F.log_softmax(torch.rand(3, 5), dim=1)
output = kl_loss(input, log_target)
```

The following theorem proves that, provided an extra regularity condition of the statistical model and thanks to the CRB, the MLE estimator converges in standard normal probability to $\boldsymbol{\theta}^*$.

Theorem 17. ***MLE convergence in standard normal probability*** *[Cou20; Cam22]. Provided a statistical model* $\mathcal{H}_{\boldsymbol{\theta}} := \left\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_\Theta}\right\}$ *with regularity conditions C1, C2, C3, C4, an i.i.d. dataset* $\mathcal{D}_X = \{\boldsymbol{x}_i\}_{0<i\leq N}$ *generated by a parameter* $\boldsymbol{\theta}^* \in Int(\Theta)$*, a Maximum Log-likelihood Estimator* $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\ \ell(\boldsymbol{\theta}; \mathcal{D}_X)$ *and a further regularity condition C5 of* $\mathcal{H}_{\boldsymbol{\theta}}$ *that reads:*

$$\|\boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}\| < M(\boldsymbol{X}) \quad \forall \boldsymbol{\theta} \in \Theta, \mathbb{E}_{\boldsymbol{x}\in p_{\boldsymbol{\theta}}}[M(\boldsymbol{x})] < +\infty \tag{34}$$

all MLE sequences $\hat{\boldsymbol{\theta}}_N$ *converge in standard normal probability to* $\boldsymbol{\theta}^*$*:*

$$\lim_{N\to+\infty} \mathbb{P}\left[\left|p_{\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*)} - \mathcal{N}\left(\mathbf{0}, \mathbb{I}_F^{-1}(\boldsymbol{\theta}^*; \mathcal{D}_X)\right)\right|\right] = 1 \tag{35}$$

Therefore, because of the CRB, the MLE is an asymptotically "optimum" estimator, i.e. its efficiency $\eta_{\hat{\boldsymbol{\theta}}_N}$ *(defined in Equation (32)) converges in probability to 1:*

$$\lim_{N\to\infty} \mathbb{P}\left[\left|\frac{1}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1) \cdot \mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}(\mathrm{A}(\boldsymbol{X}))} - 1\right|\right] = 1 \tag{36}$$

Proof. According to Remark 11, the MLE sequence $\hat{\boldsymbol{\theta}}_N$ exists and it is unique. Moreover, the estimator maximizes the log-likelihood, so its score is nihil in $\hat{\boldsymbol{\theta}}_N$ but not necessarily in $\boldsymbol{\theta}^*$. Therefore, a Taylor expansion around $\hat{\boldsymbol{\theta}}_N$ (see the regularity conditions C1 to C5) reads:

$$\boldsymbol{s}(\boldsymbol{\theta}^*; \mathcal{D}_X) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}^*; \mathcal{D}_X) = \mathbf{H}_\ell(\boldsymbol{\theta}^*; \mathcal{D}_X)\left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_N\right) + o\left(\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_N\|\right)$$

The score $\boldsymbol{s}(\boldsymbol{\theta}^*; \mathcal{D}_X)$ (a gradient, therefore linear) is the sum of N i.i.d. scores with zero mean (see Equation (18)) and variance equal to the FIM. In this case, the central limit Theorem 8 states that the probability distribution of the variable $\sqrt{\frac{N}{\mathbb{I}_F(\boldsymbol{\theta}^*; \mathcal{D}_X)}}\boldsymbol{s}(\boldsymbol{\theta}^*; \mathcal{D}_X)$ converges in distribution (see Definition 7)

to $\mathcal{N}(\mathbf{0}, \mathbf{I})$, or alternatively, the the probability distribution of the variable $\sqrt{N}\boldsymbol{s}\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right)$ converges in distribution to $\mathcal{N}(\mathbf{0}, \mathbb{I}_F\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right))$. Now, the Taylor expansion of the Hessian around $\boldsymbol{\theta}^*$ reads:

$$\mathbf{H}_\ell\left(\hat{\boldsymbol{\theta}}_N; \mathcal{D}_X\right) = \mathbf{H}_\ell\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right) + \left(\boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{s}\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right)\right) . \left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*\right) +$$

$$+ o\left(\|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*\|\right)$$

with $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}; \mathcal{D}_X\right) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{s}\left(\boldsymbol{\theta}; \mathcal{D}_X\right)$. Provided C3 and C4, the FIM $N\mathbb{I}_F\left(\boldsymbol{\theta}^*; X_1\right)$ is equal to $\mathbf{H}_\ell\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right)$ (see Equation (27)). Provided C5, the latter term $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \otimes \boldsymbol{s}\left(\hat{\boldsymbol{\theta}}; \mathcal{D}_X\right)$ is bounded by $M\left(\mathcal{D}_X\right)$ and therefore it tens to $\mathbf{0}$ for $N \to +\infty$ which implies the limit :

$$\lim_{N \to +\infty} \sqrt{N}\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*\right) = \lim_{N \to +\infty} \frac{\sqrt{N}\boldsymbol{s}\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right)}{N\mathbb{I}_F\left(\boldsymbol{\theta}^*; X_1\right)}$$

converges in distribution to $\mathcal{N}\left(\mathbf{0}, \mathbb{I}_F^{-1}\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right)\right)$, which proves the statement.

□

Remark 18. Theorem 17 establishes the confidence intervals of the MLE algorithm, providing the possibility of assessing the probability (standard normal) that the MLE approaches the "true" probability distribution $p_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}\right)$ [Cam22].

Remark 19. For $\mathcal{NN}$, the Fisher approach can fail because $\boldsymbol{\theta}^*$ is not unique, since the loss function $L_{\mathcal{D}_{XY}}$ defined in $(\mathcal{P})$ in Section 2 is not convex [Cam22]. Moreover, in practice, the dimension of the probability space $m \gg d_X$, which implies that the problem is over-parametrized, which makes it difficult to find unbiased estimators [Cam22].

1.4 The Shannon's approach

The concept of *information* was better defined by Shannon, in the framework of the so called *source coding problem*, i.e. the problem of a source emitting a message and a receiver capturing it. The message is *meaningful* only if the receiver had no *a priori* knowledge of the message itself. Part of the information attached to the message is lost if the receiver has already have a clue of it [Clo22]. Deterministic messages bare zero information. Essentially, the whole *theory of information* relies on the basic concept of *entropy* that - in analogy with thermodynamics - provides an average level of *information* or *uncertainty* of a random variable's value. In its discrete form, the *entropy* of a discrete random variable X that can assume any value in $\mathcal{D}_X = \{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1}^N$ with the discrete probability distributions $p(\mathcal{D}_X) = (p_1, \ldots, p_N)$ reads:

$$\mathbb{H}(p(\mathcal{D}_X)) = -\sum_{i=0}^{N} p(X = \boldsymbol{x}_i) \cdot \ln(p(X = \boldsymbol{x}_i)) = -\mathbb{E}_{x\sim p}\left[\ln p(x)\right] \tag{37}$$

Some examples:

- The entropy of the discrete uniform distribution $\mathcal{U}$ reads:

$$\mathbb{H}(\mathcal{U}(\mathcal{D}_X)) = -\sum_{i=1}^{N} \frac{1}{N} \cdot \ln\left(\frac{1}{N}\right) = \ln N = \ln|\mathcal{D}_X| \tag{38}$$

  In this case, each $\boldsymbol{x}_i \in \mathcal{D}_X$ has the same probability $\frac{1}{N}$

- The entropy of a deterministic variable of distribution $p(\boldsymbol{x}_i)\delta(\boldsymbol{x}_i - x_j)$ reads:

$$\mathbb{H}(\delta) = -\sum_{i=1}^{N} \delta(\boldsymbol{x}_i - x_j) \cdot \ln \delta(\boldsymbol{x}_i - x_j) = -1 \cdot \ln 1 = 0 \tag{39}$$

- The discrete uniform distribution has the largest entropy:

$$0 \leq \mathbb{H}(p) \leq \mathbb{H}(\mathcal{U}(\mathcal{D}_X)) = \ln N \tag{40}$$

  As a matter of fact,

$$\max_{\sum_{i=1}^{N} p_i = 1} \mathbb{H}(p_1, \ldots, p_N) = \sum_{i=1}^{N} \min_{\sum_{i=1}^{N} p_i = 1} p_i \ln p_i \qquad \arg\min_{p_i \in [0,1]} p_i \ln p_i = \frac{1}{e}$$

  Unfortunately, choosing $p_i = \frac{1}{e}$ does not necessarily yield $\sum_{i=1}^{N} p_i = 1$, but, recalling that $p(\boldsymbol{x}_i) = \frac{1}{\alpha} p(\alpha \boldsymbol{x}_i)$ and assuming $p(\boldsymbol{x}_i) = p(x_j), \quad \forall i, j \in N$, the uniform probability distribution $p = p_i = p(\alpha \boldsymbol{x}_i)$ can be normalized by $\alpha = \frac{e}{N}$, i.e. $p_i = \frac{1}{N}$ with $\boldsymbol{x}_i \sim \mathcal{U}(\mathcal{D}_X)$, whose entropy is $\ln N$ (see Equation (38)) [Cam22].

A practical explanation of what the Shannon's entropy means is provided in [Rio18] and adapted to mechanics in [Clo22]. Consider the Darcy's law for a fixed total discharge Q across a set of N independent and laterally isolated parallel channels of length L and area A, with 1D laminar flow and random hydraulic conductivity. The total discharge

$$Q = \left(\sum_{i=1}^{N} K_i\right) \frac{\Delta h \cdot A}{L}$$

with an overall hydraulic head Δh. The random hydraulic conductivity of each channel K_i is considered as a random variable, sampled from a set of values

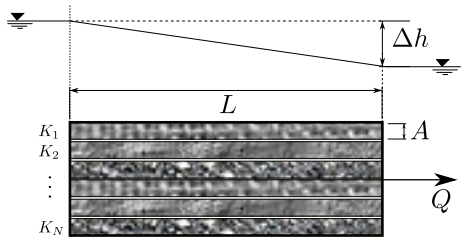

Figure 3: Example of the Darcy's law in a set of N parallel channels with random hydraulic conductivity K_i.

$\mathcal{D}_K = \{K_1, \ldots, K_b\}$ with cardinality $|\mathcal{D}_K| = b$ and $p(K) = (p_1, \ldots, p_b)$. What is then the average number of channels with random hydraulic conductivity $K_i \in \mathcal{D}_K$ that delivers the total discharge Q? If one considers a uniform discrete distribution of hydraulic conductivity $p_i = \frac{1}{b}$, $K \sim \mathcal{U}(\mathcal{D}_K)$, given the fact that one could pick N times any value of K_i in $\mathcal{D}_K$, the probability of a set of N independent channels is defined as $p\left(\{K_i\}_{i=1}^N\right) = b^{-N}$ and the average number of channels reads:

$$\mathbb{E}_{n\sim p}[N] = -\mathbb{E}_{n\sim p}\left[\log_b p\left(\{K_i\}_{i=1}^N\right)\right] = H(p_1, \ldots, p_N) - \frac{1}{\ln b} \tag{41}$$

For uniform distribution, $H(p_1, \ldots, p_N)$ is a positive monotonic function of N. $-\frac{1}{\ln b}$ plays the role of normalization constant. Instead, for a generic distribution of the K_i, the likelihood of N reads:

$$p\left(\{K_i\}_{i=1}^N\right) = \prod_{i=1}^N p_i^{b_i} = b^{N\left(\sum_{i=1}^N f_i \log_b p_i\right)} = b^{-N\cdot(\mathbb{D}_{KL}((f_1,\ldots,f_N)\|(p_1,\ldots,p_N))+\mathbb{H}(f_1,\ldots,f_N))} \tag{42}$$

with b_i being the number of times that K_i is picked and $f_i = \frac{b_i}{N}$ its empirical frequency ($\sum_{i=1}^b f_i = 1$). For the law of large numbers (by Bernoulli), $f_i \underset{N\to\infty}{\longrightarrow} p_i$, so in this case, when the number of channels is very large, the likelihood of N reads:

$$p\left(\{K_i\}_{i=1}^N\right) = b^{-N\cdot\mathbb{H}(p_1,\ldots,p_N)} \tag{43}$$

The expected number of channels is the Shannon's entropy of the set of pipes. A large entropy $H(p_1, \ldots, p_N) \to 0$ leads to a low likelihood of N, since the total discharge value remains highly uncertain, yet less likely to occur, whereas a decrease in $H(p_1, \ldots, p_N)$ translates into a more likely-to-occur configuration of N channels. In a deterministic case, only one possible configuration is possible, with probability $p\left(\{K_i\}_{i=1}^N\right) = 1$.

Moreover, any b_i follows the Binomial distribution $b_i \sim \mathcal{B}(N, p_i) = \binom{N}{b_i} p_i^{b_i}(1 -$

$p_i)^{N-b_i}$ and $\mathbb{E}_{f_i}(f_i) = \frac{1}{N}\mathbb{E}_{b_i\sim\mathcal{B}(N,p_i)}[b_i] = p_i$[7]. Therefore, Equation (43) corresponds to the likelihood for an average value of b_i, i.e., of having $p\left(\{K_i\}_{i=1}^N\right) \approx \prod_{i=1}^N p_i^{\mathbb{E}[b_i]} = \prod_{i=1}^N p_i^{Np_i}$. Random configurations of N channels with large entropy deliver a highly unknown total discharge, but they occur with low probability. Instead, "typical" configurations, with average number of channels with hydraulic conductivity K_i have low entropy and high probability to occur. The number of possible combinations (with repetitions) of values of hydraulic conductivity is $N_b = \frac{(N+b-1)!}{b!\cdot(N-1)!}$ and it correspond to the number of possible total discharge values $Q = \left(\sum_{i=1}^N f_i K_i\right)\frac{\Delta h\cdot A}{L}$. Each value Q is attained $N_Q = \frac{N!}{\prod_{i=1}^N b_i!}$ unique times with a probability that is the sum over N_Q the disjoints configurations $p_Q = N_Q \cdot p\left(\{K_i\}_{i=1}^N\right)$. If one considers the "typical" configurations, corresponding to N very very large, $p_Q = N_Q \cdot p\left(\{K_i\}_{i=1}^N\right) \approx 1$ [Clo22][8]. This implies $\frac{1}{N}\frac{\ln N_Q}{\ln b} \approx \mathbb{H}(p\left(\{K_i\}_{i=1}^N\right))$, which means that the entropy approximates - in logarithmic scale - the average number of i.i.d. configurations (or *states*) that delivers the total discharge Q, normalized with the respect to the base b and of the number of channels. In a non-"typical" configuration, $N_Q \leq b^{-N\cdot\mathbb{H}((f_1,\ldots,f_N)\|(p_1,\ldots,p_N))}$ with $\mathbb{H}((f_1,\ldots,f_N)\|(p_1,\ldots,p_N))$ representing the cross-entropy between the empirical frequencies approximating the probability distribution of K_i. The cross-entropy is defined as:

$$\mathbb{H}((p_1,\ldots,p_N)\|(q_1,\ldots,q_N)) = -\sum_{i=1}^N p_i\cdot\ln q_i \tag{44}$$

1.4.1 Why the logarithm?

Shannon's observed that, being the *uncertainty* or *surprise* associated to a random variable X inversely proportional to its probability of occurrence $P(\omega)$, deterministic events provide poor information and two events measured separately provide a total amount of information equal to the sum of the two single contributions. Based on this evidence, Shannon defined the *self-information* of an event $\omega \in (\Omega, \mathcal{E}, P)$ as a strictly decreasing monotonic function of the probability $P(X)$:

$$\mathbb{I}(\omega) = f(P(\omega)) \geq 0 \tag{45}$$

[7] $b_i \sim \mathcal{B}(N, p_i)$. The expression $g(t) = \mathbb{E}_{b_i\sim\mathcal{B}(N,p_i)}\left[e^{t\cdot b_i}\right] = \sum_{b_i=1}^N e^{t\cdot b_i}\binom{N}{b_i}p_i^{b_i}(1-p_i)^{N-b_i} = (e^t p_i + (1-p_i))^N$ allows to compute the expected value of b_i, since $g'(0) = \mathbb{E}_{b_i\sim\mathcal{B}(N,p_i)}[b_i] = N\cdot p_i$

[8] When N is very large, the Stirling's approximation applies: $N! \approx (2\pi N)^{\frac{1}{2}}\cdot N^N\cdot e^{-N}$ and $\ln b_i! = b_i\cdot\ln b_i - b_i = Np_i\cdot\ln(Nb_i) - Np_i$. With this approximation, $N_Q \approx (2\pi N)^{\frac{1}{2}} b^{N\cdot\log_b(e)\cdot(\mathbb{H}(f_1,\ldots,f_N))+\mathbb{D}_{KL}((f_1,\ldots,f_N)\|(p_1,\ldots,p_N))} = (2\pi N)^{\frac{1}{2}} p^{-\log_b(e)}(N)\cdot e^{N\mathbb{D}_{KL}((f_1,\ldots,f_N)\|(p_1,\ldots,p_N))}$ [Clo22]. More at https://michael-franke.github.io/intro-data-analysis/the-maximum-entropy-principle.html

with $f : [0,1] \mapsto [0,\infty)$ such that $\mathbb{I}(\omega) = 0$ if $P(\omega) = 1$ and $\mathbb{I}(\omega) = \infty$ if $P(\omega) = 0$. f has to be additive for two independent events $(\omega_1, \omega_2) \in \Omega^2$, i.e. $I(\omega_1 \cap \omega_2) = \mathbb{I}(\omega_1) + \mathbb{I}(\omega_2)$ with $P(\omega_1 \cap \omega_2) = P(\omega_1) \cdot P(\omega_2)$. The additive property implies that:

$$\mathbb{I}(\omega_1 \cap Y) = f(P(\omega_1 \cap \omega_2)) = f(P(\omega_1) \cdot P(\omega_2)) = \mathbb{I}(\omega_1) + \mathbb{I}(\omega_2) = f(P(\omega_1)) + f(P(\omega_2)) \tag{46}$$

Equation (46) happens to be a Cauchy's logarithmic functional equation, whose only monotone solution is in the form $f(x) = -\log_b(x) = -\frac{\ln x}{\ln b}$, with $b > 1$ (since $f : [0,1] \mapsto \infty$) and

$$\mathbb{I}(\boldsymbol{\omega}) = f(P(\boldsymbol{\omega})) = -\log_b P(\boldsymbol{\omega}) = \frac{-\ln P(\boldsymbol{\omega})}{\ln b}, \quad b > 1 \tag{47}$$

Therefore, the Shannon's entropy is the expected self-information of a random variable, quantifying how surprising the random variable is *on average*:

$$\mathbb{H}(p(\omega_1), \dots, p(\omega_N)) = \sum_{\omega_i \in \omega}^{n} p(\omega_i) \cdot I(\omega_i) = \mathbb{E}_{\omega_i \sim p_i}\left[I(\boldsymbol{\omega})\right] \tag{48}$$

1.4.2 Shannon's entropy: some fundamental property

The Shannon's entropy of random variable X can be defined as the limit (in probability sense) of the self-information of a uniform random variable. This is proven by the following theorem:

Theorem 20 (Mallat and Campagne [Cam22])**.** *If X_i are i.i.d., with probability law P_X, with a set of realization $\boldsymbol{x}_i \in \mathcal{D}_X$, with probability distribution $p(\boldsymbol{x}_i)$:*

$$\forall \epsilon > 0, \lim_{N \to \infty} \mathbb{P}\left[\left| -\frac{1}{N} \ln P(X_1, \dots, X_N) - \mathbb{H}(P(X)) \right| \leq \epsilon\right] = 1 \tag{49}$$

This result is rather fundamental. Mallat provides an insightful interpretation of Theorem 20: the self-information $-\frac{1}{N} \ln P_X(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)$ is concentrated on a surfaces defined by the equation $-\frac{1}{N} \ln P_X(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) \approx \mathbb{H}(X)$, provided an arbitrary small thickness ϵ. In machine learning practice, algorithms are trained by assuming datasets of i.i.d. samples, trying to approximate the underlying - yet unknown - entropy $\mathbb{H}(P(\boldsymbol{X}))$, associated with the true data probability distribution. However, in order to achieve a good approximation of the real entropy, several i.i.d. examples are required, so to refine the thickness ϵ.

- The joint entropy of two discrete random variables X and Y reads:

$$\begin{aligned} \mathbb{H}(p(\boldsymbol{X}, \boldsymbol{Y})) &= -\sum_{i,j=1}^{N} p(\boldsymbol{X} = \boldsymbol{x}_i, \boldsymbol{Y} = \boldsymbol{y}_j) \ln p(\boldsymbol{X} = \boldsymbol{x}_i, \boldsymbol{Y} = \boldsymbol{y}_j) = \\ &= -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim p}\left[\ln(p(\boldsymbol{X}, \boldsymbol{Y}))\right] \end{aligned} \tag{50}$$

- The conditional entropy of two discrete random variables X and Y reads:

$$\begin{aligned}\mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}_i)) &= -\sum_{j=1}^{N} p(\boldsymbol{Y}=\boldsymbol{y}_j|\boldsymbol{X}=\boldsymbol{x}_i)\ln p(\boldsymbol{Y}=y_j|\boldsymbol{X}=\boldsymbol{x}_i) = \\ &= -\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\left[\ln(p(\boldsymbol{Y}|X=\boldsymbol{x}_i))\right]\end{aligned} \tag{51}$$

and

$$\begin{aligned}&\mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X})) = -\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathbb{H}(\boldsymbol{p}(Y|\boldsymbol{X}=\boldsymbol{x}))\right] = \\ &= -\sum_{i,j=1}^{N} p(\boldsymbol{Y}=\boldsymbol{y}_j|\boldsymbol{X}=\boldsymbol{x}_i)p(X=\boldsymbol{x}_i)\ln \boldsymbol{p}(\boldsymbol{Y}=\boldsymbol{y}_j|\boldsymbol{X}=\boldsymbol{x}_i) = \\ &= -\sum_{i,j=1}^{N} p(\boldsymbol{X}=\boldsymbol{x}_i,\boldsymbol{Y}=y_j)\ln p(\boldsymbol{Y}=\boldsymbol{y}_j|\boldsymbol{X}=\boldsymbol{x}_i) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p}\left[\ln p(\boldsymbol{Y}|\boldsymbol{X})\right]\end{aligned} \tag{52}$$

-

$$\mathbb{H}(p(\boldsymbol{X},\boldsymbol{Y})) = \mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X})) + \mathbb{H}(p(\boldsymbol{X})) = \mathbb{H}(p(\boldsymbol{X}|\boldsymbol{Y})) + \mathbb{H}(p(\boldsymbol{Y})) \tag{53}$$

- The Kullback-Leibler distance $\mathbb{D}_{KL}(p\|q)$ has the following properties:
  - $\mathbb{D}_{KL}(p\|q) \geq 0$, since, due to the concavity of $\ln(x)$, $\ln(y) - \ln(x) \leq (\ln x)' \cdot (y - x)$ and setting $x = p$ and $y = q$ the following expression holds:
$$\mathbb{E}_p\left[\ln\frac{q}{p}\right] = -\mathbb{D}_{KL}(p\|q) \leq \mathbb{E}_p\left[\frac{q-p}{p}\right] = 0$$
  - $\mathbb{D}_{KL}(p\|q) = 0$ implies that $p \equiv q$ if and only if
$$\operatorname{supp}(p) \cap \operatorname{supp}(q) \neq \emptyset$$
Otherwise, when p and q have two disjoint supports, it often occurs that $\mathbb{D}_{KL}(p\|q) = 0$ or $\mathbb{D}_{KL}(p\|q) = 0 \to \infty$. This pathological situation occurs quite often in practice, especially for genereative algorithms such as GAN (see [Goo+14]).
  - The Kullback-Leibler distance between $p(\boldsymbol{X},\boldsymbol{Y})$ and $p(\boldsymbol{X})\cdot p(\boldsymbol{Y})$ is called *mutual information* $\mathbb{I}(X,Y)$ [Che+16; Cam22]:

$$\begin{aligned}\mathbb{I}(p(\boldsymbol{X},\boldsymbol{Y})) &= \mathbb{H}(p(\boldsymbol{Y})) - \mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X})) = \\ &= \mathbb{H}(p(\boldsymbol{X})) - \mathbb{H}(p(\boldsymbol{X}|\boldsymbol{Y})) = \mathbb{D}_{KL}(p(\boldsymbol{X},\boldsymbol{Y})\|p(\boldsymbol{X})\cdot p(\boldsymbol{Y}))\end{aligned} \tag{54}$$

The proof is straightforward, since $\mathbb{H}(p(\boldsymbol{X},\boldsymbol{Y})) = \mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X})) + \mathbb{H}(p(\boldsymbol{X})) = \mathbb{H}(p(\boldsymbol{X}|\boldsymbol{Y})) + \mathbb{H}(p(\boldsymbol{Y})) \geq 0$ and

$$\mathbb{I}(p(\boldsymbol{X},\boldsymbol{Y})) = -\mathbb{H}(p(\boldsymbol{X},\boldsymbol{Y})) + \mathbb{H}(p(\boldsymbol{X})) + \mathbb{H}(p(\boldsymbol{Y})) =$$

$$= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p}\left[\frac{p(\boldsymbol{X},\boldsymbol{Y})}{p(\boldsymbol{X})\cdot p(\boldsymbol{Y})}\right] = \mathbb{D}_{KL}\left(p(\boldsymbol{X},\boldsymbol{Y})\|p(\boldsymbol{X})\cdot p(\boldsymbol{Y})\right) \geq 0$$

- $\mathbb{I}(p(\boldsymbol{X},\boldsymbol{Y}))$ is strictly connected to the notion of *self-information* presented in Section 1.4.1. $\mathbb{I}(p(\boldsymbol{X},\boldsymbol{Y}))$ if $\boldsymbol{X},\boldsymbol{Y}$ are two independent events.
- The notion of mutual information is widely used in practice. Maximizing $\mathbb{I}(p(\boldsymbol{X},\boldsymbol{Y}))$ at constant $\mathbb{H}(p(\boldsymbol{X}))$ requires $\mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X})) \to 0$, i.e., forcing the *disentanglement* between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [Che+16].
- The mutual information enters in one of the alternative definitions of the ELBO, in Equation (88):

$$\begin{aligned}\text{ELBO} = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{Z}|\boldsymbol{x})}\left[\ln p_{\boldsymbol{\theta}}(\boldsymbol{Z}|\boldsymbol{X})\right]\right] - \\ - \mathbb{D}_{KL}\left(p_{\boldsymbol{\theta}}(\boldsymbol{Z})\|q_{\boldsymbol{\phi}}(\boldsymbol{Z})\right) - \mathbb{I}(\boldsymbol{Z},\boldsymbol{X})\end{aligned} \tag{55}$$

  Maximize the ELBO demands to minimize $\mathbb{I}(\boldsymbol{Z},\boldsymbol{X}) = \mathbb{H}(q_{\boldsymbol{\phi}}(\boldsymbol{X})) - \mathbb{H}(q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{Z}))$ [Mak19], i.e. find a *joint data distribution* $q_{\boldsymbol{\phi}}(\boldsymbol{X},\boldsymbol{Z})$ "highly generative" , i.e. with a large conditional entropy $\mathbb{H}(q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{Z}))$. In this sens, the statistical model associate to $q_{\boldsymbol{\phi}}$ can widely span the dimensionality of whole dataset $\mathcal{X}$.

1.4.3 From discrete to continuous

The definition of entropy in Equation (37) can be easily extended to continuous random variables with $\mathcal{X}$ isomorphic to $\mathbb{R}^{d_X}$. In this case, it is called *differential entropy* [Cam22]:

$$\mathbb{H}_d(p(\boldsymbol{X})) = -\int_{\mathcal{X}} p(\boldsymbol{x})\cdot\ln(p(\boldsymbol{x}))\cdot\mu(d\boldsymbol{x}) \tag{56}$$

Compared to $\mathbb{H}(p(\boldsymbol{X}))$, $\mathbb{H}_d(p(\boldsymbol{X}))$ is not always positive. For instance, $\mathbb{H}_d\left[\mathcal{U}([0,b])\right] = \log b < 0$ if $b < 1$. The differential entropy of a multivariate normal distribution reads:

$$\begin{aligned}\mathbb{H}_d(\mathcal{N}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\boldsymbol{X})) = -\ln\left((2\pi)^{-\frac{n}{2}}\det(\boldsymbol{\Sigma})^{-\frac{1}{2}}\right)\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}}\left[(\boldsymbol{x}-\boldsymbol{\mu})\otimes\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] = \\ = \frac{d_X}{2}\ln 2\pi + \frac{1}{2}\text{Tr}(\boldsymbol{\Sigma})\end{aligned} \tag{57}$$

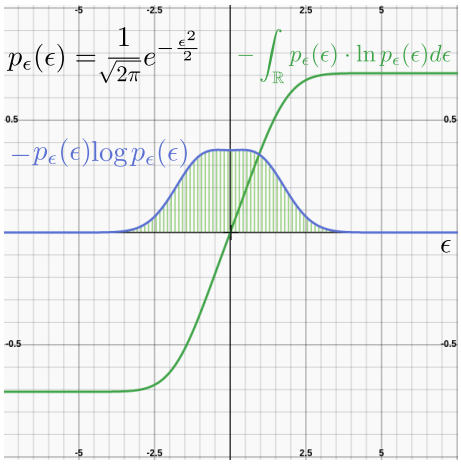

Figure 4: Differential entropy $\mathbb{H}_d(p_\epsilon)$ of a normal distribution $p_\epsilon = \mathcal{N}(0, 1)$.

The differential entropy of a normal distribution increases with the dimension of the data space d_X.

Moreover, $\mathrm{Tr}(\boldsymbol{\Sigma})$ is dominated by the spectral radius $\rho(\boldsymbol{\Sigma}) = \sup_{\lambda \in \sigma(\boldsymbol{\Sigma})} |\lambda|$. In other words, the surprise related to a random multivariate normal variables increase with the dimension of the space in which the variable lives and with the correlation between its components $\boldsymbol{\Sigma}$. The factor $\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma})$ represents a factor of "scale", since $\mathbb{H}_d(b\boldsymbol{X}) = \mathbb{H}_d(\boldsymbol{X}) + \ln b$, in analogy with what presented for its discrete counterpart in Section 1.4.

1.4.4 Typical set

In the discrete case, the approximation evoked in the Section 1.4.2 for large number of observations holds because of the well-know law of large numbers (the weak form in Theorem 3), which implies that the estimated mean of a set of i.i.d. variables with the same average converges in probability to the true mean. In other words, Equation (43) is formally expressed as (see [Cam22]):

$$\lim_{N \to +\infty} \mathbb{P}\left[\left|\frac{1}{N} \log_b p\left(\{K_i\}_{i=1}^N\right) + \mathbb{H}(p_1, \ldots, p_N)\right| \leq \varepsilon\right] = 1, \quad \forall \varepsilon > 0 \tag{58}$$

Equation (58) defines the "typical" set as:

$$\begin{aligned} A_\varepsilon^{(N)} = \big\{&(K_1, K_2, \ldots, K_N) \in \mathcal{D}_K^N | \\ &\mathbb{P}\left[\left|\frac{1}{N} \log_b p\left(\{K_i\}_{i=1}^N\right) + \mathbb{H}(p_1, \ldots, p_N)\right| > 1 - \varepsilon\right]\big\} \end{aligned} \tag{59}$$

Equation (58) represents the extension of the typical set to a large number of observations. In this case, almost all the observations fall belong to the typical set, $\forall \varepsilon > 0$. Therefore, the concept of typical set allows defining the notion on Shannon's entropy for a discrete variable as well as to provide the lower and upper bounds of its probability distribution that read:

$$b^{-N(\mathbb{H}(p)+\varepsilon)} \leq p\left(\{K_i\}_{i=1}^N\right) \leq b^{-N(\mathbb{H}(p)-\varepsilon)} \tag{60}$$

Equation (60) provided that for the typical set, the additivity of the entropy function allows to write $N\mathbb{H}(p) = \mathbb{H}(p_1, p_2, \ldots, p_N)$, which does not depend on the realization itself. Moreover, the observations reach the *asymptotic equipartition* of states, by randomly occupying the typical set. Moreover, the inequality in Equation (60) allows to compute the cardinality of the typical set, since:

$$\begin{aligned}
1 = & \sum_{\{K_i\}_{i=1}^N \in \mathcal{D}_K^N} p\left(\{K_i\}_{i=1}^N\right) \geq \sum_{\{K_i\}_{i=1}^N \in A_\varepsilon^{(N)}} p\left(\{K_i\}_{i=1}^N\right) \geq \\
\geq & \sum_{\{K_i\}_{i=1}^N \in A_\varepsilon^{(N)}} b^{-N(\mathbb{H}(p)+\varepsilon)} = \operatorname{card}\left(A_\varepsilon^{(N)}\right) b^{-N(\mathbb{H}(p)+\varepsilon)} \\
1 - \varepsilon \leq & \sum_{\{K_i\}_{i=1}^N \in A_\varepsilon^{(N)}} p\left(\{K_i\}_{i=1}^N\right) \leq \\
\leq & \sum_{\{K_i\}_{i=1}^N \in A_\varepsilon^{(N)}} p\left(\{K_i\}_{i=1}^N\right) b^{-N(\mathbb{H}(p)-\varepsilon)} = \operatorname{card}\left(A_\varepsilon^{(N)}\right) b^{-N(\mathbb{H}(p)-\varepsilon)}
\end{aligned} \tag{61}$$

In the continuous case, the equivalent of Equation (58) reads:

$$\lim_{N \to +\infty} \mathbb{P}\left[\left|\frac{1}{N} \ln p\left(\{K_i\}_{i=1}^N\right) + \mathbb{H}_d\left(p\left(\boldsymbol{X}\right)\right)\right| \leq \varepsilon\right] = 1, \quad \forall \varepsilon > 0 \tag{62}$$

since:

$$\lim_{N \to +\infty} \frac{1}{N} \ln p\left(\{K_i\}_{i=1}^N\right) = \mathbb{E}_{\boldsymbol{k} \sim p}\left[\ln p\left(\boldsymbol{X}\right)\right] \tag{63}$$

However, the number of elements in the typical set is infinite, in the continuous case. Therefore, one should define the typical volume $\Omega\left(A_\varepsilon^{(N)}\right) = \int_{A_\varepsilon^{(N)}} d\boldsymbol{x}$, in the Lebesgue measure, bounded as follows (see Equation (61), with $b = 2$):

$$(1 - \varepsilon) \cdot 2^{N(\mathbb{H}_d(K)-\varepsilon)} \leq \Omega\left(A_\varepsilon^{(N)}\right) \leq 2^{N(\mathbb{H}_d(K)+\varepsilon)} \tag{64}$$

since

$$\begin{aligned}
1 = \int_{\Omega_K} p\left(k\right) dk \geq \int_{A_\varepsilon^{(N)}} p\left(k\right) dk \geq 2^{-N(\mathbb{H}_d(p)+\varepsilon)} \\
1 - \varepsilon \leq \int_{A_\varepsilon^{(N)}} p\left(k\right) dk \leq 2^{-N(\mathbb{H}_d(p)-\varepsilon)} \cdot \Omega\left(A_\varepsilon^{(N)}\right)
\end{aligned} \tag{65}$$

Therefore, thanks to Equation (65), one assumes that the typical volume has a size $\approx \frac{1}{p(K)}$, the probability being constant regardless the realization and close to the inverse of the size of the volume set (uniform), and that the differential entropy $\mathbb{H}_d(p)$ can be seen as the logarithm to the base 2 of the characteristic N-dimensional length of the typical volume tessellation.

1.5 The principle of maximum entropy (MaxEnt)

Any statistical model conceived to infer a physical problem must be compatible with the available observations $(\boldsymbol{x}_k, y_k)$. The latter serve as verification baseline. In particular, the easiest strategy to calibrate a statistical model is to match the average observations $\mu_{y_k}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}\sim p}[y_k(\boldsymbol{x})]$, with $p(\boldsymbol{x})$ being the true (yet unknown) data probability distribution. Any statistical model $\boldsymbol{h}_{\boldsymbol{\theta}}$ formulated to infer new samples of the quantity of interest y, is associate to a parametric probability distribution $p_{\boldsymbol{\theta}}$, whose differential entropy must be compatible with the constraints imposed by the observations. However, these constraints on the observations are not sufficient to approximate the "true" probability distribution p, based on the statistical model $p_{\boldsymbol{\theta}}$. According to [Jay57], $p_{\boldsymbol{\theta}}$ must maximize its differential entropy, so that the typical volume is the largest as possible and the probability distribution on it is the most uniform possible. This framework is resolved by the following theorem [Cam22]:

Theorem 21. ***Gibbs-Boltzmann theorem***
Provided a family of probability distributions $q \in \mathcal{H}$, the solution to the following constrained optimization problem:

$$\begin{cases} \max\limits_{q\in\mathcal{H}} \mathbb{H}_d(q) = \max\limits_{q} \left(-\int_{\Omega_X} \ln q(\boldsymbol{x}) \cdot q(\boldsymbol{x})\, d\boldsymbol{x} \right) & (66) \\ c_k(q) = \mu_{y_k} - \int_{\Omega_X} y_k(\boldsymbol{x}) \cdot q(\boldsymbol{x})\, d\boldsymbol{x} = 0, \quad c_k : \mathbb{R}^n \to \mathbb{R}, 1 \le k \le K & (67) \end{cases}$$

if it exists it is defined as $p_{\boldsymbol{\theta}} \in \mathcal{H}$, which reads:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{q\in\mathcal{H}, \mathbf{c}:\mathbf{c}=\mathbf{0}}{\arg\max}\ \mathbb{H}_d(q) = \frac{e^{-\sum_{k=1}^{K} y_k(\boldsymbol{x})\cdot\theta_k}}{Z}$$

with Z a normalization constant. Moreover, $\mathbb{H}_d\left(p_{\boldsymbol{\theta}}\right) \ge \mathbb{H}_d(p)$, $\forall \boldsymbol{\theta} \in \mathcal{H}_{\boldsymbol{\theta}} \subset \mathcal{H}$ and if $p_{\boldsymbol{\theta}} = p$, then $\mathbb{H}_d\left(p_{\boldsymbol{\theta}}\right) = \mathbb{H}_d(p)$

Proof. The proof of the Gibbs-Boltzmann theorem is proven thanks to the Kuhn-Tucker theorem. The set

$$\mathcal{K} = \left\{ q | c(q) = 0, c_k \in \mathcal{C}^1\left(\mathbb{R}^n\right), \forall k = 0, \ldots, K \right\}$$

represents the set of continuous and differentiable constraint functions. If the derivatives $\frac{dc_k}{dq}$ are linearly independent, and if $\exists w \in \mathbb{R}$ such that its Gâteaux derivative[9] along w satisfies the expression:

$$D_w c_k(q) = 0, k = 0, \ldots, K$$

the constraints are "qualified" (or active). In this context, the Kuhn-Tucker theorem grants the existence of a probability distribution $p_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}} \cap \mathcal{K}$ and of a set of coefficients $\lambda_1, \ldots, \lambda_K$, with $\lambda_i \in \mathbb{R}$, for which:

$$D_q \mathbb{H}_d(p_{\boldsymbol{\theta}}) + \sum_{k=0}^{K} D_q c_k(p_{\boldsymbol{\theta}}) = 0, \quad \forall q \in \mathcal{K} \tag{68}$$

The result of Equation (68) can be recast into the condition for which the Lagrangian function defined as:

$$\mathrm{L}(q, \boldsymbol{\lambda}) = \mathbb{H}_d(q) + \sum_{k=1}^{K} \lambda_k \cdot c(q) + \lambda_0 \left(\int_{\Omega_X} q(\boldsymbol{x}) \cdot \mu(d\boldsymbol{x}) - 1 \right)$$

$p_{\boldsymbol{\theta}}$ represents a stationary point of the Lagrangian functional, since the Kuhn-Tucker theorem in Equation (68) can be reformulated as:

$$\begin{aligned} D_q \mathrm{L}(p_{\boldsymbol{\theta}}, \boldsymbol{\lambda}) &= \frac{\partial \mathrm{L}}{\partial t}(p_{\boldsymbol{\theta}} + tq, \boldsymbol{\lambda})\Big|_{t=0} = \\ &= -\int_{\Omega_X} q(\boldsymbol{x}) \cdot (\ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1) \cdot \mu(d\boldsymbol{x}) - \\ &- \sum_{k=1}^{K} \lambda_k \cdot \int_{\Omega_X} y_k(\boldsymbol{x}) \cdot q(\boldsymbol{x}) \cdot \mu(d\boldsymbol{x}) + \lambda_0 = 0, \quad \forall q \in \mathcal{K} \end{aligned} \tag{69}$$

which leads to:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\lambda}}(\boldsymbol{x}) = e^{\lambda_0 - 1} \cdot e^{-\sum_{k=1}^{K} \lambda_k \cdot y_k(\boldsymbol{x})} \tag{70}$$

Equation (70) implies that the penalty coefficients λ_i are the parameters that structure the space $\mathcal{H}_{\boldsymbol{\theta}}$. However, the definition of $p_{\boldsymbol{\theta}}$ in Equation (70) needs to be normalized, in order to assure that $\int_{\Omega_X} p_{\boldsymbol{\theta}}(\boldsymbol{x}) \mu(d\boldsymbol{x}) = 1$. This is achieved if:

$$\begin{aligned} &\int_{\Omega_X} e^{\lambda_0 - 1} \cdot e^{-\sum_{k=1}^{K} \lambda_k \cdot y_k(\boldsymbol{x})} \cdot \mu(d\boldsymbol{x}) = 1 \iff \\ \iff Z = e^{1-\lambda_0} &= \int_{\Omega_X} e^{-\sum_{k=1}^{K} \lambda_k \cdot y_k(\boldsymbol{x})} \cdot \mu(d\boldsymbol{x}) \end{aligned} \tag{71}$$

[9] $$D_w f(q) = \frac{df(q + tw)}{dt}\Big|_{t=0}$$

The stationarity of the Lagrangian function also implies that $p_{\boldsymbol{\theta}}$ must respect the constraint and force the expected value to converge to the "true" one ($\boldsymbol{\mu}$):

$$\nabla_{\lambda} \mathrm{L}\left(p_{\boldsymbol{\theta}}, \boldsymbol{\lambda}\right) = \mathbf{0} \Rightarrow \int_{\Omega_X} y_k \cdot p\left(\boldsymbol{x}\right) \mu(d\boldsymbol{x}) = \mu_{y_k}\left(\boldsymbol{x}\right) = \int_{\Omega_X} y_k \cdot p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \mu(d\boldsymbol{x}) \tag{72}$$

□

The family of exponential probability distributions Theorem 21 states that the probability distribution that maximizes the entropy belongs to the exponential family. Moreover, the probability distribution that maximizes the entropy allows to describe the log-likelihood function as:

$$\ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = -\ln Z - \sum_{k=1}^{K} y_k\left(\boldsymbol{x}\right) \cdot \theta_k = -\ln Z - \left\langle \boldsymbol{y}\left(\boldsymbol{x}\right), \boldsymbol{\theta} \right\rangle \tag{73}$$

The gradient of the term $-\ln Z$ reads:

$$-\nabla_{\boldsymbol{\theta}} \ln Z = \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] \tag{74}$$

which implies the gradient of log-likelihood function reads:

$$\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] - \boldsymbol{y}\left(\boldsymbol{x}\right) \tag{75}$$

and the following property:

$$\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \cdot \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \left(\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] - \boldsymbol{y}\left(\boldsymbol{x}\right)\right) \cdot p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \tag{76}$$

Moreover, the Fisher's information computed as the mean Hessian in Equation (28) corresponds to the covariance of the observations $y_k\left(\boldsymbol{x}\right)$ if $p_{\boldsymbol{\theta}}$ maximizes the entropy [Cam22]:

$$\begin{aligned} &\nabla_{\boldsymbol{\theta}} \otimes \nabla_{\boldsymbol{\theta}}\left(\ln p_{\boldsymbol{\theta}}\right) = \int_{\Omega_X} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \otimes \boldsymbol{y}\left(\boldsymbol{x}\right) \mu\left(d\boldsymbol{x}\right) = \\ &= \int_{\Omega_X} \left(\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] - \boldsymbol{y}\left(\boldsymbol{x}\right)\right) \otimes \boldsymbol{y}\left(\boldsymbol{x}\right) \cdot p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \cdot \mu\left(d\boldsymbol{x}\right) = \\ &= \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] \otimes \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] - \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right) \otimes \boldsymbol{y}\left(\boldsymbol{X}\right)\right] = \\ &= -\mathbb{C}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}}\left(\boldsymbol{y}\left(\boldsymbol{X}\right)\right) \end{aligned} \tag{77}$$

Remark 22. The interesting aspect of Equation (77) is that it depends only on the information provided by the observations $y_k\left(\boldsymbol{x}\right)$. In the case of neural networks, y_k depends on the weights themselves though, since $y_k = h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_k\right)$ [Cam22]. The maximum entropy principle has a strong analogy with the Fisher information theory, since the value of the "best" estimator $\hat{\boldsymbol{\theta}}$ that maximizes the

log-likelihood is unbiased (see Equation (30) for the definition). Equation (75) adds the fact that $\hat{\boldsymbol{\theta}}$ makes the observations $\boldsymbol{y}(\boldsymbol{X})$ converge towards their average values, that maximize the log-likelihood. Moreover, for a sufficiently large dataset $N \to +\infty$, the MLE $\hat{\boldsymbol{\theta}}_N$ tends to be normally distributed, i.e., its probability distribution tends to belong to the exponential family. In other words, the MLE not only maximizes the (log-)likelihood but also the entropy of the underlying dataset. In other words, the maximum entropy of a statistical model is reached by maximizing the log-likelihood of an exponential probability distribution [Cam22].

Remark 23. The quest for the best estimator can be seen as a minimax optimization problem. The optimization targets the approximation of the unknown differential entropy lower bound $\mathbb{H}_d(p)$ represented by the "true" data probability distribution ($\mathbb{H}_d\left(p_{\boldsymbol{\theta}}\right) \geq \mathbb{H}_d(p)$). In order to achieve this lower bound, a statistical model $\mathcal{H}_{\boldsymbol{\theta}}$ is chosen (the choice of the architecture of the $\mathcal{NN}$) among all the possible statistical models $\mathcal{H}$. By adjusting the parameters $\boldsymbol{\theta}$, the parametric probability distribution $p_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}$ induced by the statistical model must realize the maximum of its differential entropy and comply with the observations (see Theorem 21). To summarize, the optimization task that neural networks try to accomplish is expressed by the following minimax problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta \in \Theta} \max_{p_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}} \mathbb{H}_d\left(p_{\boldsymbol{\theta}}\left(Y|\boldsymbol{X}\right)\right) \tag{78}$$

1.6 Learning with variational inference and reconstruction

In supervised learning, the dataset is labeled, i.e.

$$\mathcal{D}_{XY} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{N}$$

The learning task is therefore to approximate the conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$ with a statistical model $\mathcal{H}_{\boldsymbol{\theta}}$. In this case the "best" approximation can be found by minimizing the conditional entropy as follows:

$$\hat{\boldsymbol{\theta}}(\mathcal{D}_{XY}) = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{(x,y) \in \mathcal{D}_{XY}} \ln \frac{1}{p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{X})} \tag{79}$$

The first order gradient method adopted to minimize the loss function reads (see Equation (63)):

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta^{(i)} \boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right), \quad \eta^{(i)} \in \mathbb{R}^{+}$$

The analogy with the concept of MaxEnt and MLE exposed in Remark 22, in order to minimize the negative log-likelihood $\mathcal{NLL}$ (equivalent to maximize

the (log-)likelihood, as expressed by Equation (20)) the MLE can be found according to the following recursive formula [Cam22]:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta^{(i)} \left(\boldsymbol{\mu}_{\boldsymbol{y}} - \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^{(i)}}} \left[\boldsymbol{y} \left(\boldsymbol{X} \right) \right] \right), \quad \eta^{(i)} \in \mathbb{R}^{+} \tag{80}$$

By identification,

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} \right) = \boldsymbol{\mu}_{\boldsymbol{y}} - \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^{(i)}}} \left[\boldsymbol{y} \left(\boldsymbol{X} \right) \right]$$

The recursive algorithm in Equation (80) ends when the entropy of the parametric probability distribution is close to its maximum value. However, in machine learning applications, the dataset at stake is sampled from an unknown probability distribution $p\left(\boldsymbol{x}\right)$. There is no certainty that $p \in \mathcal{H}_{\boldsymbol{\theta}}$. This implies that computing $\boldsymbol{\mu}_{\boldsymbol{y}} = \mathbb{E}_{\boldsymbol{x} \sim p} \left[\boldsymbol{y} \left(\boldsymbol{X} \right) \right]$ is rather intricate since $\boldsymbol{\theta}^{*} \in \Theta$ such that $\boldsymbol{x} \sim p = p_{\boldsymbol{\theta}^{*}}$ may not exist. $\boldsymbol{\mu}_{\boldsymbol{y}}$ could be estimated via its sample average, provided that enough observations are available (curse of dimensionality). As far as the computation of $\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^{(i)}}} \left[\boldsymbol{y} \left(\boldsymbol{X} \right) \right]$ goes, the parametric distribution induced by the set of weights $p_{\boldsymbol{\theta}^{(i)}}$ is known, which pave the way to the use of the family of cumbersome, yet effective, Monte Carlo methods (Importance Sampling, Metropolis-Hastings, Gibbs Sampling, Markov chains,...). Neural networks follow another strategy: a $\mathcal{NN}$ is conceived to estimate $\boldsymbol{y}\left(\boldsymbol{x}\right) = \boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$ and its statistics, depending on the weights themselves.
In unsupervised learning, in order to maximize the likelihood of the statistical model, the common approach is to consider a *latent manifold* $(\mathcal{Z}, \mathcal{E}_{Z}, P_{Z})$ with a arbitrary yet unknown probability distribution, that represents an encoded version of the data at stake [Mak19]. The joint probability distribution $p(\boldsymbol{X}, \boldsymbol{Z})$ is unknown but it can be approximated by choosing an arbitrary distribution $p(Z)$ (often Gaussian or Uniform, for the sake of simplicity) and identify a *joint model distribution* that reads

$$p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Z}) = p_{\boldsymbol{\theta}}\left(\boldsymbol{X} | \boldsymbol{Z}\right) \cdot p(\boldsymbol{Z}) \tag{81}$$

The *aggregate prior distribution*:

$$p_{\boldsymbol{\theta}}(\boldsymbol{X}) = \int_{\mathcal{Z}} p_{\boldsymbol{\theta}}\left(\boldsymbol{X} | \boldsymbol{z}\right) \cdot \mu(d\boldsymbol{z}) \tag{82}$$

corresponds to the likelihood in Equation (8), that is maximized to find the best $\boldsymbol{\theta} \in \Theta$ provided an arbitrary generative distribution $p(\boldsymbol{Z})$. However, one can also identify a *joint data distribution* as follows:

$$q_{\phi}(\boldsymbol{X}, \boldsymbol{Z}) = q_{\phi}(\boldsymbol{Z} | \boldsymbol{X}) \cdot p(\boldsymbol{X}) \tag{83}$$

The *aggregate posterior distribution* is identified by:

$$q_{\phi}(\boldsymbol{Z}) = \int_{\mathcal{X}} q_{\phi}(\boldsymbol{Z} | \boldsymbol{x}) \cdot p(\boldsymbol{x}) \mu(d\boldsymbol{x}) \tag{84}$$

The maximum likelihood matching consists into discover $p(X)$ by matching it with $p_{\boldsymbol{\theta}}(\boldsymbol{X})$. Learning the real data distribution with *variational inference* means to match $p_{\boldsymbol{\theta}}(X, Z)$ and $q_\phi(X, Z)$.
Finally, one can define the *joint reconstruction distribution* as:

$$r_{(\phi;\theta)}(\boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{X}|\boldsymbol{Z}) \cdot q(\boldsymbol{Z}) \neq q_\phi(\boldsymbol{X}, \boldsymbol{Z}) \neq p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Z}) \tag{85}$$

and its *aggregate reconstruction distribution* [Mak19] as:

$$r_{(\phi;\theta)}(\boldsymbol{X}) = \int_{\mathcal{X}} r(\boldsymbol{X}, \boldsymbol{z}) \cdot \mu(d\boldsymbol{z}) \tag{86}$$

There are different ways to maximize the log-likelihood. The most intuitive one is the maximize its *variational lower bound* or *Evidence Lower BOund* (ELBO), as proposed by Kingma and Welling [KW22]. The authors stated the following inequality[10]:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln p_{\boldsymbol{\theta}}(\boldsymbol{X})\right] \geq \\
&\quad \geq \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{Z}|\boldsymbol{x})}\left[\ln p_{\boldsymbol{\theta}}(\boldsymbol{Z}|\boldsymbol{X})\right]\right] - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{D}_{KL}\left(q_\phi(\boldsymbol{Z}|\boldsymbol{X})\|p_{\boldsymbol{\theta}}(\boldsymbol{Z})\right)\right] = \text{ELBO}
\end{aligned} \tag{87}$$

According to Equation (87), maximizing the ELBO maximizes the log-likelihood, towards $p_{\boldsymbol{\theta}}(\boldsymbol{X})$ matching the real data distribution $p(\boldsymbol{X})$. In order to maximize the ELBO, one needs to:

- Make the posterior $q_\phi(\boldsymbol{Z}|\boldsymbol{X})$ match the arbitrary probability distribution on the latent manifold $p_{\boldsymbol{\theta}}(\boldsymbol{Z})$. In this way, the negative term in ELBO represented by the Kullback-Leibler distance between the two above mentioned distributions $\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{Z}|\boldsymbol{x})}\left[-\ln p_{\boldsymbol{\theta}}(\boldsymbol{Z}|\boldsymbol{X})\right]\right]$ goes to zero;
- Maximize the log-likelihood of the conditional distribution $p_{\boldsymbol{\theta}}(X|Z)$

For both tasks, the learning algorithm has to perform the maximization "in average" over the available dataset. However, the ELBO maximization can be reformulated in other ways [Mak19], such as :

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{Z}|\boldsymbol{x})}\left[\ln p_{\boldsymbol{\theta}}(\boldsymbol{Z}|\boldsymbol{X})\right]\right] - \mathbb{D}_{KL}\left(p_{\boldsymbol{\theta}}(\boldsymbol{Z})\|q_\phi(Z)\right) - \mathbb{I}(\boldsymbol{Z}, \boldsymbol{X}) = \\
&= \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\theta}}(\boldsymbol{Z})}\left[\mathbb{D}_{KL}\left(q_\phi(\boldsymbol{X}|\boldsymbol{Z})\|p_{\boldsymbol{\theta}}(X|\boldsymbol{Z})\right)\right] - \mathbb{D}_{KL}\left(q_\phi(\boldsymbol{Z})\|p_{\boldsymbol{\theta}}(Z)\right) - \mathbb{H}(p(\boldsymbol{X})) = \\
&= -\mathbb{D}_{KL}\left(q_\phi(\boldsymbol{X}, \boldsymbol{Z})\|r_{\phi;\theta}(\boldsymbol{X}, \boldsymbol{Z})\right) - \mathbb{D}_{KL}\left(q_\phi(\boldsymbol{Z})\|p_{\boldsymbol{\theta}}(\boldsymbol{Z})\right) - \mathbb{H}(p(\boldsymbol{X}))
\end{aligned} \tag{88}$$

[10] With some mathematical intricacy, one can prove that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln p_{\boldsymbol{\theta}}(X)\right] = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{D}_{KL}\left(q_\phi(Z|X)\|p_{\boldsymbol{\theta}}(Z|X)\right)\right] - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{E}_{z\sim q_\phi(Z|x)}\left[-\ln p_{\boldsymbol{\theta}}(Z|X)\right]\right] - \\
- \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathbb{D}_{KL}\left(q_\phi(Z|X)\|p_{\boldsymbol{\theta}}(Z)\right)\right]
\end{aligned}$$

Equation (88) implies the notion of Shannon's entropy $\mathbb{H}(p(\boldsymbol{X}))$ of the real data distribution and mutual information $\mathbb{I}(\boldsymbol{Z}, \boldsymbol{X})$. Those two concepts will be introduced in the following subsections.

2 Theoretical aspects of Multi-Layer Perceptrons

2.1 Why the $\mathcal{MLP}$?

Historically, the $\mathcal{MLP}$ has been largely studied (see for instance [HDD; Cyb89; Hec92; Les+93; Pin99]) because of the following fundamental result: the universal approximation theorem.

Theorem 24. *Universal approximation theorem for a 1-hidden-layered perceptron [Cam19]*
Given a function $g \in \mathcal{C}(\mathbb{R})$ and a class of function

$$\mathcal{H}_g := span\left\{g(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b), \boldsymbol{w} \in \mathbb{R}^{d_X}, b \in \mathbb{R}\right\}$$

, then $\mathcal{H}_g$ is dense for the uniform convergence on compact space, iff g is not polynomial.

This theorem implies that a $\mathcal{MLP}$ $h_\theta \in \mathcal{H}_g$, with one hidden layer, can approximate any function $f : \mathbb{R}^{d_X} \mapsto \mathbb{R}$ if g is not polynomial. In this case, Theorem 24 states that $\mathcal{H}_g$ is dense in $\mathcal{C}\left(\mathbb{R}^{d_X}\right)$ [Cam19], i.e. that

Theorem 25. $\forall \varepsilon > 0, \exists h_\theta \in \mathcal{H}_g$ *such that* $\forall \mathcal{X}_\square \subset \mathbb{R}^{d_X}$ *compact set,* $\forall \boldsymbol{x} \in \mathbb{R}^{d_X} \quad |f(\boldsymbol{x}) - h_\theta(\boldsymbol{x})| \leq \varepsilon$

For the full proof of this fundamental theorem, we refer to [Cam19]. However, the proof's outline unveils some interesting aspects related to the $\mathcal{MLP}$ architecture. In particular, the proof relies on the following fundamental lemma:

Lemma 26. *Approximation with Fourier basis [Cam19]*
Any function $f \in \mathcal{C}(\mathbb{R}^{d_X})$ can be approximated on any compact set of $\mathbb{R}^{d_X}$, with an arbitrary precision ε, with a Fourier series. In other words:

$$\begin{aligned} &\forall f \in C(\mathbb{R}^{d_X}), \quad \forall \text{ compact set } \mathcal{X}_\square \subset \mathbb{R}^{d_X}, \quad \forall \varepsilon > 0, \exists N \in \mathbb{N}, \\ &\exists \{\boldsymbol{w}_n\}_{n=1}^N \in \mathbb{R}^{d_X} \text{ such that } \quad \forall \boldsymbol{x} \in \mathcal{X}_\square : \\ &\left| f(\boldsymbol{x}) - \sum_{n=1}^{N} \left(a_n \cdot \cos\left(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle\right) + b_n \cdot \sin\left(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle\right)\right) \right| \leq \varepsilon \end{aligned} \tag{89}$$

The proof is provided by Equation (210). The Fourier theory assures that Equation (89) is verified whenever $f \in C^\infty(\mathbb{R}^\infty)$. But this is a rather strong restriction for practical purposes. As outlined in the following, based on Lemma 26,

the Theorem 25 adopts a non-polynomial approximation of the harmonic functions $\cos(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle)$ and $\sin(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle)$, in order to prove the universal approximation property. In particular, the harmonic functions are approximated - in analogy with Equations (13) and (14) - by a series of K ridge functions

$$g\left(\sum_{k=1}^{K} w_k^{(o)} \langle \boldsymbol{w}_n, \boldsymbol{x} \rangle + b^{(o)}\right)$$

which concludes the proof of Theorem 24.

In particular, Mallat [Cam19] shows that the bounded compact support $\mathcal{X}_\square$ can be embedded in a hyperrectangle $[-T, T]^{d_X}$, on which a hyperrectangular regular grid of points is defined. This grids serves as support for the harmonic orthonormal basis $\mathcal{B}_{d_X} := \bigotimes_{m=1}^{d_X} \mathcal{B}_m$ of $L^2\left([-T, T]^{d_X}\right)$, obtained by tensorization of the basis $\mathcal{B}_m := \left\{e^{\frac{i\pi n_m t}{T}}\right\}_{n_m \in \mathbb{Z}}$ of $L^2([-T, T])$[11]. Thus, f can be decomposed on its basis (see Equation (210)):

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2T)^{d_X}} \sum_{(n_1,\dots,n_{d_X}) \in \mathbb{Z}^{d_X}} \hat{f}(\boldsymbol{w}_{(n_1,\dots,n_{d_X})}) \prod_{m=1}^{d_X} e^{\frac{i\pi n_m x_m}{T}} = \\ &= \frac{1}{(2T)^{d_X}} \sum_{n \in \mathbb{Z}} \hat{f}(\boldsymbol{w}_n) \cdot e^{i\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle} \end{aligned} \tag{90}$$

For practical purposes, a truncation of the infinite series in Equation (90) is required. This corresponds to filter the high frequencies by forcing $\{\boldsymbol{w}_n\}_{n > N_0} = \boldsymbol{0}$. The truncation error e, in $L^2\left([-T, T]^{d_X}\right)$ norm is defined as[12]:

[11] The standard scalar product in $L^2([-T, T])$ reads

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{L^2([-T,T])} = \int_{-T}^{T} \langle \boldsymbol{f}^*(t), \boldsymbol{g}(t) \rangle \, dt$$

For any two values of the basis $\mathcal{B}_m$, it holds that:

$$\left\langle e^{\frac{i\pi n_j t}{T}}, e^{\frac{i\pi n_k t}{T}} \right\rangle_{L^2([-T,T])} = \int_{-T}^{T} e^{\frac{-i\pi n_j t}{T}} \cdot e^{\frac{i\pi n_k t}{T}} \, dt = \delta_{jk}$$

Which proves their orthonormality.

[12] According to [Cam18], the proof of the Lemma is based on a uniform convergence norm, but $\|\cdot\|_{L^2}$ and $|\cdot|$ are not equivalent in infinite dimension. However, if $f \in C^\infty$, the uniform convergence is proved, by Theorem 40. Unfortunately, this is not true if $f \in C^0$, but the Stone-Weierstrass theorem allows to approximate (via uniform convergence) the function with a series of polynomial functions $p(x)$ that can be differentiated infinite times, whose Fourier coefficients $\hat{p}(k)$ decay fast and therefore the approximation is still valid.

$$
\begin{aligned}
e^2 =& \frac{1}{(2T)^{d_X}} \int_{[-T,T]^{d_X}} \left| f(\boldsymbol{x}) - \frac{1}{(2T)^{d_X}} \sum_{n=1}^{N_0} \hat{f}(\boldsymbol{w}_n) \cdot e^{i\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle} \right|^2 dt \leq \\
\leq& \frac{1}{(2T)^{d_X}} \sum_{\|\boldsymbol{w}_n\|_2 \geq C(\varepsilon)} |\hat{f}(\boldsymbol{w}_n)|^2
\end{aligned}
\tag{91}
$$

The truncation threshold $C(\varepsilon)$ depends on the accuracy of the approximation expressed by $e \leq \varepsilon$ and it implies to limit the norm of the weights:

$$
\|\boldsymbol{w}_n\|_2 \leq C(\varepsilon) \qquad \forall n > N_0
$$

In practice, this truncation is imposed by adding a penalty to the empirical loss in Equation (92), corresponding to the L^p-norm of the weights (often L^2 or L^1):

$$
L_{\mathcal{D}_{XY}}(\boldsymbol{h}_\theta) = \frac{1}{N} \sum_{(\boldsymbol{x}_k, \boldsymbol{y}_k) \in \mathcal{D}_{XY}} \ell(\boldsymbol{h}_\theta(\boldsymbol{x}_k), \boldsymbol{y}_k) + \lambda \cdot \|\boldsymbol{w}\|_p \tag{92}
$$

Another interesting aspect of Theorem 24 and inherently of Lemma 26 is related to the behaviour of the Fourier coefficients $\hat{f}(\boldsymbol{w}_n)$. The latter are the interpolation weights for reconstructing the signal with the Inverse Fourier Transform (see Equation (210). Therefore, their values are intimately related to the function regularity. Theorem 40 implies that, provided an arbitrary precision ε, the higher the regularity of f, the faster the decay of its Fourier coefficients with $\|\boldsymbol{w}\|_2$. This means that a lower number of wavelengths is demanded to approximate f. This result allows to estimate the number of weights needed by a 1-hidden layer $\mathcal{MLP}$ to approximate any regular function $f \in C^p$, based on the Theorem 24. In particular, the number of weights $N \propto M^{d_X}$ increases with the dimension of the input data d_X. Another reason why the number of weights in the hidden layer can easily become very very large is the irregularity of the function f to approximate: which implies to employ a larger number of high-frequency Fourier coefficient to reach the desired precision $e \leq \varepsilon$ [Cam19]. As a matter of fact, irregular functions have a broader-band Fourier spectrum (see for instance the Dirac delta $\delta(x - x_0)$). In this case, M can become very large because of the *Heisenberg uncertainty principle* defined by the Theorem 43 that explicitly states the underlying trade-off between time and frequency localization [Mal09].

Controlling the error via a L^2 norm as done in Equation (91) does not necessarily imply the uniform convergence stated in Theorem 25, unless $f \in C^\infty(\mathbb{R}^{d_X})$. However, even for $f \in C^0(\mathbb{R}^{d_X})$, the Stone-Weierstrass theorem that states the possibility of approximating f with a series of functions $p_i(\boldsymbol{x}) = \int_{\mathcal{X}_\square} N_i(\boldsymbol{x} - \boldsymbol{y}) \cdot f(\boldsymbol{y}) d\boldsymbol{y}$, with N_i polynomials of order q, provided that $\int_{\mathcal{X}_\square} N_i(\boldsymbol{x}) d\boldsymbol{x} = 1$.

In this case,

$$\forall \varepsilon > 0, \forall \boldsymbol{x} \in \mathcal{X}_{\square}, \exists R(\varepsilon) \text{ such that } , \forall \boldsymbol{x} \in [-R, R]^{d_X} \quad |f(\boldsymbol{x}) - \sum_{i=1}^{N} \alpha_i \cdot p_i(\boldsymbol{x})| \leq \varepsilon$$

One can easily recognize here the basics of the Finite Element Method. Moreover, the Fourier coefficients $\hat{N}_i(k)$ decay as fast as $\|\boldsymbol{w}\|^{-q}$, but, compared with Equation (89), the series of polynomes N_i does not represent the Fourier transform of f [Cam19]. Plus, Theorem 25 demands non-polynomial functions in order to assure the uniform convergence.

Finally, the Theorem 24 adopts ridge functions such as those presented in Section 2.2 to assure the universal approximation power of a 1-hidden-layer $\mathcal{MLP}$. First, based on the Hölder's inequality, each argument in the approximating the Fourier series can be bounded as follows on $\mathcal{X}_{\square}$:

$$\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle = \sum_{i=1}^{d_X} w_{n,i} x_i \leq \left(\sum_{i=1}^{d_X} w_{n,i}^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{i=1}^{d_X} x_i^2 \right)^{\frac{1}{2}} = \|\boldsymbol{w}_n\|_2 \cdot \|\boldsymbol{x}\|_2 \leq C(\varepsilon) \cdot T \cdot d_X \tag{93}$$

Recalling the expression of the truncation error $e \leq \varepsilon$ in Equation (91), the approximation based on ridge functions employed to approximate the harmonic functions can be expressed as[13]:

$$\begin{aligned} |\hat{f}(\boldsymbol{w}_n)| \cdot |\cos(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle) - \sum_{k=1}^{K} w_k^{(o)} \cdot g(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle + b_n)| \leq \\ \leq \frac{M \cdot |\cos(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle + b_n) - \sum_{k=1}^{K} w_k^{(o)} \cdot g(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle + b_n)|}{1 + \|\boldsymbol{w}_n\|^{p+1+\varepsilon}} \\ \text{with } \langle \boldsymbol{w}_n, \boldsymbol{x} \rangle \leq C(\varepsilon) \cdot T \cdot d_X \end{aligned} \tag{94}$$

The approximation in Equation (94) represents the outcome of a 1-hidden-layer $\mathcal{MLP}$ with linear output activation function $g^{(o)}(a) = a$ and $b^{(o)} = 0$. For instance, if one selects a combinations of translated (via the biases b_n) *ReLU* functions for the hidden layer activation function g, the approximation in Equation (94) represents a piece-wise linear interpolation of f (see Remark 39 [Cam19]). In order to keep the error

$$\begin{aligned} e_n^{'} = |\cos(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle) - \sum_{k=1}^{K} w_k^{(o)} \cdot g(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle + b_n)| + \\ + |\sin(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle) - \sum_{k=1}^{K} w_k^{(o)} \cdot g(\langle \boldsymbol{w}_n, \boldsymbol{x} \rangle + b_n)| \end{aligned}$$

[13] For the sinus function, the same considerations apply.

small enough, i.e., $e_n^{'} \leq \frac{\varepsilon}{N}$, with N being the number of weights for which $\|\boldsymbol{w}_n\|_2$, one needs at least $K = \frac{2T \cdot N \cdot C(\varepsilon) \cdot d_X}{\varepsilon}$ per each Fourier coefficient. The total amount of ridge functions g to approximate $f(\boldsymbol{x})$ is then $K \cdot N = \frac{2T \cdot N^2 \cdot C(\varepsilon) \cdot d_X}{\varepsilon}$. In particular, if we fix $C(\varepsilon)$, $N \approx C(\varepsilon)^{d_X}$ in order to keep $e = \mathcal{O}(\frac{C(\varepsilon)}{2T})$. Therefore, the total number of ridge functions required by the $\mathcal{MLP}$ is

$$N_K = K \cdot N \approx \frac{2T \cdot C(\varepsilon)^{2d_X+1} \cdot d_X}{\varepsilon}$$

(N sine et cosine functions approximated by K ridge functions each). Whenever the Fourier coefficients decrease slowly with the frequency, because of the irregularity of the function, $C(\varepsilon)$ increases accordingly and therefore one requires a larger number of hidden neurons N to keep the a lower approximation error. The increasing dimension d_X of the dataset imply an even larger number of parameters.

Considering that $\|f\|_{L^2(\mathbb{R})} < \|f\|_{W^{k,2}(\mathbb{R})} \leq \|f\|_{L^2(\mathbb{R})} + \|f^{(k)}\|_{L^2(\mathbb{R})}$, and according to Equation (91), the L^2 approximation error is bounded as follows [Cam19]:

$$e^2 \leq \frac{1}{(2T)^{d_X}} \sum_{\|\boldsymbol{w}_n\|_2 \geq C(\varepsilon)} |\hat{f}(\boldsymbol{w}_n)|^2 < \frac{1}{(2T)^{d_X}} \sum_{\|\boldsymbol{w}_n\|_2 \geq C(\varepsilon)} \frac{1 + \|\boldsymbol{w}_n\|_2^{2k}}{C(\varepsilon)^{2k}} |\hat{f}(\boldsymbol{w}_n)|^2 \tag{95}$$

If $f \in W^{k,2}$, $\exists 0 < B < +\infty$ such that $\|f\|_{W^{k,2}(\mathbb{R})} \leq \|f\|_{L^2(\mathbb{R})} + \|f^{(k)}\|_{L^2(\mathbb{R})} < B$, which implies that the approximation error is bounded, i.e. $e < \frac{B}{C(\varepsilon)^{2k}} = \varepsilon$, which implies that [Cam19]:

$$C(\varepsilon) = \left(\frac{B}{\varepsilon}\right)^{\frac{1}{2k}} \qquad N_K \approx \frac{2T \cdot B^{\frac{2d_X+1}{2k}} \cdot d_X}{\varepsilon^{\frac{2d_X+1}{2k}}} \tag{96}$$

Equation (96) implies that, for a fixed approximation accuracy ε, the number of hidden neurons in 1-hidden-layer $\mathcal{MLP}$ increases as $\varepsilon^{\frac{2d_X+1}{2k}}$, i.e., the number of neuron remains constant whether the regularity of the function k increases with the dimension of the data space d_X. Therefore, the 1-hidden-layer $\mathcal{MLP}$ can theoretically approximate poorly regular functions in large dimension, provided that the number of hidden neurons is large enough (very large) [Cam19]. The number of hidden neurons estimated by Mallat in [Cam19] can be refined though. In particular, according to [Mai99], the following theorem holds:

Theorem 27. *Lower bounds for approximation by $\mathcal{MLP}$ neural networks [Mai99] Given a function $f : [-T,T]^{d_X} \to \mathbb{R}$, $T > 0$, with $f \in W^{k,2}$, then it exists a 1-hidden-layer $\mathcal{MLP}$ h_θ, with sigmoid activation function and N_K hidden neurons, such that the L^2-error $e = \|f - h_\theta\|_{L^2([-T,T]^{d_X})}$ is lower than a tolerance ε if:*

$$N_K \approx \varepsilon^{\frac{1-d_X}{k}} \tag{97}$$

This theorem implies that to reduce the error by an order or magnitude, the number of neurons must be multiplied by $10^{\frac{d_X-1}{k}}$. For instance, an $W^{2,2}$ function defined over $\mathbb{R}^3$ demands to multiply by a factor 10 to decrease by 10 the approximation error.

2.2 On the sparsity of the $\mathcal{MLP}$

To go further, [Bar93] proposed a strategy to weaken the assumptions on the function regularity, yet breaking the curse of dimensionality. In particular, [Bar93] proposes to focus on functions with L^1 first order derivative. As a matter of fact, if $f \in W^{1,1}\left(\mathbb{R}^{d_X}\right)$, then:

$$f \in W^{1,1}\left(\mathbb{R}^{d_X}\right) \iff \|f(\boldsymbol{x})\|_{L^1\left(\mathbb{R}^{d_X}\right)} < +\infty, \quad \|\boldsymbol{\nabla}_x f(\boldsymbol{x})\|_{L^1\left(\mathbb{R}^{d_X}\right)} < +\infty \tag{98}$$

According to Remark 41 and following the strategy based on the Fourier analysis, proposed by S.Mallat [Cam19], Equation (98) implies the norm $\|f(\boldsymbol{x})\|_{W^{1,1}\left(\mathbb{R}^{d_X}\right)} =< +\infty$ is equivalent to $\|f(\boldsymbol{x})\|_{L^1\left(\mathbb{R}^{d_X}\right)} + \|\boldsymbol{\nabla}_x f(\boldsymbol{x})\|_{L^1\left(\mathbb{R}^{d_X}\right)}$. Since $\boldsymbol{\nabla}_x f(\boldsymbol{x}) \in L^1\left(\mathbb{R}^{d_X}\right)$, its Fourier transform exists and it reads $i\boldsymbol{w}\hat{f}(\boldsymbol{w})$. Recalling the conditions of Equation (89), and bounding the compact support $\mathcal{X}_\square$ where the data live, with a regular grid discretizing the volume $[-T,T]^{d_X} \supseteq \mathcal{X}_\square$, we can locally approximate $\boldsymbol{\nabla}_x f(\boldsymbol{x})$ with its Fourier series on $[-T,T]^{d_X}$:

$$\boldsymbol{\nabla}_x f(\boldsymbol{x}) = \sum_{\boldsymbol{w}_n \in \mathbb{Z}^{d_X}} i\boldsymbol{w}_n \hat{f}(\boldsymbol{w}_n)\, e^{-i\langle \boldsymbol{w}_n, \boldsymbol{x}\rangle}, \quad \forall \boldsymbol{x} \in [-T,T]^{d_X} \tag{99}$$

On the compact support $[-T,T]^{d_X}$, the L^1-norm of $\boldsymbol{\nabla}_x f(\boldsymbol{x})$ is bounded as follows:

$$\begin{aligned} \|\boldsymbol{\nabla}_x f(\boldsymbol{x})\|_{L^1\left(\mathbb{R}^{d_X}\right)} &= \int_{[-T,T]^{d_X}} \|\boldsymbol{\nabla}_x f(\boldsymbol{x})\|_1 d\boldsymbol{x} = \\ &\int_{[-T,T]^{d_X}} \| \sum_{\boldsymbol{w}_n \in \mathbb{Z}^{d_X}} i\boldsymbol{w}_n \hat{f}(\boldsymbol{w}_n)\, e^{-i\langle \boldsymbol{w}_n, \boldsymbol{x}\rangle} \|_1 d\boldsymbol{x} \leq \\ &\leq \int_{[-T,T]^{d_X}} \sum_{\boldsymbol{w}_n \in \mathbb{Z}^{d_X}} \|\boldsymbol{w}_n\|_1 \cdot |\hat{f}(\boldsymbol{w}_n)| d\boldsymbol{x} < +\infty \end{aligned} \tag{100}$$

In the revisited version of the work of Barron [Bar93], by S. Mallat [Cam19], the regularity condition in Equation (100) is adopted show the accuracy of approximating f with the function in Equation (18).

Theorem 28. *Approximation bounds of a neural network [Bar93; Cam19] Given a function $f : \mathbb{R}^n \to \mathbb{R}$, with $f \in W^{1,1}\left(\mathbb{R}^{d_X}\right)$ and provided a constant $C_f = \sum_{\boldsymbol{w}_n \in \mathbb{Z}^{d_X}} |\hat{f}(\boldsymbol{w}_n)| < +\infty$, then $\exists \{\boldsymbol{w}_n\}_{n \leq N_K}$, with $K = \frac{C_f^2}{\varepsilon^2}$ such that*

$$\|f(\boldsymbol{x}) - \sum_{n=1}^{N_K} \hat{f}(\boldsymbol{w}_n)\, e^{-i\langle \boldsymbol{w}_n, \boldsymbol{x}\rangle}\|_{L^2\left(\mathbb{R}^{d_X}\right)} \leq \frac{C_f^2}{N_K}$$

Proof. In order to prove the revisited version of [Bar93] work, S.Mallat in [Cam19] proposes to reindexing the Fourier coefficients $\hat{f}(\boldsymbol{w}_n)$ in descending order, i.e., according to an indexing $\mathcal{I}: \mathbb{Z} \to \mathbb{N}$ such that that

$$|\hat{f}(\boldsymbol{w}_n)| > |\hat{f}(\boldsymbol{w}_m)| \implies \quad \mathcal{I}(n) < \mathcal{I}(m), \forall m, n \in \mathbb{Z}$$

C_f can therefore be rewritten as

$$C_f = \sum_{n \in \mathbb{N}} |\hat{f}(\boldsymbol{w}_n)| = \sum_{n=1}^{N_K} |\hat{f}(\boldsymbol{w}_n)| + \sum_{n > N_K} |\hat{f}(\boldsymbol{w}_n)|$$

Since the Fourier coefficients have been reordered in descending order:

$$C_f \geq \sum_{n=1}^{N_K} |\hat{f}(\boldsymbol{w}_n)| \geq N_K |\hat{f}(\boldsymbol{w}_{N_K})|$$

which implies that the N_K^{th} smaller Fourier coefficient is lower than $\frac{C_f}{N_K}$. To conclude the proof, the approximation error can be rewritten as:

$$\|f(\boldsymbol{x}) - \sum_{n=1}^{N_K} \hat{f}(\boldsymbol{w}_n)\, e^{-i\langle \boldsymbol{w}_n, \boldsymbol{x}\rangle}\|^2_{L^2(\mathbb{R}^{d_X})} = \|\sum_{n > N_K} \hat{f}(\boldsymbol{w}_n)\, e^{-i\langle \boldsymbol{w}_n, \boldsymbol{x}\rangle}\|^2_{L^2(\mathbb{R}^{d_X})} \leq$$

$$\leq \sum_{n > N_K} \frac{C_f^2}{k^2} \leq C_f^2 \int_{N_K}^{+\infty} \frac{1}{x^2} dx = \frac{C_f^2}{N_K}$$

□

Remark 29. The original version of Theorem 28 by [Bar93] proves that considering $W^{1,1}(\mathbb{R}^{d_X})$ functions, the L^2-error of approximation is bounded by above by the constant $\frac{\int_{\mathbb{R}^{d_X}} \|\boldsymbol{w}\|_1 \cdot |\hat{f}(\boldsymbol{w})| d\boldsymbol{w}}{\sqrt{N_K}}$.

The importance of Theorem 28 resides in the fact that, penalizing the learning algorithm with a ℓ^1 norm on the weights $\boldsymbol{\theta}$ allows to promote sparsity and achieve parsimony. In other words, the 1-hidden-layer $\mathcal{MLP}$ will learn how to approximate the labelling function with the least amount of hidden features (the famous above mentioned dictionary) and with an approximation error that is not dependent on the dimension of the input space (or alternatively, for a fixed accuracy, the design of the $\mathcal{NN}$ will require an amount of hidden neurons that is independent of the dimension of the data space.

3 Advanced topics in $\mathcal{CNN}$

In the following, a collection of labeled images $\mathcal{D}_{XY} = \{\mathbb{X}_k, y_k\}_{k=1}^N$ of $W \times H$ pixels each is considered. Any image $\mathbb{X}_k$ of the database is represented by a set

of pixels, each one associated to a color expressed by a 3-dimensional vector on the Red Green Blue (RGB) scale $\boldsymbol{\psi} : \mathbb{R}^3 \to \mathbb{R}^3$ (see Section 2.5). Therefore, the image $\mathbb{X}_k \in \mathbb{R}^{W \times H \times 3}$ can be alternatively indicated by the color vector $\boldsymbol{\psi}$ as:

$$\begin{aligned} \mathbb{X}\left(x_1, x_2, :\right) &= \boldsymbol{\psi}\left(x_1, x_2, 0\right) = \\ &= \psi_R\left(x_1, x_2, 0, :\right) \boldsymbol{e}_1 + \psi_G\left(x_1, x_2, 0, :\right) \boldsymbol{e}_2 + \psi_B\left(x_1, x_2, 0, :\right) \boldsymbol{e}_3, \\ &\left(x_1, x_2\right) \in [0,1]^2 \end{aligned} \tag{101}$$

For the sake of simplicity, the image is spanned over a $[0,1]^2$ domain, by a regular grid of $W \times H$ pixels with a 2D index $\boldsymbol{x} = x_1 \boldsymbol{e}_1 + x_2 \boldsymbol{e}_2, (x_1, x_2) \in [0,1]^2$. In the following, in order to simplify the mathematical formulation, it is assumed that the image $\boldsymbol{\psi} : \mathbb{R}^3 \to \mathbb{R}^3$ with $x_3 = 0$ for all 2D image and $\boldsymbol{\psi} : (x_1, x_2, 0) \mapsto \boldsymbol{\psi}\left(x_1, x_2, 0\right)$. The 0 is omitted to avoid cumbersome notation. The label y_k belongs to an alphabet A for a classification purposes, and $\boldsymbol{y}_k \in \mathbb{R}^{d_Y}$ in regression problems. The alphabet is usually coded as a *one-hot vector*, i.e. a binary vector of the size of the alphabet.

3.1 A group-theory vision of classification

If it exists a labeling function $f : \mathbb{X} \mapsto y$, we can study its regularity and invariance with respect to any transformation applied to the image, that preserves all the labeling function level sets

$$L_c(f) = \left[\mathbb{X} \in \mathcal{X} | \quad f\left(\mathbb{X}\right) = c\right], \quad \forall c \in \mathrm{Im}\left(f\right)$$

. This invariance is expressed by a symmetry group [Mal16; Cam20][14]:

$$G := \left[g | \forall \mathbb{X} \in L\left(f\right), \quad f\left(g.\mathbb{X}\right) = f\left(\mathbb{X}\right)\right] \tag{102}$$

Some standard transformation:

- Translation:

$$\begin{aligned} g\mathbb{X} = g\boldsymbol{\psi}\left(\boldsymbol{x}\right) = \boldsymbol{\psi}\left(\boldsymbol{x} - \boldsymbol{c}_g\right) \\ \text{with } \boldsymbol{c}_g \in \mathbb{R}^2 \end{aligned} \tag{103}$$

[14] One can prove that G is a groups since, $\forall g_1, g_2, g_3 \in G$ and $\forall \mathbb{X} \in \mathcal{X}$:

- $f\left(\left(g_1 g_2\right) g_3 \mathbb{X}\right) = f\left(g_3 \mathbb{X}\right) = f\left(\mathbb{X}\right) = f\left(g_3 \mathbb{X}\right) = f\left(g_2\left(g_3 \mathbb{X}\right)\right) = f\left(g_1\left(g_2 g_3\right) \mathbb{X}\right)$ (associative)
- $\exists Id$ such that $f\left(Id.\mathbb{X}\right) = f\left(\mathbb{X}\right)$ (neutral element)
- $\exists g^{-1}$ such that $f\left(g^{-1} g.\mathbb{X}\right) = f\left(g^{-1}\mathbb{X}\right) = f\left(\mathbb{X}\right)$ (inverse element)

The group si Abelian if $g_1 g_2 = g_2 g_1$ (optional).

- Rotation:

$$g\mathbb{X} = g\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{\psi}\left(\boldsymbol{R}_g\boldsymbol{x}\right) \quad \text{with } g \in [0, 2\pi] \text{ and } \boldsymbol{R}_g \in SO(2) \tag{104}$$

- Roto-translation:

$$g\mathbb{X} = g\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{\psi}\left(\boldsymbol{R}_g\left(\boldsymbol{x} - \boldsymbol{c}_g\right)\right) \quad \text{with } g \in [0, 2\pi], \boldsymbol{c}_g \in \mathbb{R}^2 \text{ and } \boldsymbol{R}_g \in SO(2) \tag{105}$$

- Normalization:

$$g\mathbb{X} = g\boldsymbol{\psi}(\boldsymbol{x}) = s_g\boldsymbol{\psi}(\boldsymbol{x}) \quad \text{with } s_g \in \mathbb{R} \tag{106}$$

- More general, the group of diffeomorphisms acts according to the following law:

$$g\left(\mathbb{X}(\boldsymbol{x})\right) = \boldsymbol{\psi}\left(g(\boldsymbol{x})\right) \tag{107}$$

If we knew all the elements of G, we would know f entirely, since we could apply several deformation to the image preserving all the level sets $f(\boldsymbol{x}) = c, \forall c$. S. Mallat, who contributed to develop this original theory, asserts that it is hard to know before hand all the possible strain fields which the labelling function is invariant to. We can guess some (the classification is insensitive to roto-translations, for instance) but definitely not all of them [Mal16; Cam20]. In other words, the dimension of G is too large. But we can infer it from a subgroup of plausible strain fields H. As a matter of fact, $\forall g \in H$, $\mathbb{X}$ and $g\mathbb{X}$ belong to the same equivalence class. For instance, the image $\mathbb{X}$ engenders the following equivalence class:

$$[\mathbb{X}] := \left\{\mathbb{X}^{'} \in L(f) \,|\quad \mathbb{X}^{'} = h\mathbb{X}, h \in H\right\}$$

Note that, since $f\left(\mathbb{X}^{'}\right) = f(h\mathbb{X}) = f(\mathbb{X})$. The ensemble quotient associated to $[\mathbb{X}]$ reads:

$$L(f)\,|H := \left([\mathbb{X}] \in \mathcal{P}(\mathcal{X}), \mathbb{X} \in \mathcal{X}\right) \tag{108}$$

with $\mathcal{P}(\mathcal{X})$ being the power set of $\mathcal{X}$. For all $\mathbb{X}_0 \in L(f)\,|H$, $f(g\mathbb{X}_0) \in [\mathbb{X}_0]$: in other words, for each image $\mathbb{X}_0$, classified, for instance, as the image of "cat", despite any roto-translation or other known transformations of the group symmetries $g \in H$, the class of the transformed image $g\mathbb{X}_0$ remains classified as "cat". The advantage of this strategy, based on the group theory, formulated by S. Mallat is that it reduces the dimensionality: as a matter of fact, we can find the subgroup of transformations H, with $\dim(L(f)|H) = \dim(L(f)) - \dim(H)$, that are common to all the images belonging to the the equivalence class, i.e. to all the images classified by f in the same way.

3.1.1 Continuum Mechanics applied to image classification?

Several invertible transformation $g_1, g_2, \ldots, g_n$ can be applied to $\mathbb{X}$, but the labelling function should not be affected. All groups of G can potentially generate subgroups $\{g^k, k \in \mathbb{Z}\}$. Any set of groups $A = \{g_1, \ldots g_n\}$ generates a subgroup $\{A\}$, whose elements are defined by permutations of the product $g_1 g_2 \ldots g_n$. In particular, it is interesting to study the group orbit $O_{\mathbb{X}} = \{g\mathbb{X}\}_{g \in G}$ for which $f\left(O_{\mathbb{X}}\right) = f\left(\mathbb{X}\right)$. We borrow from continuum mechanics (see [Hil79; TN04; KD21] the theoretical framework necessary to describe continuous transformations that can be applied to the image, provided that each infinitesimal transformation preserves the label $f\left(\mathbb{X}\right)$. This approach provides a more flexible description of the ensemble of possible group actions $g \in H$ on the image, that can be *a priori* very very large and not limited to finite roto-translations. In particular, we can think of the image pixels as a set of points $\{\boldsymbol{X}_n\}_{n=1}^{W \times H} \in \mathrm{B}$[15], disposed on a regular grid on the $\mathbb{R}^2$ space, via the *placement* function $p : \mathrm{B} \to (\mathcal{E}, \mathcal{R})$ (the 2D euclidean space, with reference system $\mathcal{R}$ defined by $(O, \boldsymbol{e}_1, \boldsymbol{e}_2)$), which is smooth and it preserves the orientation (positive Jacobian, i.e., two points cannot occupy the same position). B is defined as a "body" [TN04], an abstract manifold of dimension 2, endowed with a Fréchet derivative [KD21]. The image can be placed in $(\mathcal{E}, \mathcal{R})$ in different configurations t, occupying a volume $\mathbb{R}^3 \supset \Omega_t = \mathrm{Im}(p_t) = p_t\left(\mathrm{B}\right)$. The space of configurations is defined as the ensemble of placement functions $p_t \in C^\infty\left(\mathrm{B}, \mathcal{E}\right)$ and it called *embedding* $\mathrm{Emb}^\infty\left(\mathrm{B}, \mathcal{E}\right) \subset C^\infty\left(\mathrm{B}, \mathcal{E}\right)$, which is an infinite-dimensional differentiable manifold. In particular, the placement $p_0\left(\boldsymbol{p}_n\right)$ places any pixel of any image in the reference configuration p_0, in a coordinate of the $\mathbb{R}^3$ space referred as to

$$\boldsymbol{x}_0 = p_0\left(\boldsymbol{p}\right), \qquad \forall \boldsymbol{p} \in \mathrm{B}$$

and the same pixel in a configuration p_t,

$$\boldsymbol{x} = p_t\left(\boldsymbol{p}\right), \qquad \forall \boldsymbol{p} \in \mathrm{B}$$

More generically, in analogy with mechanics, the color can be seen as a spatial pixel-wise vector field $\boldsymbol{\psi}\left(\boldsymbol{x}; t\right)$, where the "time" t stands for the non-dimensional continuum index defining successive transformations g_1, g_2, …, g_t, …. As a matter of fact, the image can be distorted from an arbitrary reference configuration p_0 and classified in the same way all along the orbit $O_{\mathbb{X}} = \{g_t\mathbb{X}\}_{g_t \in G, t \in \mathbb{R}}$. Moreover, data augmentation techniques are directly based upon enlarging the dataset at stake by applying arbitrary roto-translations or other kinds of diffeomorphims to the original image dataset (see Figure 5). Now, we should introduce a transformation from two different configurations, in order to deform or roto-translate the image, i.e., a diffeomorphisms between two configurations, called *deformation* $\boldsymbol{\varphi} : \Omega_0 \to \Omega_t$ and defined as:

$$\boldsymbol{x} = \boldsymbol{\varphi}\left(\boldsymbol{x}_0\right) \qquad \boldsymbol{\varphi} = p_t \circ p_0^{-1}$$

[15] $\boldsymbol{X}$ are denoted as *material coordinates*

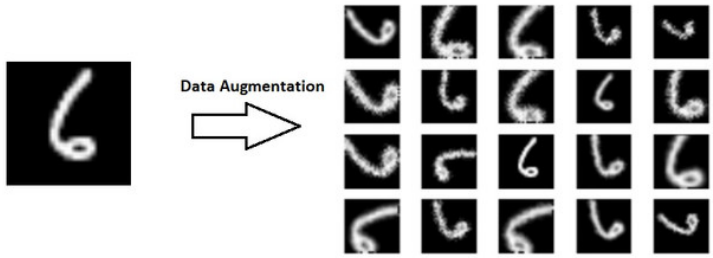

Figure 5: Data augmentation on Modified National Institute of Standards and Technology database (MNIST) dataset [LeC98]

The image in the arbitrary reference configurations occupies a volume Ω_0 and, when distorted by a diffeomorphisms, it occupies a volume Ω_t, so that $\boldsymbol{\varphi} : \Omega_0 \rightarrow \Omega_t$. One can notice that the phase flow associated to $\boldsymbol{\varphi}(\boldsymbol{x})$ and denoted as $\boldsymbol{\varphi}(\boldsymbol{x};t)$ has the following properties:

$$\boldsymbol{\varphi}(\boldsymbol{x}_0,0) = Id(\boldsymbol{x}_0) = \boldsymbol{x}_0, \qquad \boldsymbol{\varphi}(\boldsymbol{\varphi}(\boldsymbol{x}_0,s),t) = \boldsymbol{\varphi}(\boldsymbol{x}_0,s+t) \tag{109}$$

If $\boldsymbol{\varphi} \in C^1(\Omega_0)$, its gradient exists and it defined as:

$$\boldsymbol{dx}(t) = \mathbb{F}(\boldsymbol{x}_0,t)\,\boldsymbol{dx}_0 + o(\|\boldsymbol{dx}_0\|) \qquad \mathbb{F}(\boldsymbol{x}_0,t) = \sum_{n=1}^{2} \frac{\partial \boldsymbol{\varphi}(\boldsymbol{x}_0,t)}{\partial x_n} \otimes \boldsymbol{e}_n \tag{110}$$

The linearization or variation of a spatial tensor field $\boldsymbol{\psi}(\boldsymbol{x})$ is formally equivalent, in mechanics, to the Lie derivative of the spatial field itself. The Lie algebra is based on the well know *pull-back* operation by any $p \in \mathrm{Emb}^{\infty}(\mathrm{B},\mathcal{E})$ [KD21]:

$$\boldsymbol{\Psi}(\boldsymbol{X}) = p^*\boldsymbol{\psi} = J\mathbb{F}^{-1}(\boldsymbol{\psi} \circ p) \tag{111}$$

with $J = \det(\mathbb{F})$ and $\mathbb{F}^{-1}$ the inverse gradient. The inverse of the pull-back is the push-forward:

$$\boldsymbol{\psi}(\boldsymbol{x}) = p_*\boldsymbol{\Psi} = \frac{1}{J}\mathbb{F}\boldsymbol{\Psi} \circ p^{-1} \tag{112}$$

The pull-back operation allows to refer to the body B itself, where each pixel is identified in an abstract sense. The Lie algebra defines the fixed tangent planes of the differentiable geodetic $O_{\mathbb{X}}$, generated by the gradient $\mathbb{F}(\boldsymbol{X};t)$, that allow to pass from a transformation at time t to another one infinitesimally close at time $t+dt$. The *Lie derivative* is an infinitesimal version of the pull-back and it reads, for a contravariant vector $\boldsymbol{\psi}$, along the vector $\boldsymbol{X}$ of material coordinates [KD21]:

$$L_{\boldsymbol{u}}\boldsymbol{\psi} = \frac{\partial}{\partial t}\Big|_{t=0} \boldsymbol{\varphi}(\boldsymbol{x};t)^*\,\boldsymbol{\psi} \tag{113}$$

In other words, the Lie derivative consists into compute the pull-back on $\boldsymbol{\psi}$, compute its Gateaux derivative $\frac{d}{dt}\Big|_{t=0} \boldsymbol{\Psi}\left(\boldsymbol{X}+t\boldsymbol{c}\right)$ and finally push it forward to the spatial configuration. For the sake of clarity, let us define the Lie derivative for a vector field $\boldsymbol{\psi}$, by firstly defining the following phase flow in $\mathbb{R}^3$ (Landau notation):

$$\boldsymbol{\varphi}\left(\boldsymbol{x};-t\right)=\boldsymbol{x}-t\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t) \tag{114}$$

and the associated Gateaux derivative along $\boldsymbol{y}\in\mathbb{R}^3$:

$$D_{\boldsymbol{y}}\boldsymbol{\varphi}\left(\boldsymbol{x};-t\right)=\boldsymbol{y}-tD_{\boldsymbol{y}}\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t) \tag{115}$$

From now on, let us assume that $\boldsymbol{u}\left(\boldsymbol{x}\right)$ is Fréchet differentiable, with the bounded linear operator

$$\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}=\sum_{n=1}^{3}\frac{\partial\boldsymbol{u}\left(\boldsymbol{x}\right)}{\partial x_n}\otimes\boldsymbol{e}_n$$

that represents the gradient of the vector $\boldsymbol{u}$ and for which

$$D_{\boldsymbol{y}}\boldsymbol{u}\left(\boldsymbol{x}\right)=\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}\left(\boldsymbol{x}\right).\boldsymbol{y}$$

In continuum mechanics, this implies that the displacement field does include discontinuity or fractures. Plugging $\boldsymbol{y}=\boldsymbol{\varphi}\left(\boldsymbol{x};t\right)$ in Equation (115), the latter reads:

$$\begin{aligned}D_{\boldsymbol{\varphi}(\boldsymbol{x};t)}\boldsymbol{\varphi}\left(\boldsymbol{x};-t\right)&=\left(\mathbf{I}-t\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t)\right)\left(\boldsymbol{x}+t\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t)\right)=\\&=\boldsymbol{\varphi}\left(\boldsymbol{x};t\right)-tD_{\boldsymbol{x}}\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t)\end{aligned} \tag{116}$$

Moreover, the Taylor expansion of the continuous map $\boldsymbol{\psi}$ reads:

$$\boldsymbol{\psi}\left(\boldsymbol{\varphi}\left(\boldsymbol{x};t\right)\right)=\boldsymbol{\psi}\left(\boldsymbol{x}\right)+tD_{\boldsymbol{u}}\boldsymbol{\psi}\left(\boldsymbol{x}\right)+o(t) \tag{117}$$

The Lie derivative defined in Equation (113) can now be explicitly written as:

$$\begin{aligned}L_{\boldsymbol{u}}\boldsymbol{\psi}&=\lim_{t\to0}\frac{D_{\boldsymbol{\varphi}(\boldsymbol{x};t)}\boldsymbol{\varphi}\left(\boldsymbol{x};-t\right)\left(\boldsymbol{\psi}\left(\boldsymbol{\varphi}\left(\boldsymbol{x};t\right)\right)\right)-\boldsymbol{\psi}\left(\boldsymbol{x}\right)}{t}=\\&=\lim_{t\to0}\frac{\left(\mathbf{I}-t\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t)\right)\left(\boldsymbol{\psi}\left(\boldsymbol{x}\right)+tD_{\boldsymbol{u}}\boldsymbol{\psi}\left(\boldsymbol{x}\right)+o(t)\right)-\boldsymbol{\psi}\left(\boldsymbol{x}\right)}{t}=\\&=\lim_{t\to0}\frac{\left(\mathbf{I}-t\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}\left(\boldsymbol{x}\right)+o(t)\right)\left(\boldsymbol{\psi}\left(\boldsymbol{x}\right)+t\mathbb{D}_{\boldsymbol{x}}\boldsymbol{\psi}\left(\boldsymbol{x}\right)\boldsymbol{u}+o(t)\right)-\boldsymbol{\psi}\left(\boldsymbol{x}\right)}{t}=\\&=\mathbb{D}_{\boldsymbol{x}}\boldsymbol{\psi}\left(\boldsymbol{x}\right)\boldsymbol{u}\left(\boldsymbol{x}\right)-\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}\left(\boldsymbol{x}\right)\boldsymbol{\psi}\left(\boldsymbol{x}\right)=\\&=D_{\boldsymbol{u}}\boldsymbol{\psi}\left(\boldsymbol{x}\right)-D_{\boldsymbol{\psi}}\boldsymbol{u}\left(\boldsymbol{x}\right)=[\boldsymbol{u},\boldsymbol{\psi}]\end{aligned} \tag{118}$$

with $[\boldsymbol{u},\boldsymbol{\psi}]$ being called the *Lie brackets*. Coming back to images $\mathbb{X}$, considered as planar continuum bodies, for the sake of simplicity, the translation g in

Equation (103), corresponds to a deformation $\boldsymbol{\varphi}(\boldsymbol{x};-1)=\boldsymbol{x}-\boldsymbol{c}_g$, with $\boldsymbol{c}_g=\boldsymbol{u}(\boldsymbol{x})$, such that:

$$g\mathbb{X}=\boldsymbol{\psi}\left(\boldsymbol{x}-\boldsymbol{c}_g\right)=\boldsymbol{\psi}(\boldsymbol{x}-\boldsymbol{u}(\boldsymbol{x}))=\mathbb{X}+D_{-\boldsymbol{u}(\boldsymbol{x})}\boldsymbol{\psi}(\boldsymbol{x})+o(\|\boldsymbol{u}\|)=\boldsymbol{\psi}(\boldsymbol{\varphi}(\boldsymbol{x};1))$$

Locally, the group g of dimension 2 reflects a translation on a image in its actual configuration $\boldsymbol{x}$, with a hyper tangent plan defined by

$$\mathbb{D}_{\boldsymbol{x}}\boldsymbol{\psi}(\boldsymbol{x};t)=\sum_{n=1}^{2}\frac{\partial\boldsymbol{\psi}(\boldsymbol{x})}{\partial x_n}\otimes\boldsymbol{e}_n$$

The translation on the hyper tangent plane is provided by u_1 and u_2 that represents the local *displacement* components on the hyper-plane tangent to the group orbit and along which infinitesimally-close translations from the local diffeomorphism can be performed. The group action is resolved by the term $D_{-\boldsymbol{u}(\boldsymbol{x})}\boldsymbol{\psi}(\boldsymbol{x})$. S. Mallat in his works [Mal16; Cam20] proposes to refer to the reference configuration $\boldsymbol{x}_0$ and to decompose the displacement into two components: a global and a local one, obtained by approximating the displacement vector at the first order Taylor expansion around it:

$$\boldsymbol{u}(\boldsymbol{x})=\boldsymbol{u}(\boldsymbol{x}_0)+\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)$$

with an associated deformation (at first order in Equation (114), for $t=1$) that reads:

$$\boldsymbol{\varphi}(\boldsymbol{x};-1)=(\mathbf{I}-\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}(\boldsymbol{x}_0))(\boldsymbol{x}-\boldsymbol{x}_0)+\boldsymbol{x}_0-\boldsymbol{u}(\boldsymbol{x}_0) \tag{119}$$

The local action is represented by the local strain $(\mathbf{I}-\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}(\boldsymbol{x}_0))(\boldsymbol{x}-\boldsymbol{x}_0)$ and the global translation by $\boldsymbol{x}_0-\boldsymbol{u}(\boldsymbol{x}_0)$. According to Equation (117), the image transformation, based on this local diffeomorphism reads:

$$\boldsymbol{\psi}(\boldsymbol{\varphi}(\boldsymbol{x};-1))=\boldsymbol{\psi}(\boldsymbol{x})+\mathbb{D}_{\boldsymbol{x}}\boldsymbol{\psi}(\boldsymbol{x})((\mathbf{I}-\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u})(\boldsymbol{x}-\boldsymbol{x}_0)+\boldsymbol{x}_0-\boldsymbol{u}(\boldsymbol{x}_0)) \tag{120}$$

In analogy with group theory, the term $\mathbb{D}_{\boldsymbol{x}}\boldsymbol{\psi}((\mathbf{I}-\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u})(\boldsymbol{x}-\boldsymbol{x}_0)+\boldsymbol{x}_0-\boldsymbol{u}(\boldsymbol{x}_0))$ represent the group action on $\boldsymbol{\psi}$. In analogy with the infinitesimal strain theory in mechanics, the *small displacement* hypothesis is granted by

$$\sup_{\boldsymbol{x}\in\Omega_t}\|\boldsymbol{u}(\boldsymbol{x})\|\ll \mathrm{L}$$

and the *small strain* assumption by

$$\sup_{\boldsymbol{x}\in\Omega_t}\|\mathbb{D}_{\boldsymbol{x}}\boldsymbol{u}(\boldsymbol{x})\|\ll 1$$

with L being the characteristic size of the image (in this case, since the coordinates are normalized, $\mathrm{L}=1$).

3.1.2 Group invariance

The Lie algebra allow us to conceive a classifier h_θ insensitive to diffeomorphism (i.e., approximating the labelling function f, if it exists), by directly working on the grid of pixels. The dimension of such group of diffeomorphism is potentially high [Mal16; Cam20]. In the following, we attempt at approximating the labelling function f with a function $h_\theta : \mathcal{X} \to \mathbb{R}$ defined as:

$$h_\theta(\mathbb{X}) = \sum_{n=1}^{N} \phi_n(\mathbb{X}) w_n = \langle \boldsymbol{z}(\mathbb{X}), \boldsymbol{w} \rangle \tag{121}$$

The idea of S.Mallat is to learn f by making the group symmetries of h_θ to match the same invariants $g \in G$. In other words, it is necessary that the G is a symmetry group of $\boldsymbol{z}$, i.e. the ensemble quotient in Equation (108) [Mal16; Cam20]. The problem is that $\boldsymbol{z}$ cannot be constructed out of the symmetry groups of G, since those are potentially infinite and not known before hand. If we consider the group of diffeomorphisms, we hope we can learn the subgroup of diffeormophisms that do not affect the labelling function by adjusting the weights $\boldsymbol{w}$, since forcing $\boldsymbol{z}$ to learn the invariant would be too cumbersome and this strategy could lead to overfit the data. Moreover, S. Mallat showed a remarkable result in his works [Mal16; Cam20]: the canonical group invariant are not working for approximating the labelling function. In simple words, learning the canonical group invariants means to learn the parameters $s_g, \boldsymbol{c}_g, \boldsymbol{R}_g$ defined in Equations (103), (104) and (106) respectively. In other words, in analogy with mechanics, the idea behind this strategy is to refer all the configuration to the reference one $\boldsymbol{x}_0$, provided that this reference configuration is known before hand (a *deformable templates*, see [GM98]). For instance:

- Scale renormalization:

$$g\mathbb{X} = s_g \boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x}_0) \, \forall s_g \Rightarrow s_g = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} \| \boldsymbol{\psi}(x_{0,1j}, x_{0,2i}) \|}{\sum_{i=1}^{H} \sum_{j=1}^{W} \| \boldsymbol{\psi}(x_{1j}, x_{2i}) \|} \tag{122}$$

  If the average norm of $\boldsymbol{\psi}(\boldsymbol{x}_0)$ is 1, all possible scale factors s_g of the sort do not change $\boldsymbol{\psi}(\boldsymbol{x}) = g^{-1}\boldsymbol{\psi}(\boldsymbol{x}_0) = \left(\sum_{i=1}^{H} \sum_{j=1}^{W} \| \boldsymbol{\psi}(x_{1j}, x_{2i}) \| \right) \boldsymbol{\psi}(\boldsymbol{x}_0)$.

- Recentering:

$$\begin{aligned} g\mathbb{X} =& \boldsymbol{\psi}\left(\boldsymbol{x} - \boldsymbol{c}_g\right) = \boldsymbol{\psi}(\boldsymbol{x}_0) \\ \boldsymbol{c}_g =& \frac{\sum_{j=1}^{W} x_{1j} \| \boldsymbol{\psi}(x_{1j}, x_{2i}) \| \boldsymbol{e}_1 + \sum_{i=1}^{H} x_{2i} \| \boldsymbol{\psi}(x_{1j}, x_{2i}) \| \boldsymbol{e}_2}{\sum_{i=1}^{H} \sum_{j=1}^{W} \| \boldsymbol{\psi}(x_{1j}, x_{2i}) \|} \\ \Rightarrow & \boldsymbol{c}_g = \boldsymbol{x} - \boldsymbol{x}_0 \end{aligned} \tag{123}$$

  with $\boldsymbol{\psi}(\boldsymbol{x}) = g^{-1}\boldsymbol{\psi}(\boldsymbol{x}_0) = \boldsymbol{\psi}\left(\boldsymbol{x}_0 + \boldsymbol{c}_g\right)$

However, this approach is doomed to fail according to S. Mallat [Mal16; Cam20], since the template image is not accessible nor defined in some cases. S. Mallat proposes an alternative approach, i.e., he proposes to remove the diffeomorphism, by selecting a feature space F orthogonal to the weight space Θ:

$$\begin{aligned} h_\theta(\mathbb{X}) = \langle \boldsymbol{z}(g\mathbb{X}), \boldsymbol{w} \rangle = h_\theta(g\mathbb{X}) = \langle \boldsymbol{z}(g\mathbb{X}), \boldsymbol{w} \rangle \\ \iff \langle \boldsymbol{z}(\mathbb{X}) - \boldsymbol{z}(g\mathbb{X}), \boldsymbol{w} \rangle \quad \forall \boldsymbol{w} \in \Theta \end{aligned} \tag{124}$$

Retrieving the infinitesimal strain theory in Equation (120) and the Taylor expansion in Equation (117) with $t = -1$, Equation (124) can be linearized as follows:

$$\begin{aligned} \langle \boldsymbol{z}(\mathbb{X}) - \boldsymbol{z}(g\mathbb{X}), \boldsymbol{w} \rangle = \langle \boldsymbol{z}(\boldsymbol{\psi}(\boldsymbol{x})) - \boldsymbol{z}(\boldsymbol{\psi}(\boldsymbol{\varphi}(\boldsymbol{x}; -1))), \boldsymbol{w} \rangle \approx \\ \approx -\mathbb{D}_{\boldsymbol{\psi}} \boldsymbol{z}(\boldsymbol{\psi}(\boldsymbol{x}))(\boldsymbol{\psi}(\boldsymbol{\varphi}(\boldsymbol{x}; -1))) \quad \forall \boldsymbol{w} \in \Theta \end{aligned} \tag{125}$$

In other words, Equation (125) highlights the fact that the gradient of the hidden features $\boldsymbol{z}$ engenders the space $\mathcal{Z} = \Theta^{\perp}$, for all images $\mathbb{X} \in L(f)$, identified via the learning process that determines the optimum weights $\boldsymbol{w}$. A trivial strategy to achieve this result would be to force the invariance of $\boldsymbol{z}(g\mathbb{X}) = \boldsymbol{z}(\mathbb{X})$ with the respect of any group action, so that $\boldsymbol{z}(\mathbb{X}) = \boldsymbol{z}(g\mathbb{X}) = \boldsymbol{z}(\boldsymbol{\psi}(\boldsymbol{x}_0))$. However, this requirement can be rather prohibitive to achieve. A weaker condition is to assure that the feature a Lipschitz functions [Mal16; Cam20], in order to assure they remains "close" for infinitesimal group action, i.e., the following property must hold:

$$\| \boldsymbol{z}(\mathbb{X}) - \boldsymbol{z}(g\mathbb{X}) \| \leq C \left(\sup_{\boldsymbol{x} \in \Omega_t} \| \boldsymbol{u}(\boldsymbol{x}) \| + \sup_{\boldsymbol{x} \in \Omega_t} \| \mathbb{D}_{\boldsymbol{x}} \boldsymbol{u}(\boldsymbol{x}) \| \right) \| \boldsymbol{z}(\mathbb{X}) \| \tag{126}$$

Therefore, provided that a small strain and/or a small translations is applied to the image, if the hidden features are Lipschitz, their variation are bounded. However, natural images can be rather discontinuous, which hinders the extraction or Lipschitz features that are stable against infinitesimal groups actions. In this case, seemingly small strain and translations can definitely cause a mismatch between the original and deformed image and the distance

$$\begin{aligned} \| \boldsymbol{z}(\mathbb{X}) - \boldsymbol{z}(g\mathbb{X}) \| = \| \boldsymbol{z}(\mathbb{X}) \| + \| \boldsymbol{z}(g\mathbb{X}) \| - \\ -2 \sum_{i=1}^{W} \sum_{j=1}^{H} \langle \boldsymbol{z}(\boldsymbol{\psi}(i,j)), \boldsymbol{z}(\boldsymbol{\psi}(i,j)) \rangle \approx \| \boldsymbol{z}(\mathbb{X}) \| \end{aligned} \tag{127}$$

because the projection of a features onto the new one is small because the strain is not "small" enough to assure the feature stability. There is no absolute limit amplitude for diffeomorphism to assure the fact that the features remain Lipschitz.

If one considers a linear diffeomorphisms, with $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\varepsilon}.\boldsymbol{x}$, with a small symmetric strain tensor $\boldsymbol{\varepsilon} = \varepsilon_{11}\boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \varepsilon_{22}\boldsymbol{e}_2 \otimes \boldsymbol{e}_2$, with $0 \leq \varepsilon_1 \ll 1$, $0 \leq \varepsilon_2 \ll 1$, representing constant bi-directional stretching, Equation (120) reads:

$$\boldsymbol{\psi}(\boldsymbol{\varphi}(\boldsymbol{x}; -1)) = \boldsymbol{\psi}(\boldsymbol{x}) + \mathbb{D}_{\boldsymbol{x}}\boldsymbol{\psi}(\boldsymbol{x})((\mathbf{I} - \boldsymbol{\varepsilon})\boldsymbol{x}) \tag{128}$$

If one considers features $\boldsymbol{z}(\mathbb{X})$ obtained by Fourier transform of the image, defined as follows:

$$\boldsymbol{z}(\mathbb{X}) = \mathcal{F}(\mathbb{X}) = \int_{\mathbb{R}^2} \boldsymbol{\psi}(\boldsymbol{x}) e^{-i\langle \boldsymbol{k}, \boldsymbol{x}\rangle} d\boldsymbol{x} = \hat{\boldsymbol{\psi}}(\boldsymbol{k}) \tag{129}$$

the features corresponding to the bi-directionally stretched image read:

$$\begin{aligned} \boldsymbol{z}(g\mathbb{X}) = \mathcal{F}(g\mathbb{X}) &= \int_{\mathbb{R}^2} \boldsymbol{\psi}(\boldsymbol{x} - \boldsymbol{\varepsilon}.\boldsymbol{x}) e^{-i\langle \boldsymbol{k}, \boldsymbol{x}\rangle} d\boldsymbol{x} = \\ &= \det{}^{-1}(\mathbf{I} - \boldsymbol{\varepsilon}) \int_{\mathbb{R}^2} \boldsymbol{\psi}(\boldsymbol{x}) e^{-i\left\langle (\mathbf{I}-\boldsymbol{\varepsilon})^{-T}\boldsymbol{k}, \boldsymbol{x}\right\rangle} d\boldsymbol{x} = \\ &= \frac{1}{(1-\varepsilon_1)(1-\varepsilon_2)} \hat{\boldsymbol{\psi}}\left((\mathbf{I} - \boldsymbol{\varepsilon})^{-T}\boldsymbol{k}\right) \end{aligned} \tag{130}$$

According to Equation (130), a bi-directional stretch smoothens the image, since high-frequencies features are contracted by a factor $(\mathbf{I} - \boldsymbol{\varepsilon})^{-T}$ towards the low frequencies and amplified by a factor $\frac{1}{(1-\varepsilon_1)(1-\varepsilon_2)}$. Therefore, even for small strain, the Fourier features are not necessarily stable by stretching (see Section 3.1.2). The Lipschitz condition is strictly related with features that assure the scale separability, which is obtained by wavelet transform in combination with non-linear activations (see [Mal09]).

3.2 Average Pooling

The most trivial strategy to force invariance by translation and dilatation ($\boldsymbol{\varphi} = \boldsymbol{x} + \boldsymbol{c} + \operatorname{diag}(\alpha_1, \alpha_2, 0)\boldsymbol{x}$), mentioned in Section 3.1.2 is to apply consecutively multiple group actions to each feature map and sum the contribution, such as :

$$\sum_{g \in G} g.\mathbb{X} = \sum_{g \in G} \boldsymbol{\psi}(\boldsymbol{x} - \boldsymbol{c}_g) \tag{131}$$

because Equation (131) represents the average image, that can be rewritten as $\sum_{g\in G} \boldsymbol{\psi}\left(\boldsymbol{x}_g^{'}\right)$ by change of variables $\boldsymbol{x}_g^{'}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{c} + \operatorname{diag}(\alpha_1, \alpha_2, 0)\boldsymbol{x} - \boldsymbol{c}_g$. Another interesting interpretation is given by the fact that the Fourier transform of a translated vector $\boldsymbol{\psi}(\boldsymbol{x} - \boldsymbol{c}_g)$ in $\mathbb{R}^2$ is independent of $\boldsymbol{c}_g$ in the following condition [Cam20]:

$$\mathcal{F}(\boldsymbol{\psi}(\boldsymbol{x} - \boldsymbol{c}_g)) = e^{i\langle \boldsymbol{k}, \boldsymbol{c}_g\rangle} \cdot \mathcal{F}(\boldsymbol{\psi}(\boldsymbol{x})) = \mathcal{F}(\boldsymbol{\psi}(\boldsymbol{x}))\, \forall \boldsymbol{c}_g \iff \boldsymbol{k} = \mathbf{0} \tag{132}$$

with $\hat{\boldsymbol{\psi}}(\mathbf{0}) = \mathcal{F}(\boldsymbol{\psi}(\boldsymbol{x}))(\mathbf{0})$ being the average of the $\boldsymbol{\psi}(\boldsymbol{x})$. Averaging the image cancels all local image features and all the previously applied group actions

along the group orbit (and of their order).

To avoid this loss of information, $\boldsymbol{z}(\mathbb{X})$ can be obtained by performing an averaging as in Equation (131) on a linear combination of N_c *feature maps* $z_c(\mathbb{X})$, corresponding to different *channels*. In other words, according to S. Mallat [Cam20], the features are obtained by *average pooling*, defined as:

$$\begin{aligned}\boldsymbol{z}(\mathbb{X}) &= \sum_{c=1}^{N_c} \sum_{g \in G} \boldsymbol{e}_c z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x} - \boldsymbol{c}_g\right)\right) = \\ &= \sum_{c=1}^{N_c} \left(\sum_{g \in G} g z_c(\mathbb{X})\, \boldsymbol{e}_c \right) = AvgPooling\left(\sum_{c=1}^{N_c} z_c(\mathbb{X})\, \boldsymbol{e}_c \right)\end{aligned} \tag{133}$$

Average pooling in Equation (133) grants the sought invariance by translation (or other group action) provided that each feature map z_c (each channel) is equivariant by translation:

$$\begin{aligned}\sum_{c=1}^{N_c} z_c(\mathbb{X})\, \boldsymbol{e}_c = \sum_{c=1}^{N_c} \left(\sum_{g \in G} z_c(g\mathbb{X})\, \boldsymbol{e}_c \right) = \sum_{c=1}^{N_c} \left(\sum_{g \in G} z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x} - \boldsymbol{c}_g\right)\right) \boldsymbol{e}_c \right) = \\ = \sum_{c=1}^{N_c} \left(\sum_{g \in G} g z_c(\mathbb{X})\, \boldsymbol{e}_c \right)\end{aligned} \tag{134}$$

The equivariance by translation of each feature map suggests that each channel should be rendered by a covariant operator, such as the linear convolution (see Figure 6). In other words, features $\boldsymbol{z}$ are invariant by group action (translation, rotation, diffeomorphisms) thanks to average pooling performed on feature maps that commute with the same group action, i.e., if the image is deformed, each feature map follows the same deformation, according to the following expression:

$$z_c(g\mathbb{X}) = z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x} - \boldsymbol{c}_g\right)\right) = z_c\left(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{c}_g\right) = g z_c(\boldsymbol{x}) \tag{135}$$

Pooling layers can be applied on the top of a standard $\mathcal{MLP}$ of the type described in Section 2.2, for classification purposes. However, $\mathcal{MLP}$ are not necessarily producing translation-equivariant output. The average pooling output is independent of translations and dilatation, so that the classification is not affected (see Section 2.5 for technical details about the *AvgPooling* implementation). $\boldsymbol{c}_g$ is called *stride* and it can be larger than 1 pixel, which consists into an average operation with subsampling. Moreover, the average operation does not require the previous knowledge of G such as in the approach that learns the canonical invariants (see the recentering in Equation (123)).

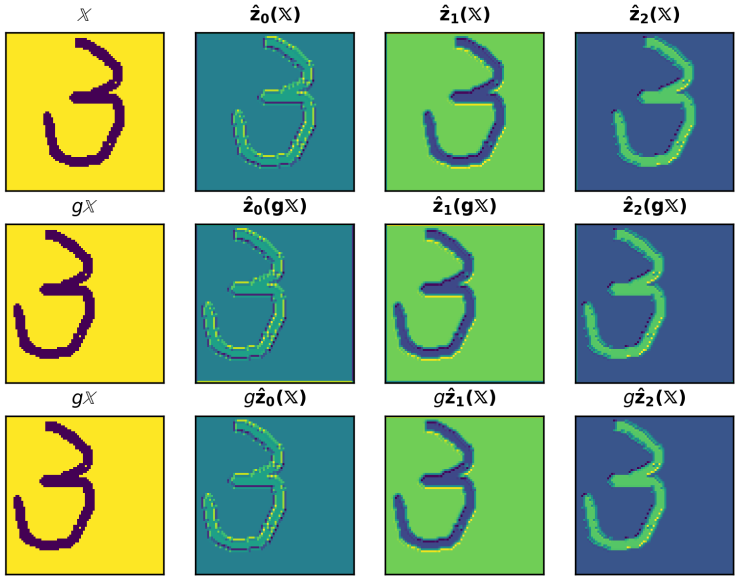

Figure 6: Convolutional feature maps extracted from a digit image belonging to the `Omniglot` dataset (https://www.omniglot.com) and translated by using the `elasticdeform` library (https://elasticdeform.readthedocs.io/en/latest/license.html).

3.3 Convolutional layer

Assuming equivariant group action on the feature maps, the effect of any further group actions g' on the feature maps vanishes (because of the sum over all the groups):

$$\boldsymbol{z}\left(g'\mathbb{X}\right)=\sum_{c=1}^{N_c}\sum_{g\in G}z_c\left(g\left(g'\mathbb{X}\right)\right)\boldsymbol{e}_c=\sum_{c=1}^{N_c}\sum_{g\in G}g'z_c\left(g\mathbb{X}\right)\boldsymbol{e}_c \tag{136}$$

In order to force the equivariance to translation (but also with the respect to rotation and other small strain) of each feature maps $\boldsymbol{z}\left(\mathbb{X}\right)=\sum_{c=1}^{N_c}z_c\left(\mathbb{X}\right)\boldsymbol{e}_c$, the latter must be the result of a *convolution*, as defined in Appendix A and reported in the following expression:

$$\boldsymbol{z}\left(\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right)=\int_{\mathbb{R}^2}\boldsymbol{H}\left(\boldsymbol{u}\right).\boldsymbol{\psi}\left(\boldsymbol{x}-\boldsymbol{u}\right)d\boldsymbol{u}+\boldsymbol{b} \tag{137}$$

with $\boldsymbol{H}$ the causal convolution *kernel* (i.e. the response function of the LTI system) and with $\boldsymbol{b}$ the filter bias. $\boldsymbol{H}$ has usually a compact support. Since the image is a discrete signal defined by via the sampling functions s_1 and s_2 as (see Equation (226)):

$$\boldsymbol{\psi}\left[i,j\right]=s_1\left(\frac{x_{1i}}{W}\right)\cdot s_2\left(\frac{x_{2j}}{H}\right)\mathbf{I}\star\boldsymbol{\psi}\left(\boldsymbol{x}\right) \tag{138}$$

the feature maps are the result of a discrete convolution that reads:

$$\boldsymbol{z}\left[j,i\right]=\sum_{u=0}^{k_W-1}\sum_{v=0}^{k_H-1}\boldsymbol{H}\left[u,v\right].\boldsymbol{\psi}\left[i-u,j-v\right]+\boldsymbol{b},\quad (i,j)\in[\![0,W]\!]\times[\![0,H]\!] \tag{139}$$

with $f\left[u,v\right]=f\left(\frac{u}{W},\frac{v}{H}\right)$ and $\boldsymbol{z}$ being stored in `C` order, contiguous along the rows. Each feature map $z_c\left(\mathbb{X}\right)=\langle\boldsymbol{H}\star\boldsymbol{\psi},\boldsymbol{e}_c\rangle+b_c$ corresponds to a LTI filter on the image, for a total of N_c filters. Refer to Section 2.5 for an extensive excursus on practical aspects related to the the implementation of discrete convolution.

Remark 30. Strided convolutions, defined as:

$$\boldsymbol{z}\left(\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right)=\int_{\mathbb{R}^2}\boldsymbol{H}\left(\boldsymbol{u}\right).\boldsymbol{\psi}\left(\mathrm{diag}\left(s_W,s_H,0\right)\boldsymbol{x}-\boldsymbol{u}\right)d\boldsymbol{u} \tag{140}$$

with strides $s_W\geq 1$ and $s_H\geq 1$, are a way of reducing the dimensionality of the data, but they represent a loss of information (since some pixels are skipped) and they disrupt the translation-equivariance (or covariance, stated in Equation (199)) because they infringe the Nyquist-Shannon theorem [AW19]. In other words, when the convolution is strided, if the underlying image is translated, the resulting feature maps do not match the translated feature maps of the original image. As a matter of fact, discrete signals should be the result of a sampling at at least twice the highest frequency in the input analog signal, in order to grant its correct reconstruction. In the case of strided convolution, the image/feature map is sampled according ot the stride values, disregarding some of the information contained in the original signal. In order to counteract this problem, the number of feature maps increases accordingly, but still they are not translation-equivariant [MMD20]. On the other hand, downsampling due to non-strided convolutions or pooling, break the hypothesis (valid for continuous signals) of translation invariance [SG20]. This effect is mitigated by adding zero-padding at the edge of the image.

Remark 31. Despite breaking feature maps' equivariance, subsampling is still largely employed with success. According to [MMD20], this is mostly because of the subsampling "shiftability" property. Shiftability is defined as the translation scaled by the associated subsampling stride, i.e.:

$$z_c(g\mathbb{X})=z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x}-\boldsymbol{c}_g\right)\right)=z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x}\right)-\mathrm{diag}\left(\frac{1}{s_W},\frac{1}{s_H},0\right)\boldsymbol{c}_g\right) \tag{141}$$

with s_W and s_H being the stride along the image width and height respectively. In other words, the subsampled feature maps obtained by strided convolution are equivariant to all group translations by a vector that is a fraction of the stride. Shiftability is a scaled version of the translation-equivariance. For

a stride couple (s_W, s_H), there exist $s_W \cdot s_H$ equivariant translations of the feature map, i.e. $\forall(\alpha_W, \alpha_H)$, with $1 \leq \alpha_W \leq s_W$ and $1 \leq \alpha_H \leq s_H$ the following expression holds:

$$z_c(g\mathbb{X}) = z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x} - \operatorname{diag}\left(\alpha_W, \alpha_H\right)\boldsymbol{c}_g\right)\right) = z_c\left(\boldsymbol{\psi}\left(\boldsymbol{x}\right) - \operatorname{diag}\left(\frac{\alpha_W}{s_W}, \frac{\alpha_H}{s_H}, 0\right)\boldsymbol{c}_g\right) \tag{142}$$

Equation (142) proves that subsampling reduces the so called "image movement" [MMD20], defined as the amount of translation of the feature map from the untranslated position $\boldsymbol{x}_0$: shiftable feature maps are equivariant to $\alpha_W \geq 1$ and $\alpha_H \geq 1$ times larger. However, subsampling reduces the so called "image similarity" due to the information loss of skipping pixels. The reduced similarity prevents the invariance to translation of feature maps. The preserved similarity depends on the "local homogeneity" of the image [MMD20]: the larger the correlation length of $\boldsymbol{\psi}\left(\boldsymbol{x}\right)$, the higher the local homogeneity and the larger will be the preserved similarity despite subsampling. However, the input image $\boldsymbol{\psi}\left(\boldsymbol{x}\right)$ should not be treated explicitly (e.g., by applying a Gaussian blur). Instead, feature maps remain "similar" by using strided pooling: the stride (subsampling) reduces the "image movement", while, at the same time, pooling increases local homogeneity, by averaging over the kernel [MMD20].

The role of non-linear activation Convolutional filters are applied to the image or to the feature maps of the previous layers, followed by non-linear activation functions γ[16] that are also equivariant by translation and small strain (see Figure 7):

$$\gamma\left(z_c\left(g\mathbb{X}\right)\right) = g\gamma\left(z_c\left(\mathbb{X}\right)\right) \tag{143}$$

The easiest choice is to adopt pixel-wise non-linear activation functions [Cam20].

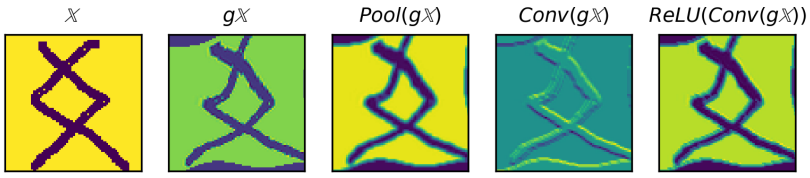

Figure 7: Action of the average pooling, convolution and convolution with $ReLU$ activation on a distorted image $g\mathbb{X}$.

Moreover, S. Mallat suggests that non-linear activation functions can be adopted to threshold local pattern (or noise).

[16] γ is used instead of the standard notation g, in order to avoid confusion with the group action.

4 Optimizing a Neural Network

4.1 Gradient Descent (GD) algorithm: first order approach

If one considers a function $L_{\mathcal{D}_{XY}} : \Theta \to \mathbb{R}$, with $\Theta \subset \mathbb{R}^{d_\Theta}$ and $f \in C^1(\Theta)$, the Taylor expansion expressed by the following expression holds:

$$T_{\hat{\boldsymbol{\theta}}} L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) = L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right) + \left\langle \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\rangle + o\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|\right) \tag{144}$$

with $o\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|\right)$ being a quantity such that $\lim_{\|\boldsymbol{\theta} - \hat{\boldsymbol{x}}\| \to 0} \frac{o\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|\right)}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|} = 0$. Being K convex, the following function is positive (by Theorem 73):

$$\left\langle \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\rangle \geq 0, \quad \boldsymbol{\theta} \in K \tag{145}$$

This means that, at the first order,

$$T_{\hat{\boldsymbol{\theta}}} L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) \geq L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right)$$

and

$$\left\langle \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\rangle = \mathcal{O}\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|\right)$$

. In this sense, to minimize $L_{\mathcal{D}_{XY}}$ on a convex set K one needs to search in the direction of $-\boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}$, i.e., at the first order:

$$\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \eta \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right) \tag{146}$$

with $\eta \in \mathbb{R}^+$ being the *learning rate*, and replacing it in Equation (145), the Euler's inequality is satisfied and $L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}(\eta)) \geq L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \quad \forall \eta \in \mathbb{R}^+$. In alternative, the following function $\xi : \mathbb{R}^+ \times \Theta \to \mathbb{R}$

$$\xi : (\eta, \boldsymbol{\theta}) \mapsto L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta} - \eta \cdot \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})\right) \tag{147}$$

gives an idea of how $L_{\mathcal{D}_{XY}}$ evolves along the half-line directed along $-\boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})$. The Taylor expansion of ξ with the respect to η reads:

$$T_0 \xi(\eta; \boldsymbol{\theta}) = L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) - \eta \|\boldsymbol{\nabla}_x L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})\|^2 + o(\eta) \tag{148}$$

In this sense, provided $\eta > 0$ small enough, $L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta} - \eta \cdot \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})\right) \leq L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})$. The choice of the gradient is even the optimal one, due to the following result (derived from Equation (145):

$$-r \frac{\boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})}{\|\boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})\|} = \arg\min_{\|\boldsymbol{\delta\theta}\| = r} L_{\mathcal{D}_{XY}}(\boldsymbol{\theta} + \boldsymbol{\delta\theta}) \tag{149}$$

Therefore, the standard Gradient Descent algorithm (GD) iteratively updates the weights according to the following scheme (inspired by Equation (146)):

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta^{(i)} \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} \right), \quad \eta^{(i)} \in \mathbb{R}^{+} \tag{150}$$

This choice is particularly attractive from a computational standpoint, since:

$$\boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} - \eta^{(i)} \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} \right) \right) = \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i+1)} \right) \tag{151}$$

However, $\eta^{(i)}$ must be small enough to grant the Euler's inequality

$$T_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}} \left(\hat{\boldsymbol{\theta}} \right) \leq L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta} \right)$$

to be satisfied at the first order. η can be updated at each iteration. For instance, the greedy choice of $\eta^{(i)}$ reads:

$$\eta^{(i+1)} = \underset{\eta > 0}{\arg\min}\ \xi \left(\eta; \boldsymbol{\theta}^{(i)} \right) \tag{152}$$

The greedy choice is reached when

$$\xi' \left(\eta^{(i)}, \boldsymbol{\theta}^{(i)} \right) = \left\langle \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} - \eta^{(i)} \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} \right) \right), \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i+1)} \right) \right\rangle = 0$$

which is reached when the gradients of the loss function at two consecutive epochs are orthogonal.

4.2 Gradient Descent algorithm: second order approach

If $L_{\mathcal{D}_{XY}}$ is twice differentiable, then its Taylor's expansion at the second order reads:

$$\begin{aligned} T_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\delta\theta} \right) = L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta} \right) + \left\langle \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta} \right), \boldsymbol{\delta\theta} \right\rangle + \\ \frac{1}{2} \left\langle \mathbf{H}_{L_{\mathcal{D}_{XY}}} \left(\boldsymbol{\theta} \right) . \boldsymbol{\delta\theta}, \boldsymbol{\delta\theta} \right\rangle + o(\| \boldsymbol{\delta\theta} \|^2) \end{aligned} \tag{153}$$

Analogously, the Hessian $\mathbf{H}_{L_{\mathcal{D}_{XY}}} \left(\boldsymbol{\theta} \right)$ appears in the Taylor's expansion of the gradient:

$$T_{\boldsymbol{\theta}} \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\delta\theta} \right) = \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta} \right) + \mathbf{H}_{L_{\mathcal{D}_{XY}}} \left(\boldsymbol{\theta} \right) . \boldsymbol{\delta\theta} + o(\| \boldsymbol{\delta\theta} \|) \tag{154}$$

The Hessian $\mathbf{H}_{L_{\mathcal{D}_{XY}}} \left(\boldsymbol{\theta} \right)$ is symmetric. Moreover, if the Hessian is positive semi-definite (i.e., non-negative eigenvalues), then $L_{\mathcal{D}_{XY}}$ is convex. $L_{\mathcal{D}_{XY}}$ is strictly convex is the Hessian's eigenvalues are strictly positive (see Remark 58). Therefore, the Hessian allows to decide whether $\boldsymbol{\theta}$ is a local minimum (positive semi-definite, see Theorem 75) or a global minimum (positive definite, see Theorem 78). Moreover, second order methods are based on the Hessian, i.e.,

the so called *Newton's method*, which are based on the general gradient descent algorithm:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mathbf{H}_{L_{\mathcal{D}_{XY}}}^{-1}\left(\boldsymbol{\theta}^{(i)}\right).\boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right) \tag{155}$$

Second order methods, if the Hessian is positive definite, converge faster than the first method in Equation (150), but more computationally expensive, since the Hessian must be computed and inverted at each epoch [Pey20]. The first order approximation defined by Equation (150) is a special case of the generalized Gradient Descent in Equation (150), for $\mathbf{H}_{L_{\mathcal{D}_{XY}}}\left(\boldsymbol{\theta}^{(i)}\right) = \eta^{(i)}\mathbf{I}$.

Remark 32. The optimal conditioning is achieved for Hessians that are close to the identity. In the supervised case, if the loss function $L_{\mathcal{D}_{XY}}$ is interpreted as the average log-likelihood of the parametric probability distribution $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}\left(\boldsymbol{y} = \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)$, under certain regularity conditions for $p_{\boldsymbol{\theta}}$ (see Section 1.3 for further details), the Hessian can be interpreted as the Fisher information associated to $\boldsymbol{h}_{\boldsymbol{\theta}}$. Moreover, if $p_{\boldsymbol{\theta}}$ belongs to the family of exponential probability distributions (among which, there it exists the probability distribution that maximizes the conditional entropy $\mathbb{H}(\boldsymbol{Y}|\boldsymbol{X})$, according to the principle of maximum entropy exposed in Section 1.5), of the type

$$p_{\boldsymbol{\theta}} = \frac{e^{-\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle}}{Z}$$

then the Hessian corresponds to the minus covariance of the observations $-Cov_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}[\boldsymbol{y}(\boldsymbol{X})]$ [Cam22]. In this case, each diagonal term of the Hessian corresponds to the variance $Var_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}(y_k(\boldsymbol{X}))$. Therefore, in order to facilitate the optimal conditioning and force the Hessian to approximate the identity, one can perform the so called *batch normalization* (BatchNorm), which consists into normalizing y_k and replace it by:

$$y_k^{'} = \frac{y_k - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}[y_k]}{Var_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}(y_k(\boldsymbol{X}))} \tag{156}$$

The BatchNorm operation eases the convergence of the gradient descent algorithm, since it replaces the diagonal terms of the Hessian by 1, without computing the Hessian explicitly. In $\mathcal{NN}$, BatchNorm is often applied after the pre-activation at each layer, so to force the Hessian diagonal to 1. However, this is not always sufficient, since in complex $\mathcal{NN}$ with non-linear activation functions, the Hessian (or the FIM, see Equation (24)) is not expressed in its orthonormal basis, which is unknown *a priori*. Therefore, there could be extra diagonal terms that prevent the optimal conditioning of the Hessian.

The analogy with the classical pre-conditioning methods in linear algebra is evident. Replacing $\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)} = \boldsymbol{\delta\theta}$ in Equation (153), the following second

order Taylor's expansion is obtained:

$$\begin{aligned} T_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i+1)}-\boldsymbol{\theta}^{(i)}\right) =& L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right)- \\ &-\left\langle \boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right), \mathbf{H}_{L_{\mathcal{D}_{XY}}}\left(\boldsymbol{\theta}^{(i)}\right).\boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right)\right\rangle + \\ &+o(\|\mathbf{H}_{L_{\mathcal{D}_{XY}}}\left(\boldsymbol{\theta}^{(i)}\right).\boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right)\|) \end{aligned} \tag{157}$$

Given the expression in Equation (157), if $\mathbf{H}_{L_{\mathcal{D}_{XY}}}\left(\boldsymbol{\theta}^{(i)}\right)$ is positive definite and $\boldsymbol{\nabla}_{\theta} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right) \neq \mathbf{0}$, then $L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i+1)}\right) < L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right)$, which means that the Newton's method are effectively attempting to iteratively minimize the Empirical Loss function [Pey20].

4.3 Convergence analysis of first order methods

Theorem 33. ***Convergence analysis of functions with Lipschitz gradients*** *[Nes83]*
A function $f : K \to \mathbb{R}$, defined over a non-empty convex set $K \subset H$, with H being a Hilbert space, f proper and strictly convex on K, with $f \in C^1(K)$ and a gradient $\boldsymbol{\nabla}_x f \in Lip(K)$ with Lipschitz constant β. If $\exists \delta_{\min}, \delta_{\max} \in \mathbb{R}^+$ such that:

$$0 < \frac{1}{\beta} = \delta_{\min} \leq \delta_k < \delta_{\max} = \frac{2}{\beta}$$

then $\exists \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \delta_k \boldsymbol{\nabla}_x f(\boldsymbol{x}_k)$ that converges to $\hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in K}{\arg\min} f(\boldsymbol{x})$ and there exists a constant $A > 0$ such that

$$f(\boldsymbol{x}_k) - f(\hat{\boldsymbol{x}}) \leq \frac{cst.}{k+1} \tag{158}$$

Furthermore, if f is strongly convex of coefficient α, there exists $0 \leq \rho < 1$ such that:

$$\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\| \leq \rho^k \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\| \tag{159}$$

with $0 < \rho \leq 1$.

Proof. Following the proof presented by [Pey20], since f is strongly convex on a convex set K, by Theorem 77 it exists a unique minimizer $\hat{\boldsymbol{x}}$. Moreover, the strict convexity implies Item 2. The fact that $\boldsymbol{\nabla}_x f \in \text{Lip}(K)$ implies that $\forall (\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \in K^2$:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{\beta}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \tag{160}$$

as stated by the upper bound in Proposition 81. Replacing the update scheme $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\delta_k \boldsymbol{\nabla}_x f(\boldsymbol{x}_k)$ in Equation (160), one obtains the following expression:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \delta_k \left(\frac{\delta_k \cdot \beta}{2} - 1 \right) \| \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 \tag{161}$$

A necessary requirement to pursue the minimizer of f is $\delta_k < \frac{2}{\beta} = \delta_{\max}$, so that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$ and therefore $\tau_k = \delta_k \left(\frac{\delta_k \beta}{2} - 1 \right) < 0$. To prove the existence of a lower bound $\delta_{\min}$ the convexity property Item 2 must be evoked:

$$f(\hat{\boldsymbol{x}}) \geq f(\boldsymbol{x}_k) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_k), \hat{\boldsymbol{x}} - \boldsymbol{x}_k \rangle \tag{162}$$

Comparing Equation (161) and Equation (162), replacing the update scheme $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\delta_k \boldsymbol{\nabla}_x f(\boldsymbol{x}_k)$ and considering that $\tau_k < 0$ leads to:

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\hat{\boldsymbol{x}}) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_k), \boldsymbol{x}_k - \hat{\boldsymbol{x}} \rangle + \tau_k \| \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 = \\ &= f(\hat{\boldsymbol{x}}) + \frac{\| \boldsymbol{x}_k - \hat{\boldsymbol{x}} + 2\tau_k \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 - \| \boldsymbol{x}_k - \hat{\boldsymbol{x}} \|^2}{4\tau_k} = \\ &= f(\hat{\boldsymbol{x}}) + \frac{\| \boldsymbol{x}_k - \hat{\boldsymbol{x}} \|^2 - \| \boldsymbol{x}_k - \hat{\boldsymbol{x}} + 2\tau_k \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2}{4|\tau_k|} = \\ &= f(\hat{\boldsymbol{x}}) + \frac{\| \boldsymbol{x}_k - \hat{\boldsymbol{x}} \|^2 - \| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} + \delta_k (\delta_k \beta - 1) \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2}{4|\tau_k|} \end{aligned} \tag{163}$$

From Equation (163), it must be noted that, in order to assure that $f(\boldsymbol{x}_{k+1}) \geq f(\hat{\boldsymbol{x}})$, one must assure that:

$$\| \boldsymbol{x}_k - \hat{\boldsymbol{x}} \|^2 - \| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} + \delta_k (\delta_k \beta - 1) \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 \geq 0 \tag{164}$$

Adopting the well known relation $\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2 \leq \|\boldsymbol{a} - \boldsymbol{b}\|^2$, one obtains:

$$\begin{aligned} \delta_k^2 (\delta_k \beta - 2)^2 \| \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 \geq \| \boldsymbol{x}_k - \hat{\boldsymbol{x}} \|^2 - \\ - \| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} + \delta_k (\delta_k \beta - 1) \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 \geq 0 \end{aligned} \tag{165}$$

and again:

$$\begin{aligned} \| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} + \delta_k (\delta_k \beta - 1) \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|^2 = \| \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \| \cdot \left(\frac{\| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} \|^2}{\| \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \|} + \right. \\ \left. + \delta_k^2 (\delta_k \beta - 1)^2 \| \boldsymbol{\nabla}_x f(\boldsymbol{x}_k) \| + 2\delta_k (\delta_k \beta - 1) \| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} \| \right) \end{aligned} \tag{166}$$

and finally, calling $\varepsilon = \frac{\| \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} \|}{\| \boldsymbol{x}_k - \hat{\boldsymbol{x}} \|}$:

$$g(\varepsilon) = \varepsilon^2 + 2\delta_k (\delta_k \beta - 1) \varepsilon + \delta_k^2 \beta^2 \left(2\delta_k^2 \beta^2 - 6\delta_k \beta + 5 \right) - 1 \geq 0 \tag{167}$$

In order to satisfy Equation (167) for all $\varepsilon > 0$, $\delta_k > \frac{1}{\beta} = \delta_{\min}$, as shown - for different values of β - in Figure 8 (blue-graded curves). Orange-graded lines in Figure 8 corresponds to values of $\delta_k < \frac{1}{\beta}$: in this case, it is possible to achieve $g(\varepsilon) > 0$ with $\beta \leq 1$, provided that $\varepsilon > \varepsilon_{\min} > 0$. This means that $\|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}\| > \varepsilon_{\min}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|$ which implies that the GD algorithm may not converge. For $\beta > 1$, the algorithm never converges for $\delta_k < \frac{1}{\beta}$.
Summing Equation (163) n_e times, one obtains:

$$\begin{aligned} n_e \left(f\left(\boldsymbol{x}_{n_e-1}\right) - f\left(\hat{\boldsymbol{x}}\right)\right) &\leq \sum_{k=0}^{n_e-1} f\left(\hat{\boldsymbol{x}}_{k+1}\right) - n_e \cdot f\left(\hat{\boldsymbol{x}}\right) \\ &\leq \sum_{k=0}^{n_e} \frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} + \delta_k\left(\delta_k\beta - 1\right)\boldsymbol{\nabla}_x f\left(\boldsymbol{x}_k\right)\|^2}{4|\tau_k|} \end{aligned} \tag{168}$$

If for instance $\delta = \delta_{\min} = \frac{1}{\beta}$, then:

$$\begin{aligned} f\left(\boldsymbol{x}_{n_e-1}\right) - f\left(\hat{\boldsymbol{x}}\right) \leq \frac{\beta}{2n_e} \sum_{k=0}^{n_e-1} \|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}\|^2 = \\ = \frac{2}{n_e\beta}\left(\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_{n_e} - \hat{\boldsymbol{x}}\|^2\right) \leq \frac{\beta\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|^2}{2n_e} \end{aligned} \tag{169}$$

Equation (169) proves Equation (158) [Pey20].
Finally, following [Pey20], in order to prove the second part of the theorem, expressed in Equation (159), f has to be strongly convex, which implies, by Item 2 and adopting the assumption that $\boldsymbol{\nabla}_x f\left(\hat{\boldsymbol{x}}\right) = \mathbf{0}$, that :

$$\begin{aligned} \frac{\alpha}{2}\|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}\| \leq f\left(\boldsymbol{x}_{k+1}\right) - f\left(\hat{\boldsymbol{x}}\right) \leq \\ \leq \frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}} + \delta_k\left(\delta_k\beta - 1\right)\boldsymbol{\nabla}_x f\left(\boldsymbol{x}_k\right)\|^2}{4|\tau_k|} \end{aligned} \tag{170}$$

Replacing, for instance, $\delta_k = \frac{1}{\beta}$ in Equation (170) it implies that:

$$\frac{\alpha}{2}\|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}\| \leq \frac{\beta}{2}\left(\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}\|\right)$$

and the following expression proves the statement in Equation (159):

$$\|\boldsymbol{x}_{n_e} - \hat{\boldsymbol{x}}\| \leq \left(\frac{\beta}{\eta + \beta}\right)^{\frac{n_e}{2}} \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\| \tag{171}$$

with $\rho = \sqrt{\frac{\beta}{\alpha+\beta}}$ and $0 < \rho \leq 1$ since $\eta \leq 0$. □

Remark 34. As noted in Remark 82, $\frac{\beta}{\alpha}$ represents the conditioning number of the Hessian matrix $\kappa(\mathbf{H}_f)$. In this sense, Equation (171) can be rewritten as:

$$\|\boldsymbol{x}_{n_e} - \hat{\boldsymbol{x}}\| \leq \left(\frac{\kappa(\mathbf{H}_f(\boldsymbol{x}))}{1 + \kappa(\mathbf{H}_f(\boldsymbol{x}))} \right)^{\frac{n_e}{2}} \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\| \tag{172}$$

For large conditioning number, the convergence of the algorithm is harder and strongly dependent on the initial guess $\boldsymbol{x}_0$.

5 On the automatic differentiation

5.1 Backward propagation via adjoint gradient operator

In order to described how the Automatic Differentiation algorithm works, some basics must be recalled.
For any tensor field (i.e. multi-linear application) of order $m+n$, denoted as $\mathbb{T} : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}$, with $\mathbb{T} \in \mathbb{R}^n \otimes \mathbb{R}^m$ and $\forall \boldsymbol{x} \in \mathbb{R}^m$, the derivative of the tensor field $\mathbb{T}(\cdot, \boldsymbol{x})$ along the vector $\boldsymbol{v} \in \mathbb{R}^m$ reads [For+15]:

$$D_{\boldsymbol{v}}\mathbb{T}(\cdot, \boldsymbol{x}) = \lim_{h \to 0} \frac{\mathbb{T}(\cdot, \boldsymbol{x} + h\boldsymbol{v}) - \mathbb{T}(\cdot, \boldsymbol{x})}{h}, \qquad h \in \mathbb{R}, \forall \boldsymbol{v} \in \mathbb{R}^m \tag{173}$$

If it exists, $D_{\boldsymbol{v}}\mathbb{T}(\cdot, \boldsymbol{x}) \in \mathbb{R}^n$ and the partial derivative along the i^{th} component $\boldsymbol{e}_i$ of the orthonormal base $\mathcal{B}_m$ reads [For+15]:

$$\frac{\partial \mathbb{T}(\cdot, \boldsymbol{x})}{\partial x_i} = D_{\boldsymbol{e}_i}\mathbb{T}(\cdot, \boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^m \tag{174}$$

The gradient of $\mathbb{T}(\cdot, \boldsymbol{x})$ is defined as the operator $\boldsymbol{\nabla}_x : \mathbb{R}^m \mapsto \mathbb{R}^n$ [For+15]:

$$\boldsymbol{\nabla}_x \mathbb{T}(\cdot, \boldsymbol{x}) : \boldsymbol{v} \mapsto D_{\boldsymbol{v}}\mathbb{T}(\cdot, \boldsymbol{x}) \tag{175}$$

The gradient is a tensor field of order $m+n+1$ and a linear operator on $\boldsymbol{v}$:

$$D_{a\boldsymbol{v}+b\boldsymbol{u}}\mathbb{T}(\cdot, \boldsymbol{x}) = \lim_{h \to 0} \frac{\mathbb{T}(\cdot, \boldsymbol{x} + h \cdot (a\boldsymbol{v} + b\boldsymbol{u})) - \mathbb{T}(\cdot, \boldsymbol{x})}{h} = a \cdot D_{\boldsymbol{v}}\mathbb{T}(\cdot, \boldsymbol{x}) + b \cdot D_{\boldsymbol{u}}\mathbb{T}(\cdot, \boldsymbol{x}) \tag{176}$$

which implies that:

$$\begin{aligned} D_{\boldsymbol{v}}\mathbb{T}(\cdot, \boldsymbol{x}) &= \sum_{i=1}^{m} \langle \boldsymbol{v}, \boldsymbol{e}_i \rangle D_{\boldsymbol{e}_i}\mathbb{T}(\cdot, \boldsymbol{x}) = \sum_{i=1}^{m} \langle \boldsymbol{v}, \boldsymbol{e}_i \rangle \frac{\partial \mathbb{T}(\cdot, \boldsymbol{x})}{\partial x_i} = \boldsymbol{\nabla}_x \mathbb{T} . \boldsymbol{v}, \\ \boldsymbol{\nabla}_x \mathbb{T} &= \sum_{i=1}^{m} \frac{\partial \mathbb{T}(\cdot, \boldsymbol{x})}{\partial x_i} \otimes \boldsymbol{e}_i \end{aligned} \tag{177}$$

Since the gradient is a tensor $\boldsymbol{\nabla}_x \mathbb{T}(\cdot, \boldsymbol{x})$ of order $m+n+1$, it maps any vector $\boldsymbol{v} \in \mathbb{R}^m$ into its dual space $\mathbb{R}^{m*} = \mathbb{R}^n$ and the following expression holds:

$$D_{\boldsymbol{v}}\mathbb{T}^*(\boldsymbol{u}, \boldsymbol{x}) = \langle \boldsymbol{\nabla}_x \mathbb{T} . \boldsymbol{u}, \boldsymbol{v} \rangle \in \mathbb{R}, \qquad \forall \boldsymbol{v}, \boldsymbol{x} \in \mathbb{R}^m, \forall \boldsymbol{u} \in \mathbb{R}^n \tag{178}$$

In the back-propagation framework, the above mentioned basic notions related to the gradient of tensor fields can be applied to $\mathbb{T}(\cdot, \boldsymbol{x}) = L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) \in \mathbb{R}$, with $\boldsymbol{x} \equiv \boldsymbol{\theta} \in \mathbb{R}^m$. Moreover, the adjoint correspond to to transposition operator. The quest for a minimum is steered by the following well established Euler's inequality

Theorem 35. *(Necessary condition, A convex set) Given any open set $\Omega \subset \mathbb{R}^n$, a convex subset $A \subset \Omega$ and a function $f : \Omega \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$ with $f \in \mathcal{C}^1(\Omega)$, the necessary condition for $\bar{\boldsymbol{x}}$ to be a local minimum of f on A reads:*

$$\left\langle \boldsymbol{\nabla}_x f(\bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \right\rangle \geq 0, \qquad \forall \boldsymbol{y} \in A \tag{179}$$

Theorem 35 implies the iterative expression in Equation (79), provided a convex set Θ and a function $L_{\mathcal{D}_{XY}} \in \mathcal{C}^1(\Theta)$. The gradient of each weight θ_j is updated as in Equation (79):

$$\theta_j^{(i+1)} = \theta_j^{(i)} - \alpha \frac{\partial L_{\mathcal{D}_{XY}}}{\partial \theta_j}\left(\theta_j^{(i)}\right) = \theta_j^{(i)} - \alpha \cdot D_{\boldsymbol{e}_j} L_{\mathcal{D}_{XY}}(1, \boldsymbol{\theta}^{(i)}) = \theta_j^{(i)} - \alpha \cdot \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}} . \boldsymbol{e}_j \tag{180}$$

In particular, the Automatic Differentiation exploits the basics differential algebra presented above for composite non-linear functions of the type $L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) = L_{\mathcal{D}_{XY}} \circ l \circ g_i \circ a_j(\theta_k)$, adopting the chain rule in Equation (80). As a matter of fact, considering, without loss of generality, a composite function $f = h \circ g$ with $f : \mathbb{R}^m \mapsto \mathbb{R}$, $h : \mathbb{R}^m \mapsto \mathbb{R}$ and $g : \mathbb{R} \mapsto \mathbb{R}$, with $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, $g : y \mapsto z = g(y)$ and $h : \boldsymbol{x} \mapsto y = h(\boldsymbol{x})$, the chain rule in Equation (80) applies to f, g and h according to the following expressions:

$$D_{\boldsymbol{x}} f(\boldsymbol{x}) = D_y g(h(\boldsymbol{x})) . D_{\boldsymbol{v}} h(\boldsymbol{x}) = \left(\frac{dg}{dy}\right)^* (h(\boldsymbol{x})) \cdot D_{\boldsymbol{v}} h(\boldsymbol{x}) \tag{181}$$

Moreover, the adjoint derivative reads:

$$D_{\boldsymbol{v}} f^*(\boldsymbol{x}) = D_{\boldsymbol{v}} h^*(\boldsymbol{x}) . D_y g^*(h(\boldsymbol{x})) = D_{\boldsymbol{v}} h^*(\boldsymbol{x}) \frac{dg}{dy}(h(\boldsymbol{x})) \tag{182}$$

Replacing Equation (182) in Equation (178), the gradient of f is immediately computed. It is worth noticing that, once the *forward pass* $L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) = L_{\mathcal{D}_{XY}} \circ l \circ g_i \circ a_j(\theta_k)$ is performed, the *backward pass* in Equation (182) is obtained with no further computation, since $D_{\boldsymbol{v}} h^*(\boldsymbol{x}) = D_{\boldsymbol{v}} h^T(\boldsymbol{x})$ and $\left(\frac{dg}{dy}\right)^* = \frac{dg}{dy}$. Moreover, each adjoint operation is local to the tensor itself, since it performed by simply inverting the graph constructed in the *forward pass*, i.e., by flipping the input-output connection. Therefore, the standard Automatic Differentiation implementation[17] bundles the gradients of a tensor to the tensor itself, in order to reduce the memory consumption (no duplicate tensors) and exploit the same graph (see Figure 43).

[17] https://autodiff.github.io/

5.2 Counteract vanishing gradients with ELU and $SELU$ activation functions

A valid alternative to $ReLU$ is represented by Exponential Linear Unit [CUH16] ELU, depicted in Figure 9. The ELU function reads:

$$g(a) = ELU(a) = \begin{cases} a, a > 0 & (183) \\ \alpha \cdot (e^a - 1), a \leq 0 & (184) \end{cases}$$

ELU is well known to prevent the so called *bias shift* [CUH16]. In order to explain the latter, the analogy between Empirical Loss function and negative log-likelihood exposed in Remark 22 must be retrieved. Provided this analogy, if one focuses on the k^{th} hidden layer, with $N_\ell \geq k > 1$ the derivative $\frac{\partial L_{\mathcal{D}_{XY}}}{\partial \theta_c^{(k)}}$ can be seen a the derivative of the negative log-likelihood $\frac{\partial \ln p_\theta(y|\boldsymbol{x})}{\partial \theta_c^{(k)}}$, for any label y in the database.
For the sake of clarity, let us consider layer the weights and biases of each of the $u^{(k+1)}$ neurons in the k^{th} layer, defined by the vector $\boldsymbol{\theta}^{(k)}$, that reads:

$$\begin{aligned} \boldsymbol{\theta}^{(k)} &= \left[b^{(k)}, W_{11}^{(k)}, W_{12}^{(k)}, \ldots, W_{1u^{(k)}}^{(k)}, W_{21}^{(k)}, \ldots, W_{u^{(k+1)}u^{(k)}}^{(k)} \right] = \\ &= \left[b^{(k)}, \boldsymbol{w}^{(k)} \right] \end{aligned} \tag{185}$$

with a total amount of scalar weights/biases $n_\theta^{(k)} = u^{(k+1)} \cdot u^{(k)} + 1$.
To access each entry of $\boldsymbol{\theta}^{(k)}$, a scalar indexing function i is defined as follows:

$$\begin{aligned} & i(m,n) = 1 + n + u^{(k)} \cdot m, \quad 0 \leq m \leq u^{(k+1)}, \quad 0 \leq n \leq u^{(k)} \\ & \text{with} \\ & \theta_{i(0,0)}^{(k)} = b^{(k)} \\ & \theta_{i(m,n)}^{(k)} = W_{mn}^{(k)}, \quad 1 \leq m \leq u^{(k+1)}, \quad 1 \leq n \leq u^{(k)} \end{aligned} \tag{186}$$

considering Equation (97), Equation (104) and the derivative of the loss function with the respect to bias and weights expressed in Equation (99) and Equation (98), $\frac{\partial \ln p_\theta(y|\boldsymbol{X})}{\partial \theta_i^{(k)}}$ reads [CUH16]:

$$\frac{\partial \ln p_\theta(y|\boldsymbol{x})}{\partial \theta_i^{(k)}} = \frac{\partial L_{\mathcal{D}_{XY}}(h_\theta(\boldsymbol{x}), y)}{\partial a_i^{(k)}} \cdot \frac{\partial a^{(k)}}{\partial \theta_i^{(k)}} = \delta(\boldsymbol{x}) \cdot \frac{\partial a^{(k)}}{\partial \theta_i^{(k)}}(\boldsymbol{x}) \tag{187}$$

$\frac{\partial a^{(k)}}{\partial \theta_c^{(k)}}(\boldsymbol{x})$ is obtained by considering Equation (99) if $\theta_i^{(k)}$ is a bias and Equation (98) if $\theta_i^{(k)}$ is a weight:

$$\frac{\partial a^{(k)}}{\partial \theta_c^{(k)}} = \begin{cases} 1, & m = n = 0, \quad \theta_{i(0,0)}^{(k)} = b^{(k)} & (188) \\ h_n^{(k-1)}, & 1 \leq m \leq u^{(k+1)}, 1 \leq n \leq u^{(k)} \quad \theta_{i(m,n)}^{(k)} = W_{mn}^{(k)} & (189) \end{cases}$$

The Fisher Information Matrix (FIM, see Equation (24)) with the respect to the weights $\boldsymbol{\theta}^{(k)}$ of the k^{th} neuron, reads:

$$\begin{aligned}\left(\mathbb{I}_F^{(k)}\right)_{i_r i_c} &= \mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\frac{\partial \ln p_\theta\left(Y|\boldsymbol{X}\right)}{\partial \theta_{i_r}^{(k)}}\cdot\frac{\partial \ln p_\theta\left(Y|\boldsymbol{X}\right)}{\partial \theta_{i_c}^{(k)}}\right] = \\ &= \mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\left(\boldsymbol{X},Y\right)\cdot\frac{\partial a^{(k)}}{\partial \theta_{i_r}^{(k)}}\cdot\frac{\partial a^{(k)}}{\partial \theta_{i_c}^{(k)}}\right]\end{aligned} \tag{190}$$

The Fisher information of two weights within the k^{th} layer depend on the value of $\delta^2\left(\boldsymbol{x},y\right)\geq 0$, i.e., the probability of drawing samples $\left(\boldsymbol{x}_i,y_i\right)$ is higher for higher values of $\delta\left(\boldsymbol{x},y\right)$. The mean value in Equation (190) can be computed by observing that the samples y drawn via the MLP have a probability distribution $p_{\delta^2}\left(\boldsymbol{x},y\right)$ that reads:

$$p_{\delta^2}\left(\boldsymbol{x},y\right) = \frac{\delta\left(\boldsymbol{x},y\right)^2\cdot p_\theta}{\int_{\mathcal{X}\times\mathcal{Y}}\delta(\boldsymbol{u},u)^2\cdot p_\theta d\boldsymbol{u}du} = \frac{\delta\left(\boldsymbol{x},y\right)^2\cdot p_\theta}{\mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta\left(\boldsymbol{X},Y\right)^2\right]} \tag{191}$$

Plugging Equation (191) in Equation (190), the latter can be rewritten as:

$$\left(\mathbb{I}_F^{(k)}\right)_{i_r i_c} = \mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\left(\boldsymbol{X},Y\right)\right]\cdot\mathbb{E}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\frac{\partial a^{(k)}}{\partial \theta_{i_r}^{(k)}}\cdot\frac{\partial a^{(k)}}{\partial \theta_{i_c}^{(k)}}\right] \tag{192}$$

The so called *bias shift* is defined as the 0^{th} column of the FIM, considering only the bias columns, as follows:

$$\begin{aligned}\boldsymbol{b}_s &= \sum_{i_r=1}^{n_\theta^{(k)}}\left(\mathbb{I}_F^{(k)}\right)_{i_r i_c(0,0)}\boldsymbol{e}_{i_r} = \mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\right]\cdot\mathbb{E}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}a^{(k)}\right] = \\ &= \mathbb{C}_{(\boldsymbol{x},y)\sim p_\theta}\left(\delta^2,\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}a^{(k)}\right) + \mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}a^{(k)}\right]\cdot\mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\right]\end{aligned} \tag{193}$$

$\boldsymbol{b}_s$ captures the statistical change of mean due the bias $b^{(k)}$ because of to the correlation among input and output units. As a matter of fact, this is confirmed by the fact that the bias shift drops to zero whenever $\mathbb{E}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}a^{(k)}\right]=\mathbf{0}$ or, equivalently, whenever

$$\mathbb{C}_{(\boldsymbol{x},y)\sim p_\theta}\left(\delta^2,\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}a^{(k)}\right) = -\mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}a^{(k)}\right]\cdot\mathbb{E}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\right] \tag{194}$$

which implies that the derivative of the loss function at the output units δ and its gradient with the respect to the output of the input units are decorrelated [CUH16]. [CUH16] showed that, when computing the back-propagation (see solution $(\mathcal{G})$), one can mitigate the effect of the bias shift, by pre-multiplying the gradient $\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}}L_{\mathcal{D}_{XY}}$ by the inverse of the FIM $\mathbb{I}_F^{(k)-1}$, obtaining a weight update vector $\boldsymbol{\Delta\theta}^{(k)}$ that reads:

$$\begin{cases} \boldsymbol{\Delta w} = \left[\mathbb{I}_F^{(k)-1}\right]_{1:,1:} \left(\boldsymbol{\nabla}_w L_{\mathcal{D}_{XY}} - \Delta b^{(k)} \left[\boldsymbol{b}_s\right]_{1:}\right) & (195) \\ \Delta b^{(k)} = s \left(\dfrac{\partial L_{\mathcal{D}_{XY}}}{\partial b^{(k)}} - \left[\boldsymbol{b}_s\right]_{1:}^T \left[\mathbb{I}_F^{(k)-1}\right]_{1:1:} \boldsymbol{\nabla}_w L_{\mathcal{D}_{XY}}\right) & (196) \end{cases}$$

with

$$\begin{aligned} s = & \mathbb{E}^{-1}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\right] + \\ & + \mathbb{E}^{-1}_{(\boldsymbol{x},y)\sim p_\theta}\left[\delta^2\right] \cdot \mathbb{E}^{T}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}} a^{(k)}\right] \\ & \cdot \mathbb{V}^{-1}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}} a^{(k)}\right] \cdot \mathbb{E}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}} a^{(k)}\right] \end{aligned} \tag{197}$$

Whenever the bias shift drops to 0, the weight update in Equation (196) is equivalent to the standard update rule (see solution ($\mathcal{G}$)). The correction of the bias shift strongly depends on the correlation of the incoming units which is captured by $\left[\mathbb{I}_F^{(k)}\right]^{-1}_{11}$ and it is equivalent to shift the mean activations of the incoming units toward zero (by a term $-\alpha\mathbb{E}_{(\boldsymbol{x},y)\sim p_{\delta^2}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}^{(k)}} a^{(k)}\right]$) and scaling up the bias unit (by a factor α) [CUH16]. The correction of the bias shift proposed in Equation (196) is effective but cumbersome to compute, since the FIM can become extremely large whenever the number of hidden neurons in the k^{th} layer is large. Therefore, [CUH16] proposed to adopt an ELU activation function to mitigate the bias shift instead. As a matter of fact, $ELU(a) < 0$ for $a < 0$ and $\frac{\partial ELU(a)}{\partial a} \geq 0, \forall a$, saturating to 0 for $a \to -\infty$. The ELU units activate for $a > 0$, similarly to $ReLU$, but they keep active with a positive derivative even for $a < 0$ and negative small value that promote the bias shift correction toward zero mean and the bias scaling [CUH16] . Moreover, the ELU (as depicted in Figure 44) prevents the vanishing gradient at the same time.

An alternative to ELU is provided by the Scaled Exponential Linear Unit [Kla+17], $SELU$. The latter activation reads:

$$g(a) = SELU(a) = s\left(\max(0, a) + \min\left(0, \alpha\left(e^a - 1\right)\right)\right) \tag{198}$$

and it is depicted, along with its derivative, in Figure 10. $SELU$ has been conceived to prevent the occurrence of vanishing gradient phenomena by promoting the input/output self-centering of weights' means and variances towards 0 and 1 respectively, even in the presence of noise and perturbation, that could engender the overall gradient descent convergence. As a matter of fact, considering a generic $\mathcal{MLP}$ layer ℓ, with pre-activation $\boldsymbol{a}^{(\ell)} = \sum_{j_1}^{u^{(\ell+1)}} a_j^{(\ell)} \boldsymbol{e}_j$, for the sake of simplicity one can assume that all the pre-activation units have the same average $\mathbb{E}\left[a_j\right] = \mu_a$, $1 \leq j \leq u^{(\ell+1)}$ and the same variance $\mathbb{V}\left[a_j\right] = \sigma_a^2$, $1 \leq j \leq u^{(\ell+1)}$. If one defines the mean and variance of the j^{th} activation

$h_j(\ell) = SELU\left(a_j\right)$ respectively $\mu_h = \mathbb{E}\left[h_j\right]$ and $\sigma_h^2 = \mathbb{V}\left[h_j\right]$, then Figure 11 showed the effect of applying $SELU$ on the mean and variance of each activation h_j [Kla+17]:

$SELU$ makes the $\mathcal{MLP}$ layer self-normalizing itself, since it a stable attracting fixed point solution, depending on the values of $w = \sum_{j_1}^{u^{(\ell+1)}} a_j^{(\ell)}$ and $v = \sum_{j_1}^{u^{(\ell+1)}} a_j^{(\ell)2}$. Moreover, if the pre-activation mean and variance μ_a and σ_a^2 belong to a set $\mathcal{A} := \left\{\left(\mu, \sigma^2\right) | \mu \in \left[\mu_{\min}, \mu_{\max}\right], \sigma^2 \in \left[\sigma_{\min}^2, \sigma_{\max}^2\right]\right\}$, then $SELU\left(\mathcal{A}\right) \subseteq \mathcal{A}$, i.e., this inclusion is transitive across layers since $a_j^{(\ell)}$ is a function of $h_k^{(\ell-1)}$. $SELU$ dampens the noise and perturbations which make w and v, recentering each neuron activation mean and variance.

A LTI, Fourier transform and convolution [RECAP]

This recap section is a summary of chapter II *Fourier Kingdom* in [Mal09], to which we remand for further details.

A linear time-invariant (LTI) filter is an operator equivariant with the respect to translation (or covariant):

$$\boldsymbol{y}\left(t\right) = \mathcal{L}\left(\boldsymbol{x}\left(t\right)\right) \iff \boldsymbol{y}\left(t-\tau\right) = \mathcal{L}\left(\boldsymbol{x}\left(t-\tau\right)\right) \tag{199}$$

with $\boldsymbol{y}(t) \in \mathbb{R}^n$. From now on, $\mathcal{L}\left(\cdot\right)$ is assumed to be hold weak continuity properties, so as to assure its stability against small deviations $\boldsymbol{x}\left(t\right) + \boldsymbol{\epsilon}$. If $\boldsymbol{x}(t) \in C^0\left(\mathbb{R}^n\right)$, then $\boldsymbol{y} \in C^0\left(\mathbb{R}^n\right)$ and they can be represented by:

$$\begin{aligned} \boldsymbol{x}\left(t\right) &= \int_{\mathbb{R}} \boldsymbol{x}(\tau) \cdot \delta(t-\tau) dt \\ \boldsymbol{y}\left(t\right) &= \mathcal{L}\left(\int_{\mathbb{R}} \boldsymbol{x}(\tau) \cdot \delta(t-\tau) dt\right) = \int_{\mathbb{R}} \sum_{k=1}^{n} x_k(\tau) \cdot \mathcal{L}\left(\delta(t-\tau)\boldsymbol{e}_k\right) dt \end{aligned} \tag{200}$$

The term $\langle \mathcal{L}(\delta\left(t-\tau\right)\boldsymbol{e}_k), \boldsymbol{e}_h \rangle = H_{hk}(t-\tau)$ represents the impulse response of a LTI filter, whose components reads:

$$\boldsymbol{y}\left(t\right) = \int_{\mathbb{R}} \boldsymbol{H}\left(t-\tau\right).\boldsymbol{x}\left(\tau\right) d\tau = \int_{\mathbb{R}} \boldsymbol{H}\left(\tau\right).\boldsymbol{x}\left(t-\tau\right) d\tau \tag{201}$$

Therefore, the response $\boldsymbol{y}(t)$ is the *convolution* of the LTI response $\boldsymbol{H}$ with the input signal $\boldsymbol{x}(t)$, denoted as $\boldsymbol{H} \star \boldsymbol{x}\left(t\right) = \boldsymbol{x} \star \boldsymbol{H}(t)$.
A LTI filter implies the following properties:

- the convolution operator is commutative and translational covariant (see Section 3.2);

- the stability condition on Equation (201) implies that L is bounded if $\boldsymbol{x}$ is bounded, i.e. if $\exists M \in \mathbb{R}^{+}$ such that $\int_{\mathbb{R}} \|\boldsymbol{H}(\tau)\| d\tau \leq M$ then:

$$\begin{aligned} \|\mathcal{L}(\boldsymbol{x}(t))\| \leq \int_{\mathbb{R}} \|\boldsymbol{H}(\tau)\| \cdot \|\boldsymbol{x}(t-\tau)\| d\tau \leq \\ \sup_{t\in\mathbb{R}} \|\boldsymbol{x}(t)\| \cdot \int_{\mathbb{R}} \|\boldsymbol{H}(\tau)\| d\tau \leq \sup_{t\in\mathbb{R}} \|\boldsymbol{x}(t)\| \cdot M \end{aligned} \tag{202}$$

- The response function can be written as :

$$\boldsymbol{H}(t) = \sum_{k=1}^{n} \boldsymbol{H}(t) . \boldsymbol{e}_k \otimes \boldsymbol{e}_k = \sum_{k=1}^{n} \boldsymbol{h}_k(t) \otimes \boldsymbol{e}_k \tag{203}$$

- A filter is causal when $H_{hk}(t) < 0$, $\forall t < 0$.

Example 3. Some very well known LTI filters are:

- Amplification and delay:

$$\mathcal{L}(\boldsymbol{x}(t)) = \alpha \cdot \boldsymbol{x}(t-\tau), \quad \Rightarrow \quad \boldsymbol{H}(t) = \alpha\delta(t-\tau) \cdot \mathbf{Id} \tag{204}$$

- Uniform moving average:

$$\mathcal{L}(\boldsymbol{x}(t)) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} \boldsymbol{x}(u)\, du = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} \boldsymbol{x}(t-u)\, du, \quad \Rightarrow \quad \boldsymbol{H}(t) = \frac{\chi_{\left[-\frac{T}{2};\frac{T}{2}\right]}}{T}\mathbf{Id} \tag{205}$$

with $\chi_{\left[-\frac{T}{2};\frac{T}{2}\right]}$ being the indicator function in the $\left[-\frac{T}{2};\frac{T}{2}\right]$ support.

The LTI has a harmonic spectral decomposition, that relies on the following eigenvalue problem (see also Equation (203)):

$$\mathcal{L}\left(e^{i\omega t}\boldsymbol{e}_k\right) = \boldsymbol{h}_k \star e^{i\omega t} = \hat{\boldsymbol{h}}_k(\omega) \cdot e^{i\omega t} \tag{206}$$

where $e^{i\omega t}$ and $\hat{\boldsymbol{h}}_k$ are the eigenvector and eigenvalue of L, respectively. The linear operator has a harmonic decomposition, with $\hat{\boldsymbol{h}}_k(\omega)$ being the Fourier transform of $\boldsymbol{h}_k(t)$. The Fourier transform of a signal $\boldsymbol{x}(t) \in L^1(\mathbb{R})$[18], is defined as:

$$\mathcal{F}(\boldsymbol{x}(t)) = \hat{\boldsymbol{x}}(\omega) = \int_{\mathbb{R}} \boldsymbol{x}(u) \cdot e^{-i\omega u} du \tag{207}$$

If $\boldsymbol{x} \in L^1(\mathbb{R})$, then $\|\hat{\boldsymbol{x}}(\omega)\| \leq \int_{\mathbb{R}} \|\boldsymbol{x}(u)\| du < +\infty$. Moreover, since $f(t) = e^{-i\omega t} \in C^{\infty}(\mathbb{R})$ is continuous, if $\boldsymbol{x}(t) \in L^1(\mathbb{R})$, $\hat{\boldsymbol{x}}(\omega)$ is continuous, since the Fourier transform is a linear operator, which is bounded since $\|\boldsymbol{x}(t) e^{-i\omega t}\| \leq \|\boldsymbol{x}(t)\|$, so it's continuous.

[18] In $\mathbb{R}^n$, $\boldsymbol{x}(t) \in L^1(\mathbb{R})$ iff

$$\int_{\mathbb{R}} \|\boldsymbol{x}(t)\| dt < +\infty$$

Example 4. Fourier transform of a Gaussian function.

A Gaussian function (such as the standard normal distributions) $p(u) = \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}$, with $p(u) \in C^{\infty}(\mathbb{R})$ decaying very fast at infinite, has a Fourier transform that reads, by definition:

$$\hat{p}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{u^2}{2}} e^{-i\omega u} du$$

$\hat{p}(0) = 1$ and $\hat{p}(\omega)$ can be differentiated, to obtain:

$$\frac{d\hat{p}}{d\omega}(\omega) = -\frac{i}{\sqrt{2\pi}} \int_{\mathbb{R}} u e^{-\frac{u^2}{2}} e^{-i\omega u} du$$

that integrated by part over $\mathbb{R}$ provides the following differential equation solved by $\hat{p}(\omega)$:

$$\frac{d\hat{p}}{d\omega}(\omega) = -\frac{\omega}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-u^2} e^{-i\omega u} du = -\omega \cdot \hat{p}(\omega) \tag{208}$$

The solution of the simple ODE is $\hat{p}(\omega) = A \cdot e^{-\frac{\omega^2}{2}}$, with $\hat{p}(0) = A = 1$. Therefore, the Fourier transform of a Gaussian function is a Gaussian function too. With a similar approach, the function $x(t) = e^{-(a-ib)t^2}$ has a Fourier transform that reads

$$\hat{x}(\omega) = \sqrt{\frac{\pi}{a-ib}} e^{-\frac{-(a+ib)\omega^2}{4(a^2+b^2)}}$$

Therefore, if the LTI filter is stable, each eigenvalue reads:

$$\hat{\boldsymbol{h}}_k(\omega) = \int_{\mathbb{R}} \boldsymbol{h}(u) \cdot e^{-i\omega u} du < +\infty \tag{209}$$

and it represents the $L^1(\mathbb{R})$-projection of $\boldsymbol{h}(t)$ on the harmonic of frequency ω. It follows that $\hat{\boldsymbol{H}}(\omega) = \sum_{k=1}^{n} \hat{\boldsymbol{h}}_k(\omega) \otimes \boldsymbol{e}_k$.
If $\boldsymbol{x} \in L^1(\mathbb{R})$, its inverse $\hat{\boldsymbol{x}} \in L^1(\mathbb{R})$ is $\in L^1(\mathbb{R})$ too[19] and it reads:

$$\boldsymbol{x}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{\boldsymbol{x}}(\omega) \cdot e^{i\omega t} d\omega \tag{210}$$

[19] Adopting the defintion of Fourier transform in Equation (207), the inverse Fourier transform can be written as $F^{-1}(\hat{\boldsymbol{x}}(\omega)) = \frac{1}{2\pi} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \boldsymbol{x}(u) e^{i\omega(t-u)} du \right) d\omega$. However, $\int_{\mathbb{R}^2} \boldsymbol{x}(u) e^{i\omega(t-u)} du d\omega$ is not finite, so the Fubini's theorem - required to prove that the inverse Fourier transform is finite - does not apply directly. It is first necessary to multiply the integrand by a Gaussian kernel $\hat{g}_{\varepsilon}(\omega) = e^{-\frac{\varepsilon^2\omega^2}{4}}$ - of the family $\mathcal{N}\left(0, \frac{2}{\varepsilon}\right)$, that converges to 1 for $\varepsilon \to 0$ - so to obtain the function: $\boldsymbol{\phi}_{\varepsilon}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \boldsymbol{x}(u) \hat{g}_{\varepsilon}(\omega) e^{i\omega(t-u)} du d\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{\boldsymbol{x}}(\omega) \hat{g}_{\varepsilon}(\omega) e^{i\omega t}$. In this case, $\| \hat{\boldsymbol{x}}(\omega) \hat{g}_{\varepsilon}(\omega) e^{i\omega t} \| \leq \| \hat{\boldsymbol{x}}(\omega) \|$, since the Gaussian kernel decays fast and $\hat{\boldsymbol{x}}(\omega) \hat{g}_{\varepsilon}(\omega) e^{i\omega t}$ converges to $\hat{\boldsymbol{x}}(\omega) e^{i\omega t}$ for $\varepsilon \to 0$. This implies that - by the theorem of dominated convergence - $\lim_{\varepsilon \to 0} \boldsymbol{\phi}_{\varepsilon}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{\boldsymbol{x}}(\omega) e^{i\omega t}$. Moreover, as shown

Equation (210) proves that the $\boldsymbol{x}(t)$ can be decomposed on a harmonic basis $\left\{e^{i\omega t}\right\}_{\omega\in\mathbb{R}}$. Again, since $f(t)=e^{i\omega t}\in C^{\infty}(\mathbb{R})$ is continuous, if $\hat{\boldsymbol{x}}(\omega)\in L^{1}(\mathbb{R})$, $\boldsymbol{x}(\omega)$ is continuous, since the inverse Fourier transform in Equation (210) is a linear operator, which is bounded since $\|\hat{\boldsymbol{x}}(t)\,e^{i\omega t}\|\leq\|\hat{\boldsymbol{x}}(t)\|$, so it's continuous. For discontinuous functions, the harmonic reconstruction in Equation (210) is not proved.

When both $\boldsymbol{x}(t),\boldsymbol{H}(t)\in L^{1}(\mathbb{R})$, the convolution in the Fourier domain becomes a contraction (or a product)[20]:

$$\boldsymbol{y}(t)=\boldsymbol{H}\star\boldsymbol{x}(t)\iff\mathcal{F}(\boldsymbol{y}(t))=\hat{\boldsymbol{y}}(\omega)=\hat{\boldsymbol{H}}(\omega).\hat{\boldsymbol{x}}(\omega)=\mathcal{F}(\boldsymbol{H}(t)).\mathcal{F}(\boldsymbol{x}(t)) \tag{211}$$

This implies that, the response to a LTI filter is simply written as:

$$\boldsymbol{y}(t)=\frac{1}{2\pi}\int_{\mathbb{R}}\hat{\boldsymbol{H}}(\omega).\hat{\boldsymbol{x}}(\omega)\,e^{i\omega t}d\omega \tag{212}$$

Example 5. Fourier transform of a discontinuous functions

- The Fourier transform of the rectangle function $\boldsymbol{x}(t)=\frac{1}{T}\chi_{\left[-\frac{T}{2};\frac{T}{2}\right]}\boldsymbol{e}_{i}$

in Example 4, $\hat{g}_{\varepsilon}(\omega)$ is the Fourier transform of the function

$$\frac{1}{2\pi}\int_{\mathbb{R}}\hat{g}_{\varepsilon}(\omega)\,e^{i\omega t}d\omega=\frac{1}{\varepsilon\sqrt{\pi}}e^{-\frac{t^{2}}{\varepsilon^{2}}}=\frac{\sqrt{2}}{\varepsilon}g_{1}\left(\frac{\sqrt{2}}{\varepsilon}t\right)$$

which is a Gaussian kernel that approximates the Dirac delta for $\varepsilon\to0$, keeping the integral over $\mathbb{R}$ equal to 1. This allows to rewrite $\boldsymbol{\phi}_{\varepsilon}(t)=\int_{\mathbb{R}}\boldsymbol{x}(u)\,\frac{\sqrt{2}}{\varepsilon}g_{1}\left(\frac{\sqrt{2}}{\varepsilon}(t-u)\right)du$. Therefore, the following expression concludes:

$$\lim_{\varepsilon\to0}\int_{\mathbb{R}}\|\boldsymbol{\phi}_{\varepsilon}(t)-\boldsymbol{x}(t)\|dt=\int_{\mathbb{R}}\lim_{\varepsilon\to0}\|\int_{\mathbb{R}}\frac{\boldsymbol{x}(u)}{\varepsilon\sqrt{\pi}}e^{-\frac{(t-u)^{2}}{\varepsilon^{2}}}du-\boldsymbol{x}(t)\|dt$$

$$\lim_{\varepsilon\to0}\int_{\mathbb{R}}\|\boldsymbol{\phi}_{\varepsilon}(t)-\boldsymbol{x}(t)\|dt=\int_{\mathbb{R}}\|\lim_{\varepsilon\to0}\boldsymbol{\phi}_{\varepsilon}(t)-\boldsymbol{x}(t)\|dt=\int_{\mathbb{R}}\|\int_{\mathbb{R}}\boldsymbol{x}(u)\,\delta(t-u)\,du-\boldsymbol{x}(t)\|dt=\mathbf{0}$$

$$\lim_{\varepsilon\to0}\boldsymbol{\phi}_{\varepsilon}(t)=\frac{1}{2\pi}\int_{\mathbb{R}}\hat{\boldsymbol{x}}(\omega)\,e^{i\omega t}d\omega=\boldsymbol{x}(t)$$

[20] As a matter of fact, the Fourier transform of $\boldsymbol{y}(t)$ can be expanded as:

$$\hat{\boldsymbol{y}}(\omega)=\int_{\mathbb{R}}\int_{\mathbb{R}}\boldsymbol{H}(u).\boldsymbol{x}(t-u)\,e^{-i\omega t}dudt$$

This expression can be integrated by means of the Fubini's theorem because $\|\boldsymbol{H}(t-u).\boldsymbol{x}(u)\|\in L^{1}(\mathbb{R}^{2})$ so

$$\hat{\boldsymbol{y}}(\omega)=\int_{\mathbb{R}^{2}}\boldsymbol{H}(v).\boldsymbol{x}(u)\,e^{i\omega(u+v)}dudv=\hat{\boldsymbol{H}}(\omega).\hat{\boldsymbol{x}}(\omega)$$

reads:

$$\hat{\boldsymbol{x}}(\omega) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} e^{-i\omega t} dt \boldsymbol{e}_i = \frac{\sin\left(\frac{T\omega}{2}\right)}{\frac{T\omega}{2}} \boldsymbol{e}_i = \text{sinc}\left(\frac{T\omega}{2}\right) \boldsymbol{e}_i \tag{213}$$

- The Fourier transform of the Heaviside function $\boldsymbol{x}(t) = H(t-\tau)\,\boldsymbol{e}_i$:

$$\hat{\boldsymbol{x}}(\omega) = \int_{\tau}^{+\infty} e^{-i\omega t} dt = \frac{e^{-i\omega\tau}}{i\omega} \tag{214}$$

- The Fourier transform of the function $\boldsymbol{x}(t) = t \cdot H(t-\tau)\,\boldsymbol{e}_i$:

$$\hat{\boldsymbol{x}}(\omega) = \int_{\tau}^{+\infty} te^{-i\omega t} dt \boldsymbol{e}_i = -e^{-i\omega\tau}\left(\frac{i\tau}{\omega} + \frac{1}{\omega^2}\right)\boldsymbol{e}_i \tag{215}$$

  The famous $ReLU$ function can be written as $ReLU(t) = t \cdot H(t)$, whose Fourier transform is $\widehat{ReLU}(t) = -\frac{1}{\omega^2}$

- The Fourier transform of the symmetric triangle function

$$\boldsymbol{x}(t) = \left(H(-t) \cdot ReLU\left(t + \frac{T}{2}\right) + H(t) \cdot ReLU\left(-t + \frac{T}{2}\right)\right)\boldsymbol{e}_i$$

  with $T \geq 0$ reads:

$$\hat{\boldsymbol{x}}(\omega) = \boldsymbol{e}_i\left(\int_{-\frac{T}{2}}^{0}\left(t + \frac{T}{2}\right)e^{-i\omega t}dt + \int_{0}^{\frac{T}{2}}\left(\frac{T}{2} - t\right)e^{-i\omega t}dt\right)$$

  By integrating by parts, one obtains:

$$\hat{\boldsymbol{x}}(\omega) = \frac{e^{-i\omega\frac{T}{2}} + e^{i\omega\frac{T}{2}} - 2}{(i\omega)^2} = T^2\frac{\left(e^{-i\omega\frac{T}{2}} + e^{i\omega\frac{T}{2}}\right)^2}{\left(2i\omega\frac{T}{2}\right)^2} = T^2\text{sinc}^2\left(\frac{T\omega}{2}\right) \tag{216}$$

  One can notice that the triangle function can be written as a convolution between two rectangular functions, as:

$$\boldsymbol{x}(t) = \chi_{\left[-\frac{T}{2},\frac{T}{2}\right]} \star \chi_{\left[-\frac{T}{2},\frac{T}{2}\right]}\boldsymbol{e}_i$$

  This expression eases the computation of the Fourier transform, because of the convolution properties.

- The Fourier transform of the Dirac delta $\delta_\tau(t) = \delta(t-\tau)$ reads:

$$\hat{\delta}_\tau(\omega) = \int_{\mathbb{R}} \delta(t-\tau)\,e^{-i\omega t}d\tau = e^{-i\omega\tau} \tag{217}$$

and the Fourier transform of the Dirac comb

$$c_T(t) = \sum_{n\in\mathbb{Z}} \delta(t - nT) \tag{218}$$

reads:

$$\hat{c}_T(\omega) = \int_{\mathbb{R}} \sum_{n\in\mathbb{Z}} \delta(t - nT)\, e^{-i\omega t} d\tau = \sum_{n\in\mathbb{Z}} e^{-i\omega nT} \tag{219}$$

$\hat{c}_T(\omega)$ is periodic of period $\frac{2\pi}{T}$[21]. Moreover, the following theorem proves that $\hat{c}_T(\omega)$ can be rewritten as a Dirac comb in frequency domain (see Theorem 37).

In order to extend Fourier theory to discontinuous functions so to leverage its nice features, the continuity condition can be slightly released, focusing on functions defined over the Hilbert space $L^2(\mathbb{R})$ (finite energy), a complete space of functions endowed with a scalar product defined by[22]:

$$\begin{aligned} &\boldsymbol{x}(t), \boldsymbol{y}(t) \in L^2(\mathbb{R}) \Rightarrow \langle \boldsymbol{x}(t), \boldsymbol{y}(t) \rangle_{L^2(\mathbb{R})} = \int_{\mathbb{R}} \langle \boldsymbol{x}(u), \boldsymbol{y}^*(u) \rangle\, du \\ &\|\boldsymbol{x}(t)\|_{L^2(\mathbb{R})} < +\infty; \quad \|\boldsymbol{y}(t)\|_{L^2(\mathbb{R})} < +\infty \end{aligned} \tag{220}$$

All functions $\boldsymbol{x}(t), \boldsymbol{y}(t) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ inherits the following fundamental property:

Theorem 36. ***Parseval's and Plancherel's theorem***
$\forall \boldsymbol{x}(t), \boldsymbol{y}(t) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \Rightarrow$ *Parseval formula:*

$$\langle \boldsymbol{x}(t), \boldsymbol{y}(t) \rangle_{L^2(\mathbb{R})} = \frac{1}{2\pi} \langle \hat{\boldsymbol{x}}(\omega), \hat{\boldsymbol{y}}(\omega) \rangle_{L^2(\mathbb{R})} \tag{221}$$

If $\boldsymbol{x} \equiv \boldsymbol{y} \Rightarrow$ *Plancherel formula:*

$$\|\boldsymbol{x}(t)\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi} \|\hat{\boldsymbol{x}}(\omega)\|^2_{L^2(\mathbb{R})} \tag{222}$$

Proof. Considering $\boldsymbol{z}(t) = \boldsymbol{x}(t) \star \boldsymbol{y}^*(-t)$, then $\boldsymbol{z}(t) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and $\hat{\boldsymbol{z}}(\omega) = \hat{\boldsymbol{x}}(\omega) . \hat{\boldsymbol{y}}^*(\omega)$[23] with

$$\langle \boldsymbol{x}(t), \boldsymbol{y}(t) \rangle_{L^2(\mathbb{R})} = \boldsymbol{z}(0) = \int_{\mathbb{R}} \hat{\boldsymbol{z}}(\omega)\, d\omega =$$

$$= \int_{\mathbb{R}} \langle \hat{\boldsymbol{x}}(\omega), \boldsymbol{y}^*(\omega) \rangle\, d\omega = \langle \hat{\boldsymbol{x}}(\omega), \hat{\boldsymbol{y}}(\omega) \rangle_{L^2(\mathbb{R})}$$

. □

[21] $e^{-i\omega nT} \cdot 1 = e^{-i\omega nT} \cdot e^{-i2\pi n} = e^{-i\left(\omega + \frac{2\pi}{T}\right)nT}$

[22] where $*$ stands for complex conjugate

[23] $\mathcal{F}(\boldsymbol{y}^*(-t)) = \int_{-\infty}^{+\infty} \boldsymbol{y}^*(-u)\, e^{-i\omega u} du = \left(-\int_{+\infty}^{-\infty} \boldsymbol{y}(-u)\, e^{(-i\omega(-u))} d(-u) \right)^* = \boldsymbol{y}^*(\omega)$

For discontinuous functions $\boldsymbol{x}(t) \in L^2(\mathbb{R})$ but $\notin L^1(\mathbb{R})$, the inverse Fourier transform is computed as a limit of a suite of functions $\{\boldsymbol{x}_n(t)\}_{n\in\mathbb{Z}} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$[24]. An example is represented by the Dirac's comb, according to the following theorem.

Theorem 37. ***Poisson formula***
Given two distributions $d_1(\omega)$ and $d_2(\omega)$ defined as:

$$d_1(\omega) = \sum_{n\in\mathbb{Z}} e^{-i\omega nT}, \quad d_2(\omega) = \frac{2\pi}{T}\sum_{n\in\mathbb{Z}} \delta\left(\omega - \frac{2\pi n}{T}\right)$$

d_1 and d_2 are equal in the sense of distributions (or in a weak formulation), i.e.:

$$\int_{\mathbb{R}} d_1(\omega)\cdot\hat{\phi}(\omega)\,d\omega = \int_{\mathbb{R}} d_2(\omega)\cdot\hat{\phi}(\omega)\,d\omega, \quad \forall\phi\in C_0^\infty(\mathbb{R})$$

Proof. $d_1(t)$ is periodic of period $\frac{2\pi}{T}$, which implies that we can prove the theorem over the compact support $\left[-\frac{\pi}{T},\frac{\pi}{T}\right]$, i.e.:

$$\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \sum_{n\in\mathbb{Z}} e^{-i\omega nT}\cdot\hat{\phi}(\omega)\,d\omega = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \frac{2\pi}{T}\delta(\omega)\cdot\hat{\phi}(\omega)\,d\omega = \frac{2\pi}{T}\hat{\phi}(0), \quad \forall\phi\in C_0^\infty(\mathbb{R}) \tag{223}$$

If one consider the truncated geometric series:

$$\sum_{n=-N}^{N} e^{-i\omega nT} = \sum_{n=0}^{N}\left(e^{-i\omega T}\right)^n + \sum_{n=0}^{N-1}\left(e^{i\omega T}\right)^{n+1}$$

with $|e^{i\omega T}| = |e^{-i\omega T}| \leq 1$ over $\left[-\frac{\pi}{T},\frac{\pi}{T}\right]$, the sum is equal to[25]:

$$\sum_{n=-N}^{N} e^{-i\omega nT} = \frac{e^{i\omega\frac{T}{2}} + e^{-i\omega\left(N+\frac{1}{2}\right)T}}{e^{i\omega\frac{T}{2}} - e^{-i\omega\frac{T}{2}}} + \frac{e^{i\omega\frac{T}{2}} + e^{i\omega\left(N+\frac{1}{2}\right)T}}{e^{-i\omega\frac{T}{2}} - e^{i\omega\frac{T}{2}}}$$

Recalling that $2i\sin(\alpha) = \left(e^{i\alpha} - e^{-i\alpha}\right)$ the truncated series results into:

$$\sum_{n=-N}^{N} e^{-i\omega nT} = \frac{\sin\left(T\omega\left(N+\frac{1}{2}\right)\right)}{\sin\left(\frac{T\omega}{2}\right)}$$

[24] As a matter of fact $L^1(\mathbb{R})\cap L^2(\mathbb{R})$ is dense in $L^2(\mathbb{R})$, which is a Hilbert space, so complete. The suite of functions converges $\lim_{n\to+\infty}\|\boldsymbol{x}_n - \boldsymbol{x}\|_{L^2(\mathbb{R})}$ and this suite of function is a Cauchy's sequence ($\forall\varepsilon>0, \exists M\in\mathbb{N}$ such that $\|\boldsymbol{x}_n - \boldsymbol{x}_m\|_{L^2(\mathbb{R})} < \varepsilon, \quad \forall n,m\geq M$). Since the $\boldsymbol{x}_n \in L^1(\mathbb{R})$, their inverse Fourier transform exists and it is noted $\hat{\boldsymbol{x}}_n(\omega)$. Therefore, for the Plancherel formula in Equation (222), it holds that $\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi}\|\hat{\boldsymbol{x}}_n - \hat{\boldsymbol{x}}_m\|^2_{L^2(\mathbb{R})}$ and therefore $\{\hat{\boldsymbol{x}}_n\}_{n\in\mathbb{Z}}$ is a Cauchy sequence too and (being $L^2(\mathbb{R})$ complete), it converges to $\hat{\boldsymbol{x}}(\omega) \in L^2(\mathbb{R})$ which is the Fourier transform of $\boldsymbol{x}(t)$. Plancherel's, Parseval's and properties of $L^1(\mathbb{R})$ functions apply to $L^2(\mathbb{R})$ too.

[25] $\sum_{k=0}^{N} q^k = \frac{1+q^{N+1}}{1-q}$

Therefore, Equation (223) can be rewritten as:

$$\begin{aligned}\lim_{N\to+\infty}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}}\sum_{n=-N}^{N}e^{-i\omega nT}\cdot\hat{\phi}(\omega)\,d\omega=\\=\frac{2\pi}{T}\lim_{N\to+\infty}\int_{\mathbb{R}}\hat{\psi}(\omega)\cdot\hat{\xi}_N(\omega)\,d\omega\end{aligned}\tag{224}$$

with

$$\hat{\psi}(\omega)=\frac{\chi_{\left[-\frac{\pi}{T},\frac{\pi}{T}\right]}(\omega)\cdot\hat{\phi}(\omega)}{\operatorname{sinc}\left(\frac{T\omega}{2}\right)}$$

and

$$\hat{\xi}_N(\omega)=\frac{\sin\left(T\omega\left(N+\frac{1}{2}\right)\right)}{\pi\omega}=\operatorname{sinc}\left(T\omega\left(N+\frac{1}{2}\right)\right)\frac{T\left(N+\frac{1}{2}\right)}{\pi}$$

Recalling that the inverse Fourier transform of the sinc function is the rectangle function (see Equation (213)), therefore

$$\frac{1}{2\pi}\int_{\mathbb{R}}\hat{\xi}_N(\omega)\,e^{i\omega t}d\omega=\frac{\chi_{\left[T\left(N+\frac{1}{2}\right),T\left(N+\frac{1}{2}\right)\right]}}{2\pi}$$

Thanks to the Parseval formula in Equation (221), Equation (224) becomes:

$$\begin{aligned}&\frac{2\pi}{T}\lim_{N\to+\infty}\int_{\mathbb{R}}\hat{\psi}(\omega)\cdot\hat{\xi}_N(\omega)\,d\omega=\frac{4\pi^2}{T}\lim_{N\to+\infty}\int_{\mathbb{R}}\psi(t)\cdot\xi_N(t)\,dt=\\=&\frac{2\pi}{T}\lim_{N\to+\infty}\int_{T\left(N+\frac{1}{2}\right)}^{T\left(N+\frac{1}{2}\right)}\psi(t)\,dt=\frac{2\pi}{T}\hat{\psi}(0)=\int_{\mathbb{R}}\frac{2\pi}{T}\delta(u)\,\hat{\psi}(u)\,du\end{aligned}\tag{225}$$

which proves the statement. □

The Poisson formula is rather useful for discrete signals. Discrete digital signals over N points are defined through convolution with a uniform moving average function $s_N\left(\frac{n}{N}\right)$ of the type Equation (205), as follows:

$$\boldsymbol{x}_N[n]=s_N\left(\frac{n}{N}\right)\mathbf{Id}\star\boldsymbol{x}(t)\tag{226}$$

In particular, recalling Equation (205), $s_N\left(\frac{n}{N}\right)$ samples $\boldsymbol{x}(t)$ with uniform interval $T=\frac{1}{N}$:

$$s_N(t)=\chi_{\left[0,\frac{1}{N}\right]}(t)\tag{227}$$

which implies that :

$$\boldsymbol{x}_N[n]=\int_0^{\frac{1}{N}}\boldsymbol{x}\left(\frac{n}{N}-u\right)du=\int_{\frac{n-1}{N}}^{\frac{n}{N}}\boldsymbol{x}(y)\,dy\tag{228}$$

If $\boldsymbol{x}(t) \in L^1(\mathbb{R})$, then, by Equation (226), the Fourier transform of a discrete signal reads:

$$\hat{\boldsymbol{x}}_N(\omega) = \hat{s}_N(\omega) \cdot \hat{\boldsymbol{x}}(\omega) = \frac{e^{\frac{i\omega}{2N}}}{N} \operatorname{sinc}\left(\frac{\omega}{2N}\right) \cdot \hat{\boldsymbol{x}}(\omega) \tag{229}$$

Equation (229) proves that:

$$\hat{s}(\omega) = \frac{e^{\frac{i\omega}{2N}}}{N} \operatorname{sinc}\left(\frac{\omega}{2N}\right) \tag{230}$$

Analogously, if one considers the a sampling function with vanishing high frequencies, such as $\hat{s}_N(\omega) = \frac{1}{N}\chi_{\left[-\frac{N}{2},\frac{N}{2}\right]}$, since $\hat{s}_N(\omega) \in L^1 \cap L^2$, according to Equation (213):

$$s_N(t) = \operatorname{sinc}\left(\frac{tN}{2}\right) \tag{231}$$

Equation (231) paves the way to the following fundamental theorem.

Theorem 38. ***Nyquist-Shannon sampling theorem***
Consider the space $\mathcal{S}_N$ of functions $L^2(\mathbb{R})$ whose Fourier coefficients vanish at high frequencies:

$$\mathcal{S}_N := \left\{ f(t) \in L^2(\mathbb{R}) \,|\, supp\left(\hat{f}(\omega)\right) \in \left[-\frac{N}{2}, \frac{N}{2}\right] \right\}$$

Then $\forall \boldsymbol{x} \in \mathcal{S}_N^n$:

$$\boldsymbol{x}(t) = \sum_{n \in \mathbb{Z}} \boldsymbol{x}\left(\frac{2\pi n}{N}\right) sinc\left(\frac{N}{2}\left(t - \frac{2\pi n}{N}\right)\right)$$

Proof. $s_N(t) = \operatorname{sinc}\left(\frac{tN}{2}\right) \in \mathcal{S}_N$ corresponds to a rectangle function $\frac{\chi_{\left[-\frac{N}{2},\frac{N}{2}\right]}(\omega)}{N}$ in the Fourier domain. Moreover, $\boldsymbol{x}_N(t - nT) = \boldsymbol{x} \star s_N(t - nT)\,\mathbf{Id}$, with a phase shift in the corresponding Fourier transform:

$$\mathcal{F}(s_N(t - nT)) = \hat{s}_N(\omega)\, e^{-i\omega nT} = \frac{\chi_{\left[-\frac{N}{2},\frac{N}{2}\right]}(\omega)}{N} e^{-i\omega nT}$$

Therefore, on the support $\left[-\frac{N}{2}, \frac{N}{2}\right]$, $\left\{\hat{s}_N(\omega)\, e^{-i\omega nT}\right\}_{n \in \mathbb{Z}}$ is a orthonormal basis. Therefore, the Fourier transform of the functions $f \in \mathcal{S}_N$ can be decomposed on this basis:

$$\hat{f}(\omega) = \hat{s}_N(\omega) \sum_{k \in \mathbb{Z}} \hat{f}_k \cdot e^{-i\omega kT}, \quad \hat{f}_k = \int_{-\frac{N}{2}}^{\frac{N}{2}} \hat{f}(\omega) \cdot e^{i\omega t} d\omega = f \star \operatorname{sinc}\left(\frac{kTN}{2}\right)$$

Thanks to the inverse Fourier formula in Equation (210),

$$f(t) = \sum_{k \in \mathbb{Z}} \hat{f}_k \cdot s_N(t - kT)$$

It follows that $\forall \boldsymbol{x} \in \mathcal{S}_N^n$:

$$\boldsymbol{x}(t) = \sum_{n \in \mathbb{Z}} \hat{\boldsymbol{x}}_n \cdot s_N(t - nT)$$

Since $s_N(mT - nT) = \delta_{m,n}$, $\hat{\boldsymbol{x}}_n = \boldsymbol{x}(nT)$. Finally, taking $T = \frac{2\pi}{N}$ proves the statement. □

Remark 39. As a matter of fact, the Theorem 38 proves that a function with a compact support in the frequency domain can be approximated with a infinite number of cardinal sinuses. On the contrary, we could be tempted to the same thing in the inverse order, i.e., to interpolate a signal with a compact support in time, i.e., discontinuous functions in time. This is the case of the finite element method or, more in general, the spline approximation problem. The problem resides in the fact that, in analogy with Theorem 38, one needs an infinite amount of cardinal sinuses in the frequency domain to reconstruct this signal.

The continuity of a function is intimately related to the decay of its Fourier's coefficients. The following theorem [Mal09] helps understanding this point:

Theorem 40. *A function $\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^n$ is bounded and belongs to $\in C^p(\mathbb{R})$ if:*

$$\int_{\mathbb{R}} \|\hat{\boldsymbol{x}}(\omega)\| (1 + |\omega|^p) \, d\omega < +\infty \tag{232}$$

Proof. If $\hat{\boldsymbol{x}} \in L^1(\mathbb{R})$, the $\frac{d^k \boldsymbol{x}}{dt^k}$ corresponds to $(i\omega)^p \hat{\boldsymbol{x}}(\omega)$, since:

$$\frac{d^p \boldsymbol{x}}{dt^p}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} (i\omega)^p \hat{\boldsymbol{x}}(\omega) e^{i\omega t} d\omega$$

with:

$$\|\boldsymbol{x}(t)\| \leq \frac{1}{2\pi} \int_{\mathbb{R}} \|\hat{\boldsymbol{x}}(\omega)\| |e^{i\omega t}| d\omega \leq \frac{1}{2\pi} \int_{\mathbb{R}} \|\hat{\boldsymbol{x}}(\omega)\| d\omega < +\infty$$

This inequality applies to all the derivatives of order $k \leq p$, under a condition:

$$\left\| \frac{d^p \boldsymbol{x}}{dt^p}(t) \right\| \leq \frac{1}{2\pi} \int_{\mathbb{R}} |\omega|^p \cdot \|\hat{\boldsymbol{x}}(\omega)\| d\omega \leq \frac{1}{2\pi} \int_{\mathbb{R}} (1 + |\omega|^p) \cdot \|\hat{\boldsymbol{x}}(\omega)\| d\omega$$

If $\frac{1}{2\pi} \int_{\mathbb{R}} (1 + |\omega|^p) \cdot \|\hat{\boldsymbol{x}}(\omega)\| d\omega$ is bounded by assumption, then all derivatives of order $k \leq p$ are bounded, which implies that $\boldsymbol{x} \in C^p(\mathbb{R})$ □

Theorem 40 proves that $\exists M > 0$ such that , $\forall \varepsilon > 0$

$$\|\hat{\boldsymbol{x}}(\omega)\| \leq \frac{M}{1 + |\omega|^{p+1+\varepsilon}} \tag{233}$$

then $\boldsymbol{x} \in C^p(\mathbb{R})$. If $\hat{\boldsymbol{x}}$ has a compact support, then $\boldsymbol{x} \in C^\infty(\mathbb{R})$. The regularity of $\boldsymbol{x}$ depends on how fast its Fourier coefficients decay: the highest the decay rate p, the highest the degree of regularity (smoothness) of the function.

Example 6. Regularity of the function $\boldsymbol{x}(t) = \frac{1}{T}\chi_{\left[-\frac{T}{2},\frac{T}{2}\right]}(t)\,\boldsymbol{e}_i$.
In this case, $\hat{\boldsymbol{x}}(\omega) = \operatorname{sinc}\left(\frac{T\omega}{2}\right)$ and $\|\hat{\boldsymbol{x}}(\omega)\| \leq |\omega|^{-1}$. Therefore, $\boldsymbol{x} \in C^0(\mathbb{R})$.

Theorem 40 can be revisited by introducing the Sobolev space $W^{k,p}$, defined in 1D as:

$$W^{k,p}(\mathbb{R}) := \{\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^n, D^\alpha \boldsymbol{x} \in L^p(\mathbb{R}) \,\forall 0 \leq |\alpha| \leq k\} \tag{234}$$

In Equation (234), the $D^\alpha \boldsymbol{x}$ represents the weak derivative of $\boldsymbol{x}(t)$ of order $|\alpha| = k$[26], defined, $\forall \boldsymbol{\phi}(t) \in C_C^\infty$ test functions with compact support, as:

$$\int_{\mathbb{R}} \langle \boldsymbol{x}^*(t), D^\alpha \boldsymbol{\phi}(t) \rangle \, dt = (-1)^{|\alpha|} \int_{\mathbb{R}} \langle \boldsymbol{\phi}^*(t), D^\alpha \boldsymbol{x}(t) \rangle \, dt \quad \forall \phi(t) \in C_C^\infty$$

The Sobolev space defined in Equation (234) is often equipped with the following norm, valid for all $1 \leq p < \infty$:

$$\|\boldsymbol{x}\|_{W^{k,p}(\mathbb{R})} = \left(\sum_{j=0}^{k} \|\boldsymbol{x}^{(j)}(t)\|_{L^p(\mathbb{R})}^p \right)^{\frac{1}{p}} = \left(\sum_{j=1}^{k} \int_{\mathbb{R}} \|\boldsymbol{x}^{(j)}(t)\|^p dt \right)^{\frac{1}{p}} \tag{235}$$

The extension of Equation (236) to $p = \infty$ reads:

$$\|\boldsymbol{x}\|_{W^{k,\infty}(\mathbb{R})} = \max_{j \in [\![0,k]\!]} \|\boldsymbol{x}^{(j)}(t)\|_\infty \tag{236}$$

Remark 41. $W^{k,p}$ equipped with the norm $\|\cdot\|_{W^{k,p}(\mathbb{R})}$ is a Banach space, therefore a complete normed vector space. It can be proved that the norm $\|\boldsymbol{x}(t)\|_{W^{k,p}(\mathbb{R})}$ is equivalent to $\|\boldsymbol{x}(t)\|_{L^p(\mathbb{R})} + \|\boldsymbol{x}^{(k)}(t)\|_{L^p(\mathbb{R})}$.

Provided the norm equivalence stated in Remark 41, Theorem 40 can be revisited as follows:

Theorem 42. *On the regularity of a function* $\boldsymbol{x}(t) \in W^{k,2}$ *is bounded and it belongs to* $C^k(\mathbb{R})$ *if its* $W^{k,2}$ *norm is bounded.*

Proof. If $\boldsymbol{x}(t) \in W^{k,2}$, then all its derivatives of to the k^{th} one belongs to $L^2(\mathbb{R})$, with a Fourier transform that exists and it is equivalent to $(i\omega)^k \hat{\boldsymbol{x}}(\omega)$. Recalling Plancherel formula in Equation (222), the $W^{k,2}$-norm of the function $\boldsymbol{x}(t)$ can be rewritten as:

[26] In general, for a function $\boldsymbol{x} : \mathbb{R}^m \to \mathbb{R}^m$

$$D^\alpha \boldsymbol{x} = \frac{\partial^{|\alpha|} \boldsymbol{x}}{\partial^{\alpha_1} x_1 \partial^{\alpha_m} x_m \ldots \partial^{\alpha_m} x_m}$$

$$\|\boldsymbol{x}\left(t\right)\|_{W^{k,2}(\mathbb{R})}=\left(\sum_{j=0}^{k}\|\boldsymbol{x}^{(j)}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}\right)^{\frac{1}{2}}=\frac{\|\hat{\boldsymbol{x}}\left(\omega\right)\|_{L^{2}(\mathbb{R})}}{\sqrt{2\pi}}\left(1+\sum_{\substack{j=1\\k>0}}^{k}|\omega|^{2j}\right)^{\frac{1}{2}}<+\infty \tag{237}$$

Considering the equivalence of the $W^{k,2}$-norm with the norm defined as $\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}+\|\boldsymbol{x}^{(k)}\left(t\right)\|_{L^{2}(\mathbb{R})}$. Therefore, $\boldsymbol{x}$ is bounded if:

$$\|\boldsymbol{x}\left(t\right)\|_{W^{k,2}(\mathbb{R})}=\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}+\|\boldsymbol{x}^{(k)}\left(t\right)\|_{L^{2}(\mathbb{R})}=\frac{\|\hat{\boldsymbol{x}}\left(\omega\right)\|_{L^{2}(\mathbb{R})}}{\sqrt{2\pi}}\left(1+|\omega|^{k}\right)<+\infty \tag{238}$$

which implies the condition defined by Theorem 40 and proves the statement. □

The Fourier transform is affected by the *Heisenberg uncertainty*, which implies that a function $\boldsymbol{x}\left(t\right)$ cannot have both a compact support in time and in the frequency domain, as stated by the following theorem.

Theorem 43. ***Heisenberg uncertainty*** *[Mal09]*
A function $\boldsymbol{x}\left(t\right)\in L^{2}\left(\mathbb{R}\right)$ with a temporal mean u, a temporal variance σ_{t}^{2}, a temporal mean ξ and a frequency variance σ_{ω}^{2} that read:

$$u=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}}\int_{\mathbb{R}}t\|\boldsymbol{x}\left(t\right)\|^{2}dt\qquad\xi=\frac{1}{2\pi\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}}\int_{\mathbb{R}}\omega\|\hat{\boldsymbol{x}}\left(\omega\right)\|^{2}d\omega$$

$$\sigma_{t}^{2}=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}}\int_{\mathbb{R}}\left(t-u\right)^{2}\|\boldsymbol{x}\left(t\right)\|^{2}dt$$

$$\sigma_{\omega}^{2}=\frac{1}{2\pi\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}}\int_{\mathbb{R}}\left(\omega-\xi\right)^{2}\|\hat{\boldsymbol{x}}\left(\omega\right)\|^{2}d\omega$$

then

$$\sigma_{t}^{2}\cdot\sigma_{\omega}^{2}\geq\frac{1}{4}$$

Moreover, if $\exists(u,\xi,A,b)\in\mathbb{R}^{2}\times\mathbb{C}^{2}$ such that $\boldsymbol{x}\left(t\right)=A\cdot e^{i\xi t-b(t-u)^{2}}\boldsymbol{e}_{i}$, then $\sigma_{t}^{2}\cdot\sigma_{\omega}^{2}=\frac{1}{4}$

Proof. u and ξ represent the time and frequency average values. Rearranging their definition and recalling the Plancherel's formula in Theorem 36, one obtains:

$$0=\frac{1}{2\pi\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}}\int_{\mathbb{R}}\left(\omega-\xi\right)\|\hat{\boldsymbol{x}}\left(\omega\right)\|^{2}d\omega=\frac{1}{2\pi\|\boldsymbol{x}\left(t\right)\|_{L^{2}(\mathbb{R})}^{2}}\int_{\mathbb{R}}\omega\|\hat{\boldsymbol{x}}\left(\omega+\xi\right)\|^{2}d\omega$$

The Fourier spectrum $\hat{\boldsymbol{x}}\left(\omega+\xi\right)$ corresponds to a function $\boldsymbol{y}\left(t\right)=\boldsymbol{x}\left(t\right)\cdot e^{-i\xi t}$, that has a nil frequency average and whose time average reads:

$$\frac{1}{2\pi\|\boldsymbol{y}\left(t\right)\|_{L^2(\mathbb{R})}^2}\int_{\mathbb{R}}t\|\boldsymbol{y}\left(t\right)\|^2dt=\frac{1}{2\pi\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^2}\cdot\int_{\mathbb{R}}t\|\boldsymbol{x}\left(t\right)\|^2dt=u$$

Therefore, the function $\boldsymbol{z}\left(t\right)=\boldsymbol{x}\left(t+u\right)\cdot e^{-i\xi t}$ has zero temporal and frequency average, which simplifies the proof that comes next, since without loss of generality, we can prove it for $u=\xi=0$.
The product of the time and frequency variance reads:

$$\sigma_t^2\cdot\sigma_\omega^2=\frac{1}{2\pi\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\int_{\mathbb{R}}t^2\|\boldsymbol{x}\left(t\right)\|^2dt\cdot\int_{\mathbb{R}}\omega^2\|\hat{\boldsymbol{x}}\left(\omega\right)\|^2d\omega$$

By applying the Plancherel's formula to $\boldsymbol{x}^{'}\left(t\right)$, whose Fourier transform is $i\omega\hat{\boldsymbol{x}}\left(\omega\right)$, the product of the two variance can be rewritten as:

$$\sigma_t^2\cdot\sigma_\omega^2=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\int_{\mathbb{R}}t^2\|\boldsymbol{x}\left(t\right)\|^2dt\int_{\mathbb{R}}\|\boldsymbol{x}^{'}\left(t\right)\|^2dt=$$

$$=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\|t\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}\cdot\|\boldsymbol{x}^{'}\left(t\right)\|_{L^2(\mathbb{R})}$$

Because of the Hölder inequality, the product of the two variances is bounded by below as follows:

$$\sigma_t^2\cdot\sigma_\omega^2\geq\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\left(\int_{\mathbb{R}}\left\langle t\boldsymbol{x}^*\left(t\right),\boldsymbol{x}^{'}\left(t\right)\right\rangle dt\right)^2$$

$$=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\left(\int_{\mathbb{R}}\frac{t}{2}\left(\left\langle\boldsymbol{x}^*\left(t\right),\boldsymbol{x}^{'}\left(t\right)\right\rangle+\left\langle\boldsymbol{x}^{*'}\left(t\right),\boldsymbol{x}\left(t\right)\right\rangle\right)dt\right)^2=$$

$$=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\left(\int_{\mathbb{R}}\frac{t}{2}\left(\|\hat{\boldsymbol{x}}\left(\omega\right)\|^2\right)^{'}dt\right)^2=$$

$$=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\left(\frac{t}{2}\left(\|\hat{\boldsymbol{x}}\left(\omega\right)\|^2\right)|_{\mathbb{R}}-\frac{1}{2}\int_{\mathbb{R}}\left(\|\hat{\boldsymbol{x}}\left(\omega\right)\|^2\right)dt\right)^2$$

According to [Wey50], if one assumes that $\lim_{t\to+\infty}\sqrt{t}\|\boldsymbol{x}\left(t\right)\|=0$, the first integral vanishes and the previous expression becomes:

$$\sigma_t^2\cdot\sigma_\omega^2\geq\frac{1}{4}$$

that proves the first statement.
The Heisenberg inequality becomes an equality if:

$$\sigma_t^2\cdot\sigma_\omega^2=\frac{1}{\|\boldsymbol{x}\left(t\right)\|_{L^2(\mathbb{R})}^4}\left(\int_{\mathbb{R}}\left\langle t\boldsymbol{x}^*\left(t\right),\boldsymbol{x}^{'}\left(t\right)\right\rangle dt\right)^2$$

An intuitive guess is the Gaussian function, i.e. $\exists(a,b) \in \mathbb{C}^2$ such that $\boldsymbol{x}(t) = Ae^{-bt^2}\boldsymbol{e}_i$ such that $\boldsymbol{x}^{'} = -2bt\boldsymbol{x}(t)$ that satisfies the equality constraint. For $u \neq \xi \neq 0$, the above mentioned translation apply. □

B Compendium of fundamental results in optimization [RECAP]

This section presents the basic fundamental results in optimization, that constitute the basis of Machine Learning algorithms. Those notes are a summary of the Optimization class notes of CentraleSupélec [CP19], Université Paris-Dauphine [Mul19], École Normale Supérieure [Pey20] and École Polytechnique [AE23]. Another fundamental reference is [CP11].

Definition 44. Domain of a function $f : K \to \mathbb{R}$, with $K \subset \mathbb{R}^n$

$$\text{dom}(f) := \{\boldsymbol{x} \in K | f(\boldsymbol{x}) < +\infty\} \tag{239}$$

Definition 45. Epigraph of a function epi(f)=$\{(\boldsymbol{x}, y) \in \text{dom}(f) \times \mathbb{R} | f(\boldsymbol{x}) \leq y\}$

Definition 46. A function $f : K \to \mathbb{R}$, with $K \subset \mathbb{R}^n$, is said to be proper iff $\text{dom}(f) \neq \emptyset$

Definition 47. A set $K \subset H$, with H being a Hilbert space, is bounded iff[27]

$$\exists r > 0, \boldsymbol{x}_0 \in H | K \subset B_r(\boldsymbol{x}_0)$$

Definition 48. A set $K \subset H$, with H being a Hilbert space, is closed iff

$$\forall (\boldsymbol{x}_n)_{n\in\mathbb{N}} \in K \qquad \lim_{n\to+\infty} x_n = x \in K$$

Definition 49. A set $K \subset H$, with H being a Hilbert space, is compact iff

$$\forall (\boldsymbol{x}_n)_{n\in\mathbb{N}} \in K \qquad \exists\, (\boldsymbol{x}_{n_k})_{k\in\mathbb{N}} \text{ such that } \exists \boldsymbol{x} \in K \quad \lim_{k\to+\infty} \boldsymbol{x}_{n_k} = \boldsymbol{x}$$

Proposition 50. *If $K \subset H$, with H being a Hilbert space, is compact $\Rightarrow$ closed and bounded. The opposite holds iff $dim H < +\infty$*

Definition 51. Convex set $K \subset \mathbb{R}^n$:

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2, \forall t \in [0,1] \qquad \boldsymbol{x}_t = t\boldsymbol{y} + (1-t)\boldsymbol{x} \in K \tag{240}$$

[27] $B_r\left(\boldsymbol{x}_0\right)$ represents a open ball of center $\boldsymbol{x}_0$ and radius r, i.e.

$$B_r(\boldsymbol{x}_0) := \{\boldsymbol{x} \in H \| \boldsymbol{x} - \boldsymbol{x}_0 \| < r\}$$

.

Definition 52. Convex functions $f : K \to \mathbb{R}$ on convex sets $K \subset \mathbb{R}^n$:

$$\forall(\boldsymbol{x},\boldsymbol{y}) \in K^2, \forall t \in [0,1] \qquad f(\boldsymbol{x}_t) = f(t\boldsymbol{y} + (1-t)\boldsymbol{x}) \leq tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{y}) \tag{241}$$

Definition 53. Strictly convex functions $f : K \to \mathbb{R}$ on convex sets $K \subset \mathbb{R}^n$:

$$\forall(\boldsymbol{x},\boldsymbol{y}) \in K^2, \forall t \in [0,1] \qquad f(\boldsymbol{x}_t) = f(t\boldsymbol{y} + (1-t)\boldsymbol{x}) < tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{y}) \tag{242}$$

Definition 54. Strongly convex functions $f : K \to \mathbb{R}$ on convex sets K:

$$\forall(\boldsymbol{x},\boldsymbol{y}) \in K^2, \forall t \in [0,1] \qquad f(\boldsymbol{x}_t) = f(t\boldsymbol{y}+(1-t)\boldsymbol{x}) \leq tf(\boldsymbol{y})+(1-t)f(\boldsymbol{y})-\alpha t(1-t)\|\boldsymbol{y}-\boldsymbol{x}\|^2, \qquad \alpha > 0 \tag{243}$$

Proposition 55. *f convex on convex set $K \subset \mathbb{R}^n \Leftarrow f$ strictly convex on convex set $K \subset \mathbb{R}^n \Leftarrow f$ strongly convex on convex set $K \subset \mathbb{R}^n$*

Proposition 56. *f convex on convex set $K \subset \mathbb{R}^n$, $f \in C^1(K) \Leftrightarrow$*

1. $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \qquad \forall(\boldsymbol{x},\boldsymbol{y}) \in K^2$

   Proof. $\Rightarrow$
   f convex on convex set $K \subset \mathbb{R}^n$, $f \in C^1(K)$ then by Definition 52:
   $\forall(\boldsymbol{x},\boldsymbol{y}) \in K^2, \forall t \in [0,1] \qquad f(\boldsymbol{x}_t) = f(t\boldsymbol{y}+(1-t)\boldsymbol{x}) \leq tf(\boldsymbol{y})+(1-t)f(\boldsymbol{x})$
   If $f \in C^1(K)$:

   $$\lim_{t\to 0^+} \frac{f(\boldsymbol{x}_t) - f(\boldsymbol{x})}{t} = \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) \tag{244}$$

   □

   Proof. $\Leftarrow$
   $f \in C^1(K)$ and $\langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$, then:

   $$\begin{aligned} f(\boldsymbol{x}) &\geq f(\boldsymbol{x}_t) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle \\ f(\boldsymbol{y}) &\geq f(\boldsymbol{x}_t) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_t), \boldsymbol{y} - \boldsymbol{x}_t \rangle \end{aligned} \tag{245}$$

   and therefore:

   $$(1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y}) \geq f(\boldsymbol{x}_t) \tag{246}$$

   □

2. $\left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{y}) - \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \geq 0 \qquad \forall(\boldsymbol{x},\boldsymbol{y}) \in K^2$

Proof. $\Leftrightarrow$
f convex on convex set $K \subset \mathbb{R}^n$, $f \in C^1(K)$ then:

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2 \\ f(\boldsymbol{x}) &\geq f(\boldsymbol{y}) + \left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \right\rangle \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2 \end{aligned} \tag{247}$$

By summing the two inequalities, the proposition is proven. □

From a geometric standpoint, Proposition 56 states that a convex function is always above its tangent plane in $\boldsymbol{x}$.

Proposition 57. *f strongly convex on convex set $K \subset \mathbb{R}^n$, $f \in C^1(K)$ (elliptic function)* $\Leftrightarrow$

1. $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2$

Proof. $\Rightarrow$
f strongly convex on convex set $K \subset \mathbb{R}^n$, $f \in C^1(K)$ then by Definition 54:
$\forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2, \forall t \in [0,1] \qquad f(\boldsymbol{x}_t) = f(t\boldsymbol{y} + (1-t)\boldsymbol{x}) \leq tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{x}) - \alpha t(1-t)\|\boldsymbol{y} - \boldsymbol{x}\|^2$
If $f \in C^1(K)$:

$$\lim_{t \to 0^+} \frac{f(\boldsymbol{x}_t) - f(\boldsymbol{x})}{t} + \alpha(1-t)\|\boldsymbol{y} - \boldsymbol{x}\|^2 = \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \alpha\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) \tag{248}$$

which proves the statement. □

Proof. $\Leftarrow$
$f \in C^1(K)$ and $\frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$, then:

$$\begin{aligned} f(\boldsymbol{x}) &\geq f(\boldsymbol{x}_t) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}_t\|^2 \\ f(\boldsymbol{y}) &\geq f(\boldsymbol{x}_t) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_t), \boldsymbol{y} - \boldsymbol{x}_t \rangle + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}_t\|^2 \end{aligned} \tag{249}$$

and therefore:

$$(1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y}) - \alpha t(1-t)\|\boldsymbol{y} - \boldsymbol{x}\|^2 \geq f(\boldsymbol{x}_t) \tag{250}$$

□

2. $\left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{y}) - \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \geq \alpha\|\boldsymbol{y} - \boldsymbol{x}\|^2 \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2$

Proof. $\Leftrightarrow$
f convex on convex set $K \subset \mathbb{R}^n$, $f \in C^1(K)$ then:

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2 \\ f(\boldsymbol{x}) &\geq f(\boldsymbol{y}) + \left\langle \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \right\rangle + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2 \end{aligned} \tag{251}$$

By summing the two inequalities, the proposition is proven. □

From a geometric standpoint, Item 2 states that a convex function is always above a quadratic function in $\boldsymbol{y}$, which is above the tangent plane in $\boldsymbol{x}$.
Elliptic functions have higher curvature (if they belong to $C^2(K)$) since in this case property 1 in Item 2 adds to the Taylor expansion at the second order[28], as follows:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{252}$$

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2}\langle \mathbf{H}_f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + o\left(\|\boldsymbol{y} - \boldsymbol{x}\|^2\right) \tag{253}$$

which implies that, at the second order:

$$\frac{1}{2}\langle \mathbf{H}_f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \qquad \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2 \tag{254}$$

Remark 58. Equation (254) proves that C^2 strongly convex functions have "higher" curvature then a quadratic polynomial with coefficient α (whose Hessian reads $\alpha\mathbf{I}$). Moreover, since $\mathbf{H}_f(\boldsymbol{x})$ is real and symmetric, Equation (254) implies that $\mathbf{H}_f(\boldsymbol{x})$ is positive definite.

Proposition 59. *The following properties holds for convex functions :*

- *A function $f : K \to \mathbb{R}$ is convex $\iff$ $epi(f)$ is convex*
- *A function $f : K \to \mathbb{R}$ is convex $\iff$ $dom(f)$ is convex*
- *All linear combinations $\sum_{i=1}^{N} a_i f_i(\boldsymbol{x})$ of convex functions $f_i : K \to \mathbb{R}$ defined on a convex set K and with positive coefficients $a_i > 0$ is convex (proof straightforward)*
- *For a set $(f_i)_{i \in I}$ of convex functions $f_i : K \to \mathbb{R}$, then $\sup_{i \in I} f_i$ is convex.*

[28] In this case, $\lim_{\boldsymbol{y} \to \boldsymbol{x}} \frac{o(\|\boldsymbol{y} - \boldsymbol{x}\|)}{\|\boldsymbol{y} - \boldsymbol{x}\|} = 0$

- *All linear combinations* $\sum_{i=1}^{N} a_i f_i(\boldsymbol{x})$ *of convex functions* $f_i : K \to \mathbb{R}$ *defined on a convex set* K *and with positive coefficients* $a_i = 1$ *is convex (proof straightforward)*
- *The composition* $h = g \circ f$ *of a monotonically increasing convex function* $g : \mathbb{R} \to \mathbb{R}$ *and a convex function* $f : K \to \mathbb{R}$ *on a convex set* K *is convex:*

  Proof. Since f is convex, $f(\boldsymbol{x}_t) \leq tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{x})$. Since g is monotonically increasing, $g(f(t(\boldsymbol{y}) + (1-t)(\boldsymbol{x}))) \leq g(tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{x}))$. Since g is also convex, $g(f(t(\boldsymbol{y}) + (1-t)(\boldsymbol{x}))) \leq g(tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{x})) \leq tg(f(\boldsymbol{y})) + (1-t)g(f(\boldsymbol{x}))$ □

Definition 60. Coercive functions f on unbounded domains $K \subset \mathbb{R}^n$:

$$\lim_{\|\boldsymbol{x}_n\|_K \to +\infty} f(\boldsymbol{x}_n) = +\infty \tag{255}$$

Which implies that $\forall M \in \mathbb{N}, \exists N \in \mathbb{N}$ such that $f(\boldsymbol{x}_n) \geq M$, $\forall \boldsymbol{x}_n$ such that $\|\boldsymbol{x}_n\|_K \geq N$

Proposition 61. *f strongly convex on convex unbounded set $K \subset \mathbb{R}^n$, $f \in C^1(K)$ (elliptic function) $\Rightarrow$ f is coercive*

Proof. f elliptic:

$$f(\boldsymbol{y}_n) \geq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y}_n - \boldsymbol{x} \rangle + \frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{x}\|^2 \geq f(\boldsymbol{x}) - \|\boldsymbol{\nabla}_x f(\boldsymbol{x})\| \cdot \|\boldsymbol{y}_n - \boldsymbol{x}\| + \frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{x}\|^2$$

If $\|\boldsymbol{x}_n\| = \|\boldsymbol{y}_n - \boldsymbol{x}\| \to +\infty$, then $f(\boldsymbol{x} + \boldsymbol{x}_n) \geq f(\boldsymbol{x}) - \|\boldsymbol{\nabla}_x f(\boldsymbol{x})\| \cdot \|\boldsymbol{x}_n\| + \frac{\alpha}{2}\|\boldsymbol{x}_n\|^2 \to +\infty$ □

Example 7. Some example of convex and coercive functions:

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ is convex on $\mathbb{R}$ ($\frac{df}{dx} = 2x \Rightarrow \left(\frac{df}{dx}(y) - \frac{df}{dx}(y)\right)(y - x) \geq 0$)
- $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ is strongly convex on $\mathbb{R}$ ($\boldsymbol{\nabla}_x f(\boldsymbol{x}) = 2\boldsymbol{x} \Rightarrow \langle \boldsymbol{\nabla}_x f(\boldsymbol{y}) - \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = 2\|\boldsymbol{y} - \boldsymbol{x}\|^2 \geq 0$)
- $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x}) = \frac{1}{2}\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{b}, \boldsymbol{x} \rangle + c$ with $\mathbb{A} \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n, c \in \mathbb{R}$ is convex on $\mathbb{R}^n$ iff $\mathbb{A}$ is semi-positive definite ($\boldsymbol{\nabla}_x f(\boldsymbol{x}) = \mathbb{A}\boldsymbol{x} + \boldsymbol{b} \Rightarrow \langle \boldsymbol{\nabla}_x f(\boldsymbol{y}) - \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = 2\|\mathbb{A}(\boldsymbol{y} - \boldsymbol{x})\|^2 \geq 0$)
- $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x}) = \frac{1}{2}\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{b}, \boldsymbol{x} \rangle + c$ with $\mathbb{A} \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n, c \in \mathbb{R}$ is strongly convex on $\mathbb{R}^n$ iff $\mathbb{A}$ is positive definite and its minimum eigenvalue $\lambda_{\min} > 0$ ($\boldsymbol{\nabla}_x f(\boldsymbol{x}) = \mathbb{A}\boldsymbol{x} + \boldsymbol{b} \Rightarrow \langle \boldsymbol{\nabla}_x f(\boldsymbol{y}) - \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = 2\|\mathbb{A}(\boldsymbol{y} - \boldsymbol{x})\|^2 = 2\sum_{i=1}^{n} \lambda_i |y_i - x_i|^2 \geq \lambda_{\min} \sum_{i=1}^{n} |y_i - x_i|^2 = \lambda_{\min} \|\boldsymbol{y} - \boldsymbol{x}\|^2 (\alpha = \lambda_{\min})$)

- $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x}) = \frac{1}{2}\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x}\rangle + \langle \boldsymbol{b}, \boldsymbol{x}\rangle + c$ with $\mathbb{A} \in \mathcal{M}_n(\mathbb{R}), \boldsymbol{b} \in \mathbb{R}^n, c \in \mathbb{R}$ is strongly convex on $\mathbb{R}^n$, with $\mathbb{A}$ is positive definite and its minimum eigenvalue $\lambda_{\min} > 0$, f est coercive ($\mathbb{R}^n$ is unbounded, f is $C^1(\mathbb{R}^n)$ so elliptic and therefore coercive).
- All function bounded below by a coercive function are coercive

Definition 62. Infimum of a set $K \in \mathbb{R}$

$$m = \inf_{x \in K} x, \qquad m \in]-\infty, +\infty[\tag{256}$$

Iff:

- $\forall x \in K, \qquad m \leq x$
- $\forall y \in \mathbb{R}$ such that $\forall x \in K \quad y \leq x \Longrightarrow m \geq y$
  or alternatively:
  $\forall y \in \mathbb{R}$ such that $m < y \Longrightarrow \exists x \in K$ such that $y > x$
  or alternatively:
  $\forall \varepsilon > 0 \Longrightarrow \exists x \in K$ such that $x - m < \varepsilon$
  or alternatively:
  $\exists (x_n)_{n\in\mathbb{N}} \in \mathbb{R}$ such that $\forall \varepsilon > 0 \Longrightarrow \exists N \in \mathbb{N}$ such that $\forall n \geq N \quad |x_n - x| < \varepsilon$
  or alternatively:
  $\exists (x_n)_{n\in\mathbb{N}} \in \mathbb{R}, \quad \lim_{n\to+\infty} x_n = m$

The same consideration hold for $m = \inf_{\boldsymbol{x}\in K} f(\boldsymbol{x})$, with $f : K \to \mathbb{R}$ and $K \subset \mathbb{R}^n$.

Definition 63. Infimum limit and supremum of a set $K \in \mathbb{R}$

$$\forall (x_n)_{n\in\mathbb{N}} \in \mathbb{R}, \quad \liminf x_n = \lim_{n\to+\infty} \inf_{k\geq n} (x_k) \tag{257}$$

$$\limsup x_n = -\liminf -x_n \tag{258}$$

Proposition 64. $\forall (x_n)_{n\in\mathbb{N}} \in \mathbb{R}, \quad \lim_{n\to+\infty} x_n = m \quad \Longleftrightarrow \quad \liminf x_n = \limsup x_n = m$ *and* $\liminf x_n \leq \limsup x_n$

Proof. By definition, $\liminf x_n = m$ iff:

$$\forall \varepsilon > 0, \exists M_i \in \mathbb{N} \text{ such that } \forall n \geq M_i \quad |\inf_{k\geq n} x_k - m| < \varepsilon$$

By definition of $\limsup x_n = m$ iff:

$$\forall \varepsilon > 0, \exists M_s \in \mathbb{N} \text{ such that } \forall n \geq M_s \quad |\sup_{k\geq n} x_k - m| < \varepsilon$$

Since $\liminf x_n \leq \limsup x_n$, the following inequality holds:

$$\forall \varepsilon > 0, \exists M \geq \max\{M_s, M_i\} \in \mathbb{N} \text{ such that } \forall n \geq M \quad m+\varepsilon \leq \inf_{k\geq n} x_k \leq (x_k)_{k\geq n} \sup_{k\geq n} x_k < \varepsilon - m$$

which proves the statement. □

Definition 65. Lower semi-continuous functions (lsc) [CP19]
$f : K \in \mathbb{R}$ with $K \in \mathbb{R}^n$ is lsc iff:

$$\exists (\boldsymbol{x}_n)_{n\in\mathbb{N}} \in K \text{ such that when } \lim_{n\to+\infty} \boldsymbol{x}_n = \boldsymbol{x} \Rightarrow \liminf f(\boldsymbol{x}_n) \geq f(\boldsymbol{x}) \quad (259)$$

Proposition 66. *$f : K \in \mathbb{R}$ with $K \in \mathbb{R}^n$ is lsc iff epi(f) is close. All continuous functions are lsc and so are the sum of lsc functions and $\sup_i f_i$ with f_i lsc [CP19].*

Definition 67. $\hat{\boldsymbol{x}}$ is local minimizer of a proper function $f : K \to \mathbb{R}$ (see Definition 46) with a non-empty $K \subset H$ and H a Hilbert space if $\exists$ open neighborhood $O(\boldsymbol{x}) \in H$ such that:

$$f(\hat{\boldsymbol{x}}) \leq f(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in O(\boldsymbol{x}) \cap K \quad (260)$$

$\hat{\boldsymbol{x}}$ is a strict local minimizer iff:

$$f(\hat{\boldsymbol{x}}) < f(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in (O(\boldsymbol{x}) \cap K)/\{\boldsymbol{x}\} \quad (261)$$

Definition 68. $\hat{\boldsymbol{x}}$ is global minimizer of a proper function $f : K \to \mathbb{R}$ (see Definition 46) with a non-empty $K \subset H$ and H a Hilbert space if:

$$f(\hat{\boldsymbol{x}}) \leq f(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in K \quad (262)$$

$\hat{\boldsymbol{x}}$ is a strict global minimizer iff:

$$f(\hat{\boldsymbol{x}}) < f(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in K/\{\boldsymbol{x}\} \quad (263)$$

Definition 69. The minimum of a function $f : K \to \mathbb{R}$ with $K \subset \mathbb{R}^n$ is the value $m \in \mathbb{R}$ - if it exists - for which $\exists \hat{\boldsymbol{x}} \in K$ such that:

$$\forall \boldsymbol{x} \in K \quad f(\boldsymbol{x}) \geq m \quad \text{and} \quad \min_{\boldsymbol{x}\in K} f(\boldsymbol{x}) = m = f(\hat{\boldsymbol{x}}) \quad (264)$$

The following theorems represent the theoretical framework within which the Machine Learning algorithms are defined. In particular, the following theorems assess the sufficient and necessary conditions to have local or global minimizers [CP19; Mul19].

Theorem 70. ***Bolzano-Weierstrass theorem (Rein analytischer Beweis, Bolzano, 1817 and Weierstrass, 1870)**: Sufficient condition of existence of a minimizer on a compact set*
Given a function $f : K \to \mathbb{R}$ defined over a non-empty compact set $K \subset H$, with H being a finite dimensional Hilbert space, f proper and lsc on $K \Longrightarrow \exists \hat{\boldsymbol{x}} \in K$ such that

$$f(\hat{\boldsymbol{x}}) = \inf_{\boldsymbol{x}\in K} f(\boldsymbol{x}) = m \in \mathbb{R}$$

Proof. If $\exists m = \inf_{\boldsymbol{x}\in K} f(\boldsymbol{x})$, $m \in \mathbb{R}$, by Definition 62 and by Proposition 64 $\exists (f(\boldsymbol{x}_n))_{n\in\mathbb{N}} \in H$ such that $m = \lim_{n\to+\infty} f(\boldsymbol{x}_n)$. Since K is compact $\exists (\boldsymbol{x}_{n_k})_{k\in\mathbb{N}}$ such that $\lim_{k\to+\infty} \boldsymbol{x}_{n_k} = \hat{\boldsymbol{x}} \in K$ (see Definition 49). Finally, since f is lsc (see Definition 65) $\liminf f(\boldsymbol{x}_{k\geq n}) = \liminf f(\boldsymbol{x}_{n_k}) \geq f(\hat{\boldsymbol{x}})$. But, by the property in Proposition 64,

$$\lim_{n\to\infty} f(\boldsymbol{x}_n) = m = \inf_{\boldsymbol{x}\in K} f(\boldsymbol{x}) \iff m = \liminf f(\boldsymbol{x}_n) \geq f(\hat{\boldsymbol{x}})$$

Therefore, by definition of $\inf_{\boldsymbol{x}\in K} f(\boldsymbol{x})$ as the largest minimizer of $f(\boldsymbol{x})$ over K, it follows that $f(\hat{\boldsymbol{x}}) = m = \inf_{\boldsymbol{x}\in K} f(\boldsymbol{x})$ which proves the theorem. □

When K is closed but not bounded, the sufficient condition of existence of a minimizer requires f to be coercive. The following theorem is thus fundamental to prove the existence of minimizers on unbounded sets.

Theorem 71. *Given a function $f : H \to \mathbb{R}$ defined over a non-empty finite dimensional Hilbert space H, f proper coercive and lsc on $H \implies C = \arg\min_{\boldsymbol{x}\in K} f(\boldsymbol{x})$ is a non-empty compact set.*

Proof. Since f is proper, $\exists \boldsymbol{x}_0 \in K | f(\boldsymbol{x}_0) < +\infty$. In this case, considering $B_r(\boldsymbol{x}_0) \in K$ the open ball in H of radius $r > 0$ and center $\boldsymbol{x}_0$, then being f coercive (see Definition 60) then $\forall \boldsymbol{x} \notin B_r(\boldsymbol{x}_0)$ for which $\|\boldsymbol{x} - \boldsymbol{x}_0\| > r$ it holds that $f(\boldsymbol{x}) > f(\boldsymbol{x}_0)$. On the contrary, $\bar{B}_r(\boldsymbol{x}_0) \cup K$ is a compact, on which the Weierstrass theorem applies, i.e., it exists a minimizer $\hat{\boldsymbol{x}} = \arg\inf_{\boldsymbol{x}\in K} f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)$. Since f is coercive, then $\hat{\boldsymbol{x}}$ is a minimizer over H:

$$f(\hat{\boldsymbol{x}}) = \inf_{\boldsymbol{x}\in H} f(\boldsymbol{x})$$

This result implies that $\arg\min_{\boldsymbol{x}\in H} f \subset K$ is bounded. Moreover, any sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ converging to $\hat{\boldsymbol{x}} \in H$ satisfies the following property:

$$f(\hat{\boldsymbol{x}}) \leq \liminf f(\boldsymbol{x}_n) = \inf_{\boldsymbol{x}\in H} f(\boldsymbol{x})$$

which implies that $\arg\min_{\boldsymbol{x}\in H} f(\boldsymbol{x})$ is also closed and then (since H is finite dimensional) $\arg\min_{\boldsymbol{x}\in H} f(\boldsymbol{x})$ is compact. □

For any other open set $K \subset H$ is an open set, Theorem 71 does not apply. Therefore, does a minimizer of $f : K \to \mathbb{R}$ exist? The answer is positive, but under certain conditions.

Theorem 72. *Sufficient condition of existence of a minimizer on an open set Given a function $f : K \to \mathbb{R}$ defined over a non-empty bounded open set $K \subset H$, with H being a finite dimensional Hilbert space, f proper and lsc on $\bar{K}$ and $\exists \boldsymbol{x}_0 \in K$ such that:*

$$\forall \boldsymbol{x} \in \partial K, f(\boldsymbol{x}) > f(\boldsymbol{x}_0)$$

$\Longrightarrow \exists \hat{\boldsymbol{x}}$ *such that*

$$f(\hat{\boldsymbol{x}}) = \inf_{\boldsymbol{x} \in K} f(\boldsymbol{x}) = m \in \mathbb{R}$$

Proof. $\bar{K}$ is closed and bounded, therefore (since H is finite dimensional) $\bar{K}$ is compact and the Weierstrass Theorem 70 applies:

$$\exists \hat{\boldsymbol{x}} \in \bar{K} \text{ such that } f(\hat{\boldsymbol{x}}) = \inf_{\boldsymbol{x} \in \bar{K}} f(\boldsymbol{x}) = m \qquad m \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \bar{K}$$

$\hat{\boldsymbol{x}}$ belongs to K=int(K) (open set). The proof is obtained by reducing to absurd, i.e. by assuming that $\hat{\boldsymbol{x}} \in \partial K$. If it were so, provided the assumption that $\exists \boldsymbol{x}_0 \in K$ such that $\forall \boldsymbol{x} \in \partial K, f(\boldsymbol{x}) > f(\boldsymbol{x}_0)$, then

$$f(\hat{\boldsymbol{x}}) > f(\boldsymbol{x}_0) \quad \text{and} \quad f(\hat{\boldsymbol{x}}) = m \leq f(\boldsymbol{x}_0), \boldsymbol{x}_0 \in \bar{K}$$

The conditions are incompatible, which proves that $\hat{\boldsymbol{x}} \in K$ and $f(\boldsymbol{x}) \geq f(\hat{\boldsymbol{x}})$ $\forall \boldsymbol{x} \in \bar{K}$, included $\boldsymbol{x}_0$. □

The previous theorems assure the existence of a minimizer, but not the necessary condition required to search for one, nor the uniqueness of the minimizer. For such conditions, the convexity properties should be considered. The following results introduce such conditions.

Theorem 73. ***Euler's inequality****: necessary condition for a minimizer on a convex set*
Given a function $f : K \to \mathbb{R}$ defined over a non-empty convex set $K \subset H$, with H being a Hilbert space, f proper on K and $f \in C^1(K)$, if $\hat{\boldsymbol{x}}$ is a local minimizer of f on $K \Longleftarrow$

$$\langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{y} - \hat{\boldsymbol{x}} \rangle \geq 0, \qquad \forall \boldsymbol{y} \in K$$

Proof. Provided that K is convex $\forall t \in [0,1]$ $\boldsymbol{x}_t = t\boldsymbol{y} + (1-t)\hat{\boldsymbol{x}} \in K$, $\forall \boldsymbol{y} \in K$. Moreover, $f \in C^1(K)$ which implies that the following Taylor expansion on f holds:

$$f(\boldsymbol{x}_t) = T_f \hat{\boldsymbol{x}}(\boldsymbol{x}_t) + o(\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}\|) = f(\hat{\boldsymbol{x}}) + \langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{x}_t - \hat{\boldsymbol{x}} \rangle + o(\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}\|) \geq f(\hat{\boldsymbol{x}})$$

such that $\lim_{t \to 0^+} \frac{o(\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}\|)}{\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}\|} = 0$. Taking the following limit for $t \to 0$, the statement is proven:

$$\langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{y} - \hat{\boldsymbol{x}} \rangle =$$

$$= \lim_{t \to 0} \frac{\langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{x}_t - \hat{\boldsymbol{x}} \rangle + o(\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}\|)}{t} = \lim_{t \to 0} \frac{f(\boldsymbol{x}_t) - f(\hat{\boldsymbol{x}})}{t} \geq 0 \qquad \forall \boldsymbol{y} \in K$$

□

If K is open, the inequality in Theorem 73 becomes an equality and a sufficient and necessary condition for $\hat{\boldsymbol{x}}$ to be a minimizer. However, one must first acknowledge this preliminary result:

Theorem 74. ***Euler's equality****: necessary condition for a minimizer of a function on a convex open set*
Given a function $f : K \to \mathbb{R}$ defined over a non-empty convex set $K \subset H$, with H being a finite dimension Hilbert space, f proper on K and $f \in C^1(K)$, if $\hat{\boldsymbol{x}}$ is a local minimizer of f on $K \Longleftarrow$

$$\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}) = \mathbf{0}$$

Proof. $\Rightarrow$
Since $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \in K} f(\boldsymbol{x})$, $\exists O(\hat{\boldsymbol{x}})$ open neighborhood that, according to Equation (260) in Definition 67:

$$f(\hat{\boldsymbol{x}}) \leq f(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in O(\boldsymbol{x}) \cap K$$

Therefore, if one considers $\boldsymbol{x} = \hat{\boldsymbol{x}} + h\boldsymbol{y}$, $h \in]0, r_0]$[29] such that $\boldsymbol{x} \in O(\boldsymbol{x}) \cap K, \forall h \in]0, r_0]$ and $\forall \boldsymbol{y} \in K$, then :

$$\lim_{h \to 0^+} \frac{f(\hat{\boldsymbol{x}} + h\boldsymbol{y}) - f(\hat{\boldsymbol{x}})}{h} = \langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{y} \rangle \geq 0 \quad \forall \boldsymbol{y} \in K$$

But $\langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), -\boldsymbol{y} \rangle \geq 0$ too, which implies the only solution: $\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}) \equiv \mathbf{0}$. □

Proof. $\Leftarrow$
If $\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}) \equiv \mathbf{0}$, the statement is proven. □

If K and f are both convex, the inequality in Theorem 73 becomes a sufficient and necessary condition for $\hat{\boldsymbol{x}}$ to be a minimizer.

Theorem 75. ***Euler's inequality for convex functions****: sufficient and necessary condition for the existence of a global minimizer of a convex function on a convex set*
Given a function $f : K \to \mathbb{R}$ defined over a non-empty convex set $K \subset \mathbb{R}^n$, with $\mathbb{R}^n$ being a Hilbert space of finite dimension n, f proper and convex on K and $f \in C^1(K)$

$$\hat{\boldsymbol{x}} \text{ is a global minimizer of } f \text{ on } K \iff \langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{y} - \hat{\boldsymbol{x}} \rangle \geq 0, \quad \forall \boldsymbol{y} \in K$$

Proof. $\Rightarrow$
See Theorem 73 □

Proof. $\Leftarrow$
Since f is convex, Proposition 55 holds:

$$f(\boldsymbol{y}) \geq f(\hat{\boldsymbol{x}}) + \langle \boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}), \boldsymbol{y} - \hat{\boldsymbol{x}} \rangle$$

[29] When $K \subset \mathbb{R}^n$, the open neighborhood is the open ball $B_r(\hat{\boldsymbol{x}})$.

But, since $\hat{\boldsymbol{x}}$ is a minimizer, the following holds too:

$$f(\hat{\boldsymbol{x}})+2\langle\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}),\boldsymbol{y}-\hat{\boldsymbol{x}}\rangle \geq f(\boldsymbol{y})+\langle\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}),\boldsymbol{y}-\hat{\boldsymbol{x}}\rangle \geq f(\hat{\boldsymbol{x}})+\langle\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}),\boldsymbol{y}-\hat{\boldsymbol{x}}\rangle$$

which proves the statement. □

Proposition 76. *Provided that Theorem 75 holds for convex functions on convex subsets of $\mathbb{R}^n$, if K is open, Theorem 74 holds too, with the Euler's inequality that*

$$\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}) = \mathbf{0}$$

.

If f is strictly convex, the minimizer is unique.

Theorem 77. *Sufficient and necessary condition for the existence of unique minimizer of a strictly convex function on a convex set*
Given a function $f : K \to \mathbb{R}$ defined over a non-empty convex set $K \subset H$, with H being a Hilbert space, f proper and strictly convex on $K \implies$

$$\exists!\hat{\boldsymbol{x}} \text{ such that } \hat{\boldsymbol{x}} = \inf_{\boldsymbol{x}\in K} f(\boldsymbol{x})$$

Proof. The proof is obtained by reducing to absurd and assuming that there are two different minimizers $\hat{\boldsymbol{x}}_1 \neq \hat{\boldsymbol{x}}_2$. In this case, assuming $t = \frac{1}{2}$, $\frac{\hat{\boldsymbol{x}}_1+\hat{\boldsymbol{x}}_2}{2} \in K$ which is convex and therefore:

$$f(\hat{\boldsymbol{x}}_1) < f\left(\frac{\hat{\boldsymbol{x}}_1+\hat{\boldsymbol{x}}_2}{2}\right) < \frac{1}{2}f(\hat{\boldsymbol{x}}_1)+\frac{1}{2}f(\hat{\boldsymbol{x}}_2) = f(\hat{\boldsymbol{x}}_1)$$

which is not possible. □

Finally, if f is strongly convex (see Definition 54), the minimum exists and it is unique and it satisfies the Euler's inequality.

Theorem 78. ***Euler's inequality for strongly convex functions**: Sufficient and necessary condition for the uniqueness of a minimizer (if it exists) of a strongly convex function on a closed convex set*
Given a function $f : K \to \mathbb{R}$ defined over a non-empty closed and convex set $K \subset H$, with H being a Hilbert space, f proper and strongly convex on K and $f \in C^1(K)$

$$\hat{\boldsymbol{x}} \text{ is a unique minimizer of } f \text{ on } K \iff \langle\boldsymbol{\nabla}_x f(\hat{\boldsymbol{x}}),\boldsymbol{y}-\hat{\boldsymbol{x}}\rangle \geq 0, \quad \forall \boldsymbol{y} \in K$$

Proof. Since f proper and $f \in C^1$, it is lsc on K. Moreover, f is strongly convex, therefore elliptic, then it is coercive (see Proposition 61). Being coercive, Theorem 71 holds and therefore the minimizer exists. Moreover, the Theorem 71 states that $\arg\min_{\boldsymbol{x}\in K} f$ is compact. Again, since f is strongly convex,

it is also strictly convex (see Proposition 55), which implies - by Theorem 77 - that the existing minimizers is unique. Finally, f being strongly convex implies f being convex - again by Theorem 77 - which implies that Theorem 75 holds and therefore the Euler's inequality is proven. □

Remark 79. In the following we will consider $K = \mathbb{R}^n$, which is convex. Therefore, the Euler's inequality in Theorem 73 holds and the vectors $\boldsymbol{y} - \hat{\boldsymbol{x}}$ span all the affine space [Mul19].

The strong convexity of the function f has some repercussions on its smoothness, defined by its β-Lipschitz continuity [Pey20]:

Definition 80. A function $f : K \to \mathbb{R}$ is said to be β-Lipschitz iff

$$\exists \beta \in \mathbb{R}^+ \text{ such that } \forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2 \Rightarrow |f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq \beta \|\boldsymbol{y} - \boldsymbol{x}\| \tag{265}$$

The function is Lipschitz, i.e. $f \in \text{Lip}(K)$ for $\beta = \sup_{\boldsymbol{y} \neq \boldsymbol{x}} \frac{|f(\boldsymbol{y}) - f(\boldsymbol{x})|}{\|\boldsymbol{y} - \boldsymbol{x}\|}$ (Lipschitz constant).

Lipschitz strongly convex functions are sufficiently smooth since they are bounded by below and by above by a quadratic function, as shown in the following.

Proposition 81. *A strongly convex function $f : K \to \mathbb{R}$, defined over a non-empty convex set $K \subset H$, with H being a Hilbert space, f proper and strongly convex on K, with $f \in C^2(K)$ and a gradient $\boldsymbol{\nabla}_x f \in Lip(K)$ then $\forall (\boldsymbol{x}, \boldsymbol{y}) \in K^2$:*

$$\frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{266}$$

or alternatively, since the Hessian operator $\mathbf{H}_f(\boldsymbol{x})$ is symmetric [Pey20]:

$$\alpha \mathbf{I} \preceq \mathbf{H}_f(\boldsymbol{x}) \preceq \beta \mathbf{I}, \qquad \forall \boldsymbol{x} \in K \tag{267}$$

with $\preceq$ indicating the natural order of symmetric matrices, i.e.:

$$\forall (\mathbb{A}, \mathbb{B}) \in Sym(\mathbb{R}^n)^2 \Rightarrow \mathbb{A} \preceq \mathbb{B} \iff \langle \mathbb{A}.\boldsymbol{u}, \boldsymbol{u} \rangle \leq \langle \mathbb{B}.\boldsymbol{u}, \boldsymbol{u} \rangle \, \forall \boldsymbol{u} \in \mathbb{R}^n$$

Proof. The lower bound has been proved in Item 2. For the upper bound, since $f \in C^2(K)$ and with $\boldsymbol{x}_t = t\boldsymbol{y} + (1-t)\boldsymbol{x}$, $t \in [0,1]$:

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_0^1 D_{t(\boldsymbol{y}-\boldsymbol{x})} f(\boldsymbol{x})\, dt = \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle +$$

$$+ \int_0^1 \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}_t) - \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \, dt$$

Applying the Hölder inequality $\langle \boldsymbol{a}, \boldsymbol{b} \rangle \leq \|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|$ to the previous expression, coupled with the Lipschitz property of f, one obtains:

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \int_0^1 \|\boldsymbol{\nabla}_x f(\boldsymbol{x}_t) - \boldsymbol{\nabla}_x f(\boldsymbol{x})\| dt \cdot \|\boldsymbol{y} - \boldsymbol{x}\|$$

Applying the Lipschitz property of the gradient, the following expression proves the statement:

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\nabla}_x f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \beta \int_0^1 t \cdot dt \cdot \|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq \frac{\beta}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2$$

□

Remark 82. If Proposition 81 holds, equivalently Equation (267) does too. In this case, Equation (267) can be rephrased as:

$$\alpha \|\boldsymbol{u}\|^2 \leq \langle \mathbf{H}_f(\boldsymbol{x}) . \boldsymbol{u}, \boldsymbol{u} \rangle \leq \beta \|\boldsymbol{u}\|^2, \qquad \forall \boldsymbol{x} \in K, \forall \boldsymbol{u} \in H \tag{268}$$

Plus, as observed in Remark 58, $\mathbf{H}_f(\boldsymbol{x})$ is positive definite, i.e., all its eigenvalues are positive, thus, given the spectral representation of the symmetric real matrix $\mathbf{H}_f(\boldsymbol{x})$, Equation (268) reduces to:

$$\begin{cases} \min \lambda_i (\mathbf{H}_f(\boldsymbol{x})) \geq \alpha & (269) \\ \max \lambda_i (\mathbf{H}_f(\boldsymbol{x})) \leq \beta & (270) \end{cases}$$

Equation (270) implies that the Hessian matrix must be well conditioned. In other words, Equation (270) can be replaced by invoking the conditioning number of the Hessian matrix $\kappa(\mathbf{H}_f(\boldsymbol{x})) = \frac{\max \lambda_i(\mathbf{H}_f(\boldsymbol{x}))}{\min \lambda_i(\mathbf{H}_f(\boldsymbol{x}))}$ that must respect the condition $0 < \kappa(\mathbf{H}_f(\boldsymbol{x})) \leq \frac{\beta}{\alpha}$. This aspect is crucial for convergence analysis of gradient descent methods (see Section 4.3 for further details). In other words, the strongest is the convexity of the function, the more flexible is the conditioning of the Hessian matrix. On the contrary, a lower Lipschitz constant β on the gradient represents a more stringent conditioning of the Hessian matrix.

References

[AE23] **Allaire, Grégoire and Alexandre Ern**. *Optimisation et Contrôle*. Cours de l'Ecole Polytechnique. 2023. URL: http://www.cmap.polytechnique.fr/~allaire/map435/poly435.pdf.

[AW19] **Azulay, Aharon and Yair Weiss**. "Why do deep convolutional networks generalize so poorly to small image transformations?" In: *Journal of Machine Learning Research* 20 (2019), pp. 1–25.

[Bar93] **Barron, Andrew R**. "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information theory* 39.3 (1993), pp. 930–945.

[Bil95] **Billingsley, Patrick**. *Measure and Probability*. John Wiley and Sons: New York, 1995.

[Cam18] **Campagne, J.E.** *L'apprentissage face à la malédiction de la grande dimensionn*. Notes et commentaires au sujet des conférences de S. Mallat du Collè ge de France (2018). 2018. DOI: https://doi.org/10.4000/annuaire-cdf.15441.

[Cam19] **Campagne, J.E.** *L'apprentissage par réseaux de neurones profonds*. Notes et commentaires au sujet des conférences de S. Mallat du Collè ge de France (2019). 2019. DOI: https://doi.org/10.4000/annuaire-cdf.16767. URL: https://www.di.ens.fr/~mallat/College/Cours-2019-Mallat-Jean-Eric-Campage.pdf.

[Cam20] **Campagne, J.E.** *Modè les multi-échelles et réseaux de neurones convolutifs*. Notes et commentaires au sujet des conférences de S. Mallat du Collè ge de France (2020). 2020. URL: https://www.di.ens.fr/~mallat/College/Cours2020-Mallat-Jean-Eric-Campagne.pdf.

[Cam22] **Campagne, J.E.** *Information et Complexite*. Notes et commentaires au sujet des conférences de S. Mallat du Collè ge de France (2022). 2022. DOI: https://doi.org/10.4000/annuaire-cdf.18077. URL: https://www.di.ens.fr/~mallat/College/Cours-2022-Mallat-Jean-Eric-Campagne.pdf.

[Che+16] **Chen, Xi et al.** "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". en. In: arXiv:1606.03657 (June 2016). arXiv:1606.03657 [cs, stat]. URL: http://arxiv.org/abs/1606.03657.

[Clo22] **Clouteau, Didier**. *Statistical Learning for Mechanics - Lecture notes*. Cours de Modélisation des Incertitudes et Fiabilité des Ouvrages-Master 2 - Université Paris-Saclay. 2022.

[Cou20] **Cournède, Paul-Henry**. *Official handbook of the Statistics and Learning*. "Ingénieur" Curriculum, CentraleSupélec. 2020.

[CP11] **Combettes, Patrick L. and Jean-Christophe Pesquet**. "Proximal Splitting Methods in Signal Processing". en. In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Ed. by Heinz H. Bauschke et al. Vol. 49. Springer Optimization and Its Applications. New York, NY: Springer New York, 2011, pp. 185–212. ISBN: 978-1-4419-9568-1. DOI: 10.1007/978-1-4419-9569-8_10. URL: https://link.springer.com/10.1007/978-1-4419-9569-8_10.

[CP19] **Chouzenoux, Emilie and Jean-Christophe Pesquet**. *Official handbook of the Optimization*. "Ingénieur" Curriculum, Centrale-

Supélec. Class notes. 2019. URL: https://www-syscom.univ-mlv.fr/~chouzeno/ECP/index.htm.

[CUH16] **Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter**. "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". en. In: arXiv:1511.07289 (Feb. 2016). arXiv:1511.07289 [cs]. URL: http://arxiv.org/abs/1511.07289.

[Cyb89] **Cybenko, G.** "Approximation by superpositions of a sigmoidal function". en. In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314. ISSN: 0932-4194, 1435-568X. DOI: 10.1007/BF02551274. URL: http://link.springer.com/10.1007/BF02551274.

[Fis22] **Fisher, R. A.** "On the mathematical foundations of theoretical statistics". en. In: 222 (Jan. 1922). URL: https://royalsocietypublishing.org/doi/10.1098/rsta.1922.0009.

[For+15] **Forest, Samuel et al.** *Mécanique des milieux continus*. École des Mines de Paris, 2015.

[GM98] **Grenander, Ulf and Michael I. Miller**. "Computational anatomy: an emerging discipline". en. In: *Quarterly of Applied Mathematics* 56.4 (1998), pp. 617–694. ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/1668732. URL: https://www.ams.org/qam/1998-56-04/S0033-569X-1998-1668732-7/.

[Goo+14] **Goodfellow, Ian J. et al.** "Generative Adversarial Networks". In: arXiv:1406.2661 (June 2014). arXiv:1406.2661 [cs, stat]. URL: http://arxiv.org/abs/1406.2661.

[HDD] **Hecht-Nielsen, Robert, Oberlin Drive, and San Diego**. "Kolmogorov's Mapping Neural Network Existence Theorem". en. In: ().

[Hec92] **Hecht-Nielsen, Robert**. "Theory of the Backpropagation Neural Network". en. In: *Neural Networks for Perception*. Elsevier, 1992, pp. 65–93. ISBN: 978-0-12-741252-8. DOI: 10.1016/B978-0-12-741252-8.50010-8. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780127412528500108.

[Hil79] **Hill, Rodney**. "Aspects of Invariance in Solid Mechanics". en. In: *Advances in Applied Mechanics*. Vol. 18. Elsevier, 1979, pp. 1–75. ISBN: 978-0-12-002018-8. DOI: 10.1016/S0065-2156(08)70264-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S0065215608702643.

[Jay57] **Jaynes, E. T.** "Information Theory and Statistical Mechanics". In: *Phys. Rev.* 106.4 (May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620.

[KD21] **Kolev, B. and R. Desmorat**. "An Intrinsic Geometric Formulation of Hyper-Elasticity, Pressure Potential and Non-Holonomic Constraints". en. In: *Journal of Elasticity* 146.1 (Sept. 2021), pp. 29–63. ISSN: 0374-3535, 1573-2681. DOI: 10.1007/s10659-021-09853-

5. URL: https://link.springer.com/10.1007/s10659-021-09853-5.

[Kla+17] **Klambauer, Günter et al.** "Self-Normalizing Neural Networks". en. In: arXiv:1706.02515 (Sept. 2017). arXiv:1706.02515 [cs, stat]. URL: http://arxiv.org/abs/1706.02515.

[KW22] **Kingma, Diederik P. and Max Welling**. "Auto-Encoding Variational Bayes". en. In: arXiv:1312.6114 (Dec. 2022). arXiv:1312.6114 [cs, stat]. URL: http://arxiv.org/abs/1312.6114.

[LeC98] **LeCun, Yann**. "The MNIST database of handwritten digits". In: *http://yann.lecun.com/exdb/mnist/* (1998).

[Les+93] **Leshno, Moshe et al.** "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". en. In: *Neural Networks* 6.6 (Jan. 1993), pp. 861–867. ISSN: 08936080. DOI: 10.1016/S0893-6080(05)80131-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S0893608005801315.

[Mai99] **Maiorov, V.E**. "On Best Approximation by Ridge Functions". en. In: *Journal of Approximation Theory* 99.1 (July 1999), pp. 68–94. ISSN: 00219045. DOI: 10.1006/jath.1998.3304. URL: https://linkinghub.elsevier.com/retrieve/pii/S0021904598933044.

[Mak19] **Makhzani, A.** "Implicit Autoencoders". en. In: *ICLR 2019* arXiv:1805.09804 (Feb. 2019). arXiv:1805.09804. URL: http://arxiv.org/abs/1805.09804.

[Mal09] **Mallat, S. G.** *A wavelet tour of signal processing: the sparse way*. en. 3rd ed. Amsterdam; Boston: Elsevier Academic Press, 2009. ISBN: 978-0-12-374370-1.

[Mal16] **Mallat, Stéphane**. "Understanding deep convolutional networks". en. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (Apr. 2016), p. 20150203. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2015.0203. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0203.

[MMD20] **Mouton, Coenraad, Johannes C. Myburgh, and Marelie H. Davel**. "Stride and Translation Invariance in CNNs". en. In: vol. 1342. arXiv:2103.10097 [cs]. 2020, pp. 267–281. DOI: 10.1007/978-3-030-66151-9_17. URL: http://arxiv.org/abs/2103.10097.

[Mul19] **Mula, Olga**. *Optimisation et programmation dynamique*. Master mention Mathématiques Appliquées, 1è re année Université Paris Dauphine. 2019.

[Nes83] **Nesterov, Yurii**. "A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2)". In: *Doklady an ussr*. Vol. 269. 1983, pp. 543–547.

[Pey20] **Peyré, Gabriel**. *Course notes on Optimization for Machine Learning*. Notes de cours de l'École Normale Supérieure. 2020. URL: https://mathematical-tours.github.io.

[Pin99] **Pinkus, Allan**. "Approximation theory of the MLP model in neural networks". en. In: *Acta Numerica* 8 (Jan. 1999), pp. 143–195. ISSN: 0962-4929, 1474-0508. DOI: 10.1017/S0962492900002919. URL: https://www.cambridge.org/core/product/identifier/S0962492900002919/type/journal_article.

[Rio18] **Rioul, Olivier**. "Une théorie mathématique de la communication". In: *BibNum* (Jan. 2018). ISSN: 2554-4470. DOI: 10.4000/bibnum.1190. URL: http://journals.openedition.org/bibnum/1190.

[SG20] **Semih Kayhan, Osman and Jan C. van Gemert**. "On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location". en. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 14262–14273. ISBN: 978-1-72817-168-5. DOI: 10.1109/CVPR42600.2020.01428. URL: https://ieeexplore.ieee.org/document/9156444/.

[Sha48] **Shannon, C. E.** "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[TN04] **Truesdell, C. and W. Noll**. "The Non-Linear Field Theories of Mechanics". en. In: *The Non-Linear Field Theories of Mechanics*. Ed. by Stuart S. Antman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 1–579. ISBN: 978-3-642-05701-4. DOI: 10.1007/978-3-662-10388-3_1. URL: http://link.springer.com/10.1007/978-3-662-10388-3_1.

[Wey50] **Weyl, Hermann**. *The theory of groups and quantum mechanics*. Courier Corporation, 1950.

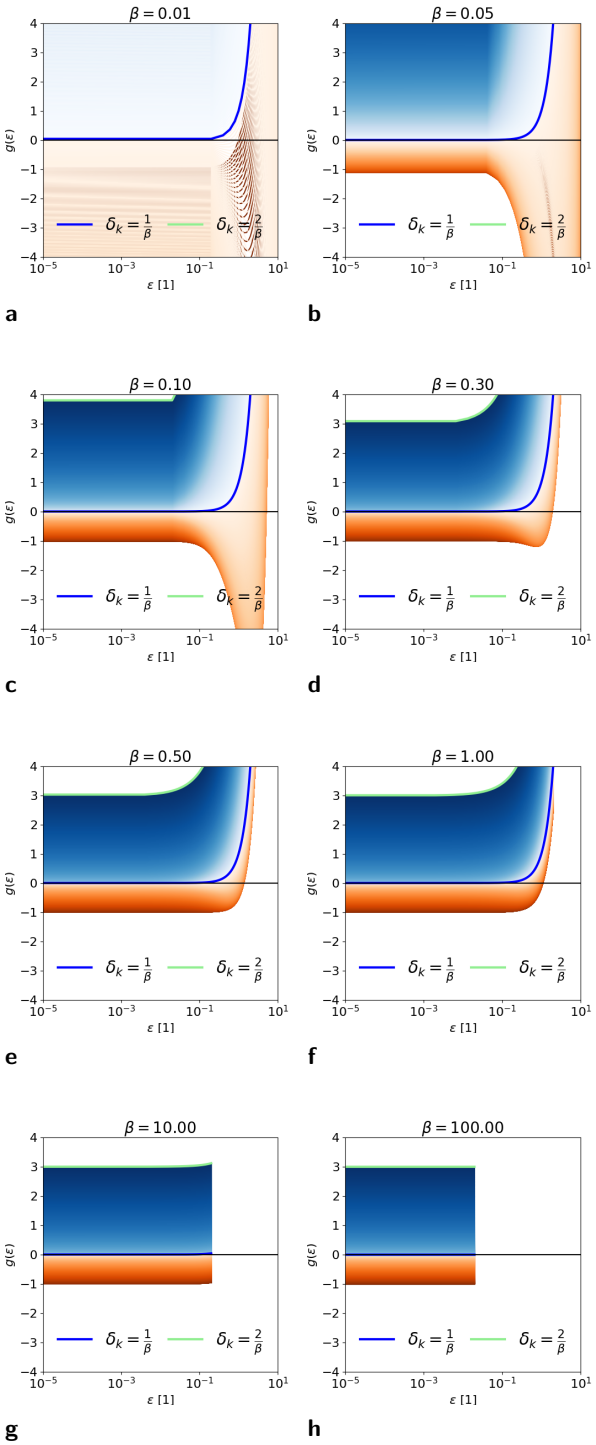

Figure 8: $g(\varepsilon)$, with $\varepsilon = \frac{\|\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}\|}{\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|}$ for different values of the Lipschitz constant β. The contour plots represent different values of δ_k. Blue-graded lines represent values of $\delta_k > \frac{1}{\beta}$. Orange-graded lines represent values of $\frac{2}{\beta} < \delta_k \leq \frac{1}{\beta}$. Blue solid line represents $\delta = \frac{1}{\beta}$ and green line $\delta_k = \frac{2}{\beta}$.

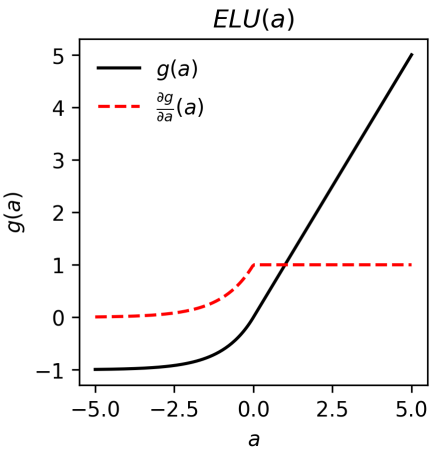

Figure 9: ELU activation function $g(a)$ and its derivative $\frac{\partial g}{\partial a}$.

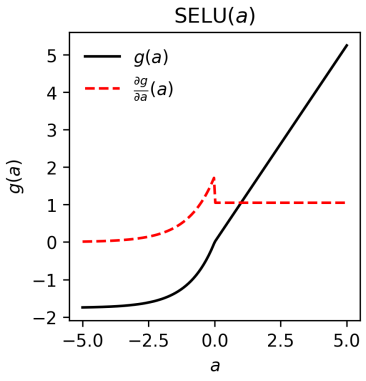

Figure 10: $SELU$ activation function $g(a)$ and its derivative $\frac{\partial g}{\partial a}$.

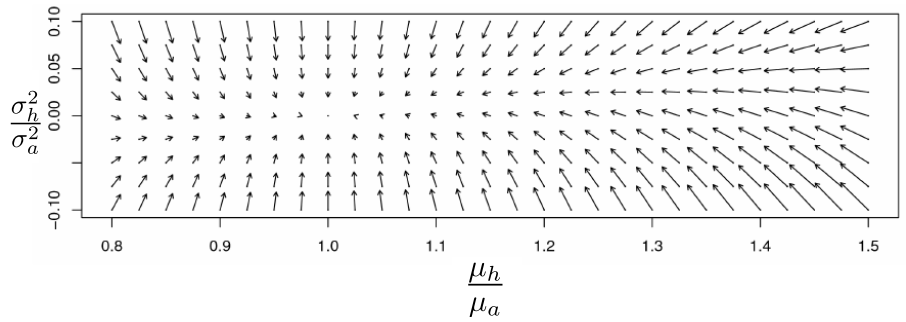

Figure 11: Self-centering properties of the $SELU$ activation function, reprinted from [Kla+17]. μ_a and μ_h represent the average pre-activation and activation of any neuron at layer ℓ of a $\mathcal{MLP}$, whereas σ_a^2 and σ_h^2 represent their variances.