UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

SYMMETRIES IN OVERPARAMETRIZED NEURAL NETWORKS: A MEAN FIELD VIEW

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO

JAVIER ESTEBAN MAASS MARTÍNEZ

PROFESOR GUÍA:
JOAQUÍN FONTBONA TORRES

MIEMBROS DE LA COMISIÓN:
FELIPE TOBAR HENRÍQUEZ
DANIEL REMENIK ZISIS
ROBERTO CORTEZ MILAN

Este trabajo ha sido parcialmente financiado por :
CMM ANID BASAL FB210005

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS Y MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO
POR: JAVIER ESTEBAN MAASS MARTÍNEZ
FECHA: 2024
PROF. GUÍA: JOAQUÍN FONTBONA TORRES

SIMETRÍAS EN REDES NEURONALES SOBREPARAMETRIZADAS: UNA MIRADA DE CAMPO MEDIO

Durante la última década, las Redes Neuronales Artificiales (NNs) han ganado gran popularidad por su éxito en aplicaciones prácticas como la visión computacional y el procesamiento de lenguaje natural; sin embargo, la comprensión teórica de estos modelos es aún escasa en general. Esta tesis pretende mejorar esta comprensión, estudiando el proceso de aprendizaje de las NNs para entender cómo estas aprovechan las simetrías de un problema para mejorar su rendimiento y poder de generalización.

Nuestro trabajo aborda dos temas principales: el análisis del límite de Campo Medio (MF) de las NNs, que provee una teoría para entender el entrenamiento de redes de ancho infinito (viéndolo como un proceso *no lineal*, *más expresivo* que otros *regímenes sobreparametrizados* de la literatura); y el uso de técnicas como Data Augmentation, Feature Averaging o las NNs Equivariantes para aprovechar las simetrías presentes en los datos de un problema.

El objetivo es comprender cómo se manifiestan las simetrías de los datos en el límite MF del entrenamiento de la NN: ¿Es también *simétrico* (en algún sentido) el proceso límite? ¿Cómo se ve la dinámica límite cuando se emplean técnicas para *aprovechar* las simetrías? ¿Se logran mejores velocidades de convergencia global? ¿Aparecen estrategias de *aprovechamiento de simetrías* significativamente mejores que otras? Nuestro trabajo proporciona las bases teóricas para responder a estas preguntas, y las aborda, en su mayoría, de forma efectiva.

La tesis se estructura en cuatro capítulos principales: una revisión bibliográfica tanto del límite MF de NNs sobreparametrizadas, como del estudio de simetrías en NNs mediante acciones de grupo en los Capítulos 2 y 3; seguido por nuestras contribuciones principales en los Capítulos 4 y 5. Nuestros aportes incluyen la formalización de la noción de *simetría* en el contexto de NNs sobreparametrizadas (permitiendo *caracterizar* las NNs equivariantes en el contexto MF), la exploración de propiedades de Transporte Óptimo para medidas invariantes (y *concentradas en subespacios*), la adaptación de técnicas tradicionales de *aprovechamiento* de simetrías al contexto MF, y el estudio exhaustivo de las propiedades de funcionales simétricos, sus minimizadores y sus flujos de gradiente de Wasserstein (WGFs). En particular, se demuestra que las funciones invariantes tienen minimizadores *invariantes* y producen WGFs con trayectorias *invariantes* cuando se inicializan correctamente.

En resumen, esta tesis profundiza la comprensión acerca de cómo influyen las simetrías de los datos en el entrenamiento de las NNs (en el límite MF). Los resultados no solo contribuyen al ecosistema de investigación teórica sobre redes neuronales, sino que también podrían dar ideas prácticas para nuevas arquitecturas de NN y/o nuevos mecanismos de optimización.

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS
Y MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO
POR: JAVIER ESTEBAN MAASS MARTÍNEZ
FECHA: 2024
PROF. GUÍA: JOAQUÍN FONTBONA TORRES

SYMMETRIES IN OVERPARAMETRIZED NEURAL NETWORKS: A MEAN FIELD VIEW

Over the last decade, Artificial Neural Networks (NNs) have gained widespread popularity due to their success in practical applications ranging from computer vision to natural language processing; the theoretical understanding of these models is, however, still largely unexplored. This thesis aims to attack this problem by delving into the learning dynamics of NNs, particularly studying how they can leverage problem-specific symmetries to improve their performance and generalization power.

Our investigation revolves around two main themes: the analysis of the Mean Field (MF) limit of NNs, which provides a simplified framework for understanding the training process of infinitely wide networks (in a non-linear and *arguably more expessive* fashion than concurrent overparametrized regimes); and the use of symmetry-leveraging techniques (such as Data Augmentation, Feature Averaging and Equivariant Architectures) to profit from the symmetries present in the training data.

Our objective is to understand how data symmetries impact the MF limit of NN training: Is the limiting process also *symmetric* (in some sense)? Do traditional symmetry-leveraging techniques *change* the limiting dynamics? Do they yield tighter convergence rates for the known global convergence results? Are some symmetry-leveraging strategies significantly *better* than others? Our work provides the theoretical grounds for answering these questions, and effectively addresses most of them.

The thesis comprises four main chapters: a literature review on the MF Limit of shallow NNs and the group-theoretical understanding of symmetries in NNs in Chapters 2 and 3, followed by the bulk of our novel contributions in Chapters 4 and 5. These include formalizing the notion of *symmetry* in NN learning tasks (allowing us to *characterize* equivariant NNs within the usual MF framework), exploring Optimal Transport properties of invariant (and *subspace-concentrated*) measures, adapting traditional symmetry-leveraging techniques to the MF setting, and thoroughly studying properties of symmetric functionals, their minimizers and their Wasserstein Gradient Flows (WGFs). Notably, we show that invariant functionals have *invariant* minimizers and produce *invariant WGF trajectories* when initialized correctly.

In summary, this thesis develops significant (and novel) theoretical contributions that deepen our understanding of how data symmetries impact NN training at the MF level. The findings not only contribute to the theoretical landscape of NN research but could also eventually offer practical insights for novel NN architectures and optimization mechanisms.

Para mi familia, mis amigos, y mi Dani...

Agradecimientos

Me gustaría comenzar por agradecerle a mi familia. Gracias a mis padres por darme todas las oportunidades y el apoyo imaginables para que me fuera bien en la vida. Gracias por sus consejos y sus lecciones; sé que siempre han buscado lo mejor para nosotros. Gracias, sobre todo, por su cariño incondicional: yo también los quiero mucho. Gracias también a mis hermanos, Ale y Seba, por haberme guiado, escuchado, aconsejado y apoyado cada vez que lo necesité. Los quiero mucho, familia, y sin ustedes jamás habría llegado a este punto.

Agradezco a la Dani, mi polola, por acompañarme durante toda la odisea universitaria. Gracias por ser mi mejor amiga desde el primer día. Gracias por acompañarme a la distancia los dos años de Francia. Gracias por estar siempre ahí para escucharme, entenderme y aconsejarme. Gracias por quererme y por ser la mejor polola del universo. La odisea seguirá más allá de la universidad y nos irá genial, como en todo lo que hacemos. Te amo.

Agradezco a todo el resto de mi (gran) familia: a mis abuelos por inspirarme con su trayectoria, a mis tíos por apoyarme cuando lo necesité, y a mis primos por acompañarme mientras crecíamos juntos. Gracias especiales a la Nancy, mi segunda madre, por haber sido una parte crucial en la construcción de quien soy hoy en día. Los quiero mucho a todos.

Agradezco también a todxs mis amigxs y colegas. Les agradezco a lxs del colegio, a lxs de la vida, a lxs de plan común, a lxs que pude conocer en Francia (toda la BàJ incluida), a lxs que conocí volviendo al DIM, y a lxs que llegaron a mi vida con ISATEC. Cada unx de ustedes me ha ayudado a construir y consolidar mi personalidad; a descubrirme, conocerme y reírme de mí mismo (las risas no faltaron). Sin duda nos seguiremos viendo continuamente.

Agradezco a Vincent y Anaëlle, mis profesores guía del *parcours recherche* en Francia. Gracias por acogerme de la mejor manera posible en un país y un sistema que me eran ajenos. Gracias por introducirme a lo que significa "investigar" y por haberme guiado, acompañado y apoyado en todo el proceso de publicar nuestro trabajo.

Gracias totales a mi profesor guía, Joaquín Fontbona, por acogerme de vuelta en Chile, introducirme al tema del *Campo Medio* y acompañarme en este año de tesis. Gracias por jugártela con el tema y por apoyarme siempre que necesité: sin duda seguiremos trabajando y encontrando resultados interesantes. Gracias también a los miembros de esta comisión (Felipe, Daniel y Roberto) por apoyarme en estos últimos pasos antes del final.

Agradezco finalmente a todxs quienes, de una forma u otra, han puesto su granito de arena en este proceso que hoy culmina con mi titulación: sin ustedes no lo habría logrado.

Table of Content

1 Introduction **1**

2 Learning with Neural Networks **4**

2.1 General Supervised Learning Problem 4

2.1.1 Multilayer Neural Networks 5

2.2 Mean Field Limit of Shallow Neural Networks 6

2.2.1 Theory of the Mean Field Limit of Shallow Neural Networks 8

2.2.2 Universality 9

2.2.3 Learning Problem 11

2.2.4 Wasserstein Spaces 14

2.2.5 Wasserstein Gradient Flows 16

2.2.6 Stochastic Gradient Descent Dynamics 21

2.2.7 SGD as a WGF: a Law of Large Numbers 23

2.2.8 Convergence to a Global Optimizer 30

2.2.9 Applicability to the NN setting 37

2.3 Mean Field Limit in *Deep Neural Networks* 38

3 Exploiting Symmetries with Neural Networks **41**

3.1 Symmetries as Group Actions 42

3.2 Theory of Invariant Measures 46

3.2.1 Base Results 46

3.2.2 Applicability in the Learning Framework 49

3.3 Symmetrization of Models 50

3.4 Techniques for Leveraging Symmetries 53

3.4.1 Data Augmentation (DA) 53

3.4.2 Feature Averaging (FA) 55

3.4.3 Equivariant Architectures (EA) 57

4 Symmetry in the study of “shallow” Neural Networks 62

4.1 Defining Symmetries in the Parameter Space 63

4.1.1 Making the Parameter Space Interesting 63

4.1.2 Making a (somewhat) interesting group representation 64

4.1.3 Generalization is key 66

4.2 Wasserstein Spaces and Invariance 68

4.2.1 Invariant Measures as Orbit Measures 72

4.3 Differentials and Integrals of Equivariant Functions 74

4.4 Optimizing a Symmetric Functional 75

5 Symmetries of the Mean Field Dynamic 82

5.1 Symmetrizing NN Models 82

5.2 Networks that exploit Symmetry 84

5.2.1 DA, FA and EA revisited 86

5.3 Symmetries for a WGF 91

5.3.1 Neural Networks that Exploit Symmetry 98

6 Conclusion 100

Bibliography 111

Annexes 113

Annex A Reading Guide and Summary of Contributions 113

Annex B Application of our framework to some EA 119

B.1 Example of *Deep Sets* 119

B.2 Some Explicit Calculations for the typical example 121

Annex C Proofs 124

C.1 Proofs for Section 2.2.2 124

C.2 Proofs for Section 2.2.3 129

C.3 Proofs for Section 2.2.4 130

C.4 Proofs for Section 2.2.5 131

C.5 Proofs for Section 2.2.7 133

C.6 Proofs for Section 2.2.8 134

C.7 Proofs for Section 3.2.2 136

C.8 Proofs for Section 3.3 138

C.9 Proofs for Section 3.4.3 140

C.10 Proofs for Section 4.1 142

C.11 Proofs for Section 4.2 144

C.12 Proofs for Section 4.3 156

C.13 Proofs for Section 4.4 158

C.14 Proofs for Section 5.1 163

C.15 Proofs for Section 5.2 164

C.16 Proofs for Section 5.3 170

Annex D Reference Results from the Literature 178

D.1 Mean Field Theory of Shallow NNs 178

D.1.1 Key Assumptions for the results of chapter 2 178

D.1.2 A note about the Central Limit Theorem of shallow NNs 181

D.1.3 Existence and Uniqueness of Solutions for the McKean-Vlasov Dynamic 183

Chapter 1

Introduction

Despite their overwhelming success in practice, artificial neural networks (NNs) remain a mystery from a theoretical point of view. At present, very few mathematical results exist to explain their effectiveness. In this context, this thesis investigates the learning process of NNs, and how they can exploit a problem's symmetries in order to improve their performance and generalization power.

In particular, the data of a problem could be known to respect some sort of *symmetry*: for instance, the task of *detecting whether a dog is present (or not) in a given image* should be *independent* of the orientation of said image. We use *group theory* (more specifically *group actions* to encode such *symmetries* in a mathematical fashion. In particular, the *dog detection* task previously described would be defined as *invariant* with respect to the action of the group of *rotations* of the plane (or simply, *rotation-invariant*). Analogously, if we cared about the *specific absolute position* of the dog in the picture (with respect to the origin), this would be a *rotation-equivariant* task, meaning that a *rotation* on the input is expected to result in an equal *rotation* of the output of the task.

There are multiple techniques that allow NNs to take advantage of a problem's symmetry, most notably we have (among many others) *Data Augmentation* (**DA**) and *Feature Averaging* (**FA**). These techniques involve trying to *drive* the NN *into respecting the symmetries*, but without necessarily harnessing this invariance to *simplify* the model at hand. NN Models that are hardwired to *respect* the symmetries of a problem (for instance, through *parameter-sharing*) correspond to what we call *equivariant architectures* of NNs (EA). These kinds of architectures are actually widely used in practice, and many well-known NN models are based upon this concept (such as CNNs, Transformers, GNNs, among many others).

On the other hand, we have what's known as the Mean Field (MF) limit of NNs; a theoretical framework that attempts to mathematically *understand how* NNs learn in an *overparametrized* setting. This formalism considers NNs which can be taken to have *infinite width*, in turn allowing for a conceptual simplification of their training process. The complex optimization procedure driven by Stochastic Gradient Descent (SGD) for *finite* networks, is replaced (in the limit) by a stochastic process driven by a partial differential equation (PDE) that's conceptually easier to solve. Particularly, under certain assumptions, it is known that this *mean field limit* training process can converge to the global optimum of the learning problem. Despite other interesting approaches for *overparametrized models* (such as *Random Features* or the famous *Neural Tangent Kernel*), the MF limit seems to give the most *meaningful insight* on how SGD *makes the parameter distribution evolve towards a global minimum* during training.

With both the concept of the MF limit of NNs and the idea of exploiting the *problem's symmetry* to build better models, we can state this work's objectives. We hope to understand how data symmetries could have an impact on the MF limit of NN training. More specifically, we seek to answer whether the limiting process could also be symmetric (in some sense) whenever the input data is; whether the use of symmetry-leveraging techniques could allow a model to *profit* from data symmetries at the MF level (e.g. Do the limiting dynamics change? Do we get tighter convergence rates for the known global convergence results?); and whether this allows us to significantly distinguish *a preferred* symmetry-leveraging strategy amongst them (i.e. is there any one that's significantly better than the rest?).

Providing satisfactory answers to these questions could eventually be useful for the design and training of more efficient and accurate NNs in the future; potentially generating a significant impact on a wide range of very relevant practical applications ranging from computer vision to natural language processing.

With this in mind, throughout our work, we will provide the necessary theoretical grounds for effectively addressing (and hopefully answering) most these questions. All in all, the *main contributions* of our work are the following:

- We provide a *unified* framework under which the MF theory of *shallow* NNs can be understood. In the process, we slightly extend some relevant results from the literature to fit into this *unified* setting (e.g. see Theorem 1, Proposition 10, among others).
- We describe some of the main elements from the theory of *symmetry-leveraging* in learning problems; notably, providing an extension of a known result regarding the *symmetrization gap* (Lemma 10).
- We *generalize* the notion of an *equivariant NN architechture* to the setting of *shallow NN models* used in the traditional MF theory. We prove that this definition is *consistent*, and that it satisfies relevant properties in our context (see Section 4.1).
- We thoroughly describe the spaces of *G-invariant probability measures* and *probability measures concentrated on $\mathcal{E}^G$ (subspace of G-equivariant parameters)*, along with some of their interesting properties (Proposition 25, Lemma 13 and Proposition 26). We also characterize G-invariant measures as *equivalent* to *measures over $G\backslash\mathcal{Z}$* (Proposition 28).
- We prove key properties of *differentials and integrals* of equivariant functions (Section 4.3).
- We prove a variant of *Jensen's inequality* (Proposition 32) and use it to show that G-invariant functionals over the space of probability measures can be minimized over the space of G-invariant measures (Proposition 33). We provide a counterexample of the analogous result for *measures concentrated on $\mathcal{E}^G$* (Proposition 35), and a way to *avoid it* when universality holds (Proposition 36).
- We translate the main *symmetry-leveraging techniques* to the MF setting (Proposition 40), and show that optimizing under **DA** or **FA** is *esentially equivalent* (Proposition 42). We also provide a bound in the case of *approximately invariant data* (Proposition 44).

- We prove that, when the initial condition (i.c.) is G-invariant, WGFs of G-invariant functionals have G-invariant trajectories (Theorem 14). We prove that an analog holds whenever the i.c. is *concentrated on* $\mathcal{E}^G$ (Theorem 15). We finally show that the **DA** and **FA** training dynamics *exactly coincide* under a G-invariant i.c. (Corollary 14).

We will establish these contributions along the main chapters of this work, which we desribe in the following paragraphs to provide the reader with a *roadmap* of this thesis:

Chapter 2 presents a thorough review of the literature on the topic of the **Mean Field (MF) Limit** in the context **Overparametrized Neural Networks (NNs)**, as well as most relevant theoretical elements for defining such an object. Many known results from the literature are presented, though some are *adapted* and *generalized* to make them fit into a *unified general setting* we try to establish.

Chapter 3 presents the ideas behind the *group theoretical* understanding of *symmetries* in the context of Neural Networks. It displays both relevant theoretical results from the recent literature (concerning *invariant/equivariant* functions and measures), as well as some of the most popular techniques used to *leverage* a problem's symmetries in practical applications. Beyond introducing many key elements from the literature, we also prove an extension of a known result regarding the *symmetrization gap* of a learning problem.

Chapter 4 starts presenting the results from our own *study* of *symmetries* in the NN context, with an Optimal transport (OT) and Mean Field view. In particular, a notion of *equivariant NN* is introduced in the *shallow NN/MF setting*, with some of its basic properties being proven. Many OT properties from invariant (and subspace-concentrated) measures are proved, and similar work is done for the derivatives/integrals of equivariant functions. Finally, functionals over the space of probability measures are heavily studied, particularly proving that *invariant functionals* have minima that correspond to *invariant measures* (however, not necessarily concentrated on the subspace of equivariant parameters).

Chapter 5 culminates our work by employing the discovered facts from all previous chapters to prove properties of the WGF of G-invariant functionals. In particular, some usual notions from **Chapter 3** (notably, model symmetrization, **DA**, **FA** and **EA**) are reintroduced in our setting and studied under this new lens. One of the main results states that the WGF of an *invariant* functional *remains invariant* overtime whenever the initialization is *invariant* as well. Furthermore, it is proven that whenever the initialization is *concentrated on the subspace of equivariant parameters*, then the subsequent flux remains concentrated there overtime. A similar result is derived in order to *compare* **DA** and **FA** under an *invariant* initialization.

Chapter 6 provides a natural conclusion to our work, gathering and summarizing the bulk of our original contributions. It also contains a compilation of open questions to be attacked in our future work.

Finally, the **Annexes** provide illustrative examples of our work, together with all the relevant proofs and technical assumptions of the presented results. In particular, **Annex A** contains a *reading guide* which provides further details about the structure and contributions of this work. The interested reader shall look into it for a more complete description.

Chapter 2

Learning with Neural Networks

2.1 General Supervised Learning Problem

Given measurable spaces $\mathcal{X}$ (the space of *features*) and $\mathcal{Y}$ (the space of *labels*), we consider data of the form $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ following a certain joint probability distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The generic *supervised learning problem*, is one in which we try to find the *best possible model* that will allow us to predict, from a given *feature* X, what the associated *label* Y is.

More specifically, we fix a subset of all possible measurable functions from $\mathcal{X}$ to $\mathcal{Y}$, $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$, (which we call the *hypothesis set*) from where we'll pick our model. To determine the *fitness* of a model $f \in \mathcal{F}$ for the task at hand, we consider a given loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, which will measure, for a given sample $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, *how far off* our model prediction, $f(X)$, is from the real value of the *label*, Y.

To find a model that will work well for *any sample drawn from* π, we will try to minimize the *expected error*, which we call the **population risk** (or also, *generalization error*), R. i.e. for a given model $f \in \mathcal{F}$ we will evaluate:

$$R(f) = \mathbb{E}_{(X,Y)\sim\pi}[\ell(f(X), Y)]$$

We say that a model which minimizes such risk *"generalizes well"* to any sample drawn from the data distribution.

In practice, however, it will be impossible to have access to the law of the data (π); so a good model must be found just from a given i.i.d. sample of data $S = (X_k, Y_k)_{k=1}^m$. That is, we have to fix some algorithm $\mathcal{A} : \bigcup_{m\in\mathbb{N}}(\mathcal{X} \times \mathcal{Y})^m \to \mathcal{F}$ that will allow us to find, for any sample S (of any size), a *good* model $\mathcal{A}(S) \in \mathcal{F}$.

The most popular heuristic for such a problem is to simply choose the model that *minimizes* the **empirical risk** with respect to the sample, $\hat{R}_S$. i.e. for a given model $f \in \mathcal{F}$ and sample S, we will evaluate:

$$\hat{R}_S(f) = \frac{1}{m}\sum_{k=1}^{m} \ell(f(X_k), Y_k)$$

The hope is that, for a large amount of collected data, this quantity will approximate R. We say that a model that minimizes $\hat{R}_S$ *adjusts well to the data* (which doesn't necessarily mean that it will *generalize well*).

With all these elements in mind, we can define a *supervised learning problem* more precisely:

Definition 2.1 ***[Supervised Learning]*** *Given a data distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we aim to find a model $\tilde{f} \in \arg\min_{f \in \mathcal{F}} R(f)$ (i.e. that generalizes well).*

As, in practice, we don't have access to the data distribution π, we approximate such a solution by minimizing the empirical risk (with respect to a given sample S drawn from π); i.e. $\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{R}_S(f)$.

In practice, in order to have computable algorithms, we usually restrict ourselves to hypothesis sets $\mathcal{F}$ that are **parametric**, and the optimization $\hat{R}_S$ is done through some sort of *gradient descent* scheme such as **Stochastic Gradient Descent (SGD)** (we will elaborate on the specifics of this method later on). This *optimization of the empirical risk* is what's known (in practice) as the **training** of a model. In particular, one of the most popular approaches for solving *supervised learning problems* (specially in recent years) has been to consider the class $\mathcal{F}_\Theta$ of **multilayer neural networks**.

2.1.1 Multilayer Neural Networks

First, consider the following definition of what we will consider as a neural network in this work:

Definition 2.2 ***[Fully-Connected Multilayer Neural Network]*** *A neural network (NN) with $L \in \mathbb{N}$ hidden layers is a function between $\mathcal{X} = \mathbb{R}^{d_0}$ and $\mathcal{Y} = \mathbb{R}^{d_L}$, composed of a collection of affine computing units combined with sequential nonlinear activation functions (see illustration in Figure 2.1). Particularly, a fully-connected multilayer NN is such that, for each layer $\ell \in [L] := \{1, \ldots, L\}$:*

- *A number of "neurons" $N_\ell \in \mathbb{N}$ in the layer (such that $N_0 = d_0$ and $N_L = d_L$).*
- *An activation function $\sigma^{(\ell)} : \mathbb{R}^{N_\ell} \longrightarrow \mathbb{R}^{N_\ell}$, which is often taken to be* ***non-linear****.*
- *"Parameters" $\theta^{(\ell)} := (W_\ell, b_\ell) \in \mathbb{R}^{N_\ell \otimes N_{\ell-1}} \times \mathbb{R}^{N_\ell}$ (which we'll use to construct our "affine computing units").*

We denote $N := (N_\ell)_{\ell=0}^L$ and $\sigma := (\sigma^{(\ell)})_{\ell=1}^L$, which are the parameters that define the ***architecture*** *of the network. We similarly define $\Theta_L(N) := \prod_{\ell=1}^L \mathbb{R}^{N_\ell \otimes N_{\ell-1}} \times \mathbb{R}^{N_\ell}$, and we consider the vector containing* ***all of the network's parameters*** *as $\theta = (\theta^{(\ell)})_{\ell=1}^L \in \Theta_L(N)$.*

Then, given a fixed architecture, $A = (N, \sigma)$, a ***Neural Network*** *with parameter θ is the function: $\Phi_\theta^A : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ such that for all $x \in \mathbb{R}^{N_0}$:*

$$x_0 = x,\ x_\ell = \sigma^{(\ell)}(W_\ell x_{\ell-1} + b_\ell)\ \forall \ell \in \{1, \ldots, L\},\ \Phi_\theta^A(x) = x_L$$

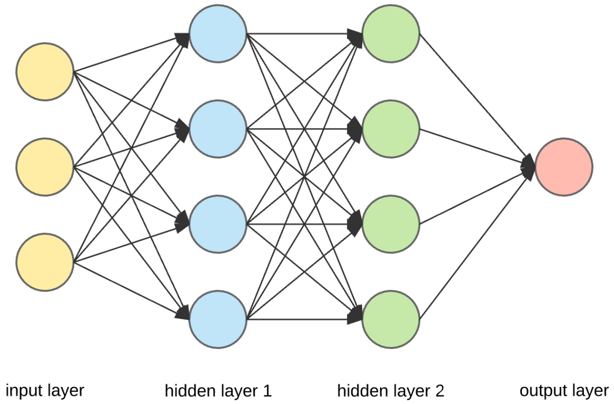


Figure 2.1: Diagram of a Multilayer Neural Network (with $L = 3$). Image taken from [23].

Neural Networks correspond to a specific *type of model*, which has received increased popularity over recent years due to its success on different kinds of complex tasks (from image classification to natural language processing). In particular, they take part in the previously described *general supervised learning setting*, as a specific kind of *parametric hypothesis set* from where we shall pick the best possible model. More concretely, we consider the hypothesis set of *all possible multilayer neural networks*, parametrized by $\Theta := \bigcup_{L\in\mathbb{N}} \bigcup_{N\in\mathbb{N}^L} \Theta_L(N)$; given by:

$$\mathcal{F}_\Theta := \{\Phi_\theta : \theta \in \Theta\}$$

It is known, as shown in Hornik et al. [37], that this collection of **neural network models with arbitrary number of parameters** possesses good properties of *universal approximation* (i.e. any continuous function on a compact set can be approximated arbitrarily well by using NNs). In particular, it is enough to consider *neural networks with a single hidden layer* ($L = 2$ in our setting) to have such a *universal approximation* property.

As for the focus of our work, we will be interested in understanding the behavior of Neural Network models during their **training process** (where data is used to *adjust* the chosen *parameters* in order to minimize the empirical risk of the problem). In particular, it will be interesting to assess **how close this training process can get us from actually minimizing the generalization error of the learning problem**.

2.2 Mean Field Limit of Shallow Neural Networks

The problem of understanding the training of NNs is highly complex. We are seeking the *optimal parameter* such that the resulting network will become a *good model* for predicting the given data distribution. Unfortunately, the function $\theta \mapsto \mathbb{E}_\pi [\ell(\Phi(X, \theta), Y)]$ we are trying to optimize is **highly non-convex**, and thus no global minimization guarantees can be easily deduced.

A strategy to *escape* the issue of *non-convexity*, is to take an asymptotic *limit of the neural network*. More concretely, consider the following setting, which has been intensely studied in the NN literature: (real-valued) neural networks with **1 hidden layer** (i.e. $L = 2$

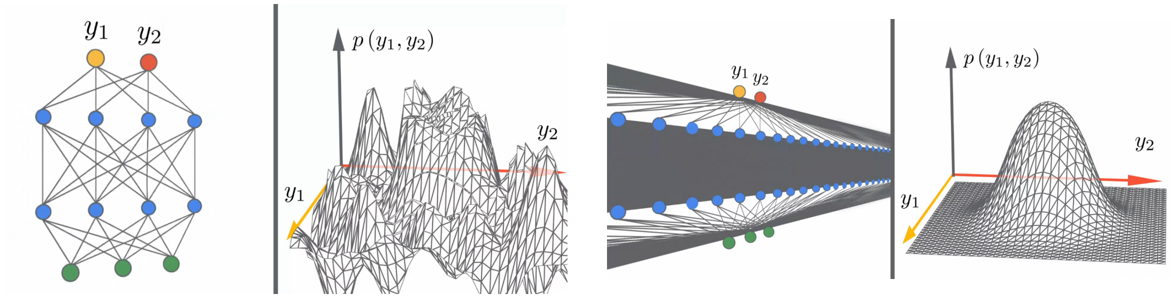


Figure 2.2: Figure taken from [75]. Here, a Bayesian Network of outputs y_1, y_2 is displayed; as the amount of hidden layers increases, the output distribution (denoted $p(y_1, y_2)$) becomes *smoother* (it actually becomes Gaussian; what's known in the literature as a *Neural Network Gaussian Process*).

in the model of Section 2.1.1). This is what's usually referred to as the setting of *shallow* NNs. Now, in order to establish some *asymptotic* result for these NN models, we can re-write them (slightly changing the notation from previous section, and considering $d_0 = d \in \mathbb{N}^*$, $d_1 = N \in \mathbb{N}^*$, $d_L = 1$ and $\sigma : \mathbb{R} \to \mathbb{R}$) as:

$$\begin{aligned} \Phi_\theta^N : \mathbb{R}^d &\to \mathbb{R}^N \to \mathbb{R} \\ x &\mapsto \frac{1}{N} \sum_{i=1}^{N} \sigma_*(x; \theta_i) \end{aligned}$$

where $\sigma_*(x; \theta_i) = w_i \sigma(A_i \cdot x + b_i)$, $\theta_i = (w_i, A_i, b_i) \in \mathbb{R}^D := \mathbb{R}^{d+2}$ and $\theta := (\theta_i)_{i=1}^N \in (\mathbb{R}^D)^N$.

This corresponds exactly to what was introduced in section 2.1.1, but rewritten in a way as to emphasize the role of the $N \in \mathbb{N}^*$ hidden units in the single hidden layer. In particular, we will be interested in taking the limit as $N \to \infty$, under which the behavior of the network will become *smoother* in some sense. This idea is illustrated in figure 2.2, where one can see how in the limit of *infinitely-many hidden units* the objects involved seem to become *better behaved* (in some sense). In this way, one would expect to leverage the *nicer* properties of the *asymptotic* network, in order to obtain better guarantees for the training of the finite-width networks.

Some of the most usual *infinite-width* limits that have been studied in the literature are the so-called **Random Features (RF)** of NNs (see Lee et al. [51], de G. Matthews et al. [22], Novak et al. [64], Garriga-Alonso et al. [35]), the **Neural Tangent Kernel (NTK)** (see Jacot et al. [41], Arora et al. [4], Li et al. [52]), and the **Mean Field (MF) limit of NNs** (driven by Mei et al. [57], Sirignano and Spiliopoulos [78], Chizat and Bach [16], Rotskoff and Vanden-Eijnden [72] and widely developed since then).

All these regimes allow us to connect the training of *wide* NNs with other mathematical objects of interest, such as *Gaussian processes* (for **RF**), the theory of *RKHS*[1] (for the **NTK**) or *Wasserstein gradient flows* (for the **MF** limit). The difference between the obtained asymptotic regimes comes from the initial **assumptions** that have to be made when taking the limit. In particular, very *restrictive* assumptions will lead to relatively limited asymptotic regime for the network. For instance, the **RF** limit involves only training the last layer of

[1] Reproducing Kernel Hilbert Space, see Hofmann et al. [36] for a reference

the network (making the problem convex, but undercutting the model's *expressivity*); and similarly, in the **NTK** limit, the asymptotic training dynamic, though simpler to theoretically understand, is one of **lazy training** (i.e. the distribution of parameters *does not globally change* overtime). Under the lens of such limitations, the **MF** limit appears as a reasonable alternative, where the limiting *training dynamic* involves a *significant evolution* of the parameter distribution (through a Wasserstein Gradient Flow), without many theoretical drawbacks. This work will therefore focus on studying the **MF** limit of neural networks and its properties (particularly under the lens of *symmetries* in the data distribution).

2.2.1 Theory of the Mean Field Limit of Shallow Neural Networks

The main theory to be considered in what follows, is that of the **mean field (MF) limit of shallow neural networks**. Though some insights will be given for the *multilayer* case, the extension of our results to the multilayer setting will be left as future work. The following analysis of the state of the art on the topic is based on the original works (in which the topic was introduced) of Mei et al. [57], Sirignano and Spiliopoulos [78], Rotskoff and Vanden-Eijnden [72] and Chizat and Bach [16]. These are complemented by more recent extensions of the original results, such as Mei et al. [58], Sirignano and Spiliopoulos [81], Chen et al. [15], Bortoli et al. [9] and Descours et al. [24]; as well as the global convergence guarantees (for the *Langevin Dynamics* of the *regularized* problem), as in Hu et al. [38], Chizat [17], Chen et al. [13] and Nitanda et al. [63]. This review tries to be as extensive as possible (with necessary complements included in Chapter D), but further insight shall be sought in the original material.

The macroscopic idea of the Mean Field limit for networks with 1 hidden layer is that, under a suitable *scaling limit* (where both the *width* of the network and the *number of SGD iterations* go to infinity), the *training* dynamics with SGD are asymptotically governed by a **non-linear** PDE corresponding to a **Wasserstein gradient flow** for a *convex* risk function in the space $(\mathcal{P}_2(\mathcal{Z}), \mathcal{W}_2)$. Under the favorable regularization conditions of the problem (e.g. considering the *Langevin Dynamics* of SGD), it can also be shown that the limit dynamic in the Wasserstein space converges (as $t \to \infty$) to the **global minimum of the regularized problem**.

More specifically, let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ (the *feature* space), $\mathcal{Z}$ be a subset of $\mathbb{R}^D$ (the *parameter* space) and $\mathcal{Y}$ be a subset of $\mathbb{R}$ (the *label* space). Consider a *shallow* NN given by:

$$\begin{aligned} \Phi_\theta^N : \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i) \end{aligned}$$

where $N \in \mathbb{N}^*$ is the number of hidden units, and $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is what the literature on the topic refers to as the *activation function* (or simply *unit*). This naturally describes a *shallow* NN (as introduced in the previous section) by setting $\sigma_*(x;\theta_i) = w_i\sigma(A_i \cdot x + b_i)$, with $\theta_i = (w_i, A_i, b_i) \in \mathbb{R}^D := \mathbb{R}^{d+2} = \mathcal{Z}$ and $\theta := (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$. However, this description allows for other settings of interesting models, as *radial basis function networks*, the *deconvolution of sparse spikes*, *density estimation via maximum mean discrepancy minimization* (see [72, 16, 82]). This model even serves to *account for deeper Neural Networks* (with a particular structure) by defining σ_* appropiately (the reader shall seek further reference in Rotskoff and Vanden-Eijnden [72]). Despite this flexibility, we will refer to the general model (with arbitrary σ_*) simply as the *shallow NN model* or the *overparametrized NN model* for the rest of the work.

2.2.2 Universality

Notice that, for a given $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$, the network Φ_θ^N can be seen as an *integral* against the **empirical measure** associated with θ, $\nu_\theta^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$. More specifically[2]:

$$\forall x \in \mathcal{X},\ \Phi_\theta^N(x) = \langle \sigma_*(x;\cdot), \nu_\theta^N \rangle$$

This immediately allows a connection to the theory of *Barron Spaces*, as described in multiple papers following Barron [6]. Consider $\mathcal{M}^S(\mathcal{Z})$ the space of **(signed) Radon measures**[3] over $\mathcal{Z}$ with **finite total variation**, and define the space of functions that can be *realized* as integrals of the *unit* against a measure in $\mathcal{M}^S(\mathcal{Z})$ (sometimes denoted $\mathcal{F}_1$):

$$\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z})) = \{f : \mathcal{X} \to \mathcal{Y} \mid \exists \gamma \in \mathcal{M}^S(\mathcal{Z}),\ f(\cdot) = \int_{\mathcal{Z}} \sigma_*(\cdot, z)\gamma(dz)\}$$

This functional space might seem somewhat *restricted*; however, as stated in Rotskoff and Vanden-Eijnden [72] (and coming previously from Cybenko [20], Barron [6], Park and Sandberg [67]), it has enough expressiveness to achieve *universality*, i.e. $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is actually dense in $L^2(\mathcal{X}, \mathcal{Y}; \mu)$. We will give a proof to a *slightly stronger* version of this fundamental result, as it will be useful in the general setting we want to establish. For this, consider the following assumptions (the original ones employed in Rotskoff and Vanden-Eijnden [72] shall be found in section C.1):

Assumption 1 *Let $\pi_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$, and $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be subsets of general separable Hilbert spaces. Consider that:*

- *$\mathcal{Z}$ is* ***compact****.*
- *$\pi_{\mathcal{X}} - a.s.\ \forall x \in \mathcal{X},\ \sigma_*(x, \cdot)$ is continuous (which we denote $\sigma_*(x, \cdot) \in C(\mathcal{Z}, \mathcal{Y})$).*
- *The function $x \mapsto \sup_{(z,z') \in \mathcal{Z}\times\mathcal{Z}} \langle \sigma_*(x; z), \sigma_*(x; z') \rangle_{\mathcal{Y}}$ is in $L^1(\mathcal{X}, \mathbb{R}; \pi_{\mathcal{X}})$*
- *σ_* is discriminating, in the sense that*[4]*:*

$$\left[\forall z \in \mathcal{Z},\ \langle g, \sigma_*(\cdot; z) \rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} = 0\right] \implies [g = 0\ \ \pi_{\mathcal{X}} - a.e.\ \ in\ \ \mathcal{X}]$$

[2] An alternative scenario would be to have $\mathcal{Z} = \mathbb{R} \times \tilde{\mathcal{Z}}$ and for any $(w_i, \tilde{\theta}_i) \in \mathcal{Z}$, $\sigma_*(x; (w_i, \tilde{\theta}_i)) = w_i \tilde{\sigma_*}(x; \tilde{\theta}_i)$ defining $\gamma_\theta^N = \frac{1}{N}\sum_{i=1}^N w_i \delta_{\tilde{\theta}_i}$, which is a signed measure on $\tilde{\mathcal{Z}}$ such that: $\forall x \in \mathcal{X},\ \Phi_\theta^N(x) = \langle \tilde{\sigma_*}(x; \cdot), \gamma_\theta^N \rangle$. To simplify notation we won't state such a difference explicitly, but rather assume clarity from context.

[3] We consider *signed* radon measures, to allow for a *free linear coefficient* in front of σ_*, letting the integral $\langle \sigma_*, \gamma \rangle$ be potentially *unbounded* even when σ_* is *bounded*.

[4] Note that the *inner product* in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ is introduced. It is defined as $\langle f, g \rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} = \int_{\mathcal{X}} \langle f(x), g(x) \rangle_{\mathcal{Y}} \pi_{\mathcal{X}}(dx)$

Remark About these assumptions, we can say that:

- Assuming the **compactness** of $\mathcal{Z}$ (which might be seen as a strong assumption) *isn't completely hurtful* in this setting, as the universality result will give us a **signed measure** to approximate the relevant functions. Therefore, despite this assumption (together with continuity) *forcing $\sigma_*(x, \cdot)$ to be bounded*, the *free* signed coefficient coming from the signed measure will allow us to be able to approximate even unbounded target functions.
- As it is stated, the *discriminating* assumption for σ_* might be different from the usual *discriminating* assumption employed in the literature, which states that a function $\sigma : \mathbb{R} \to \mathbb{R}$ is discriminating if $\forall \mu \in \mathcal{M}^S(\mathcal{X})$:

  $$\int_{\mathbb{R}^d} \sigma(a^T x + b) d\mu(x) = 0 \ \forall a \in \mathbb{R}^d, \ \forall b \in \mathbb{R} \Longrightarrow \mu \equiv 0$$

  However, such an assumption is actually *stronger* that the one we're currently considering:

  Proposition 1 *Consider traditional setting of shallow neural networks. i.e. Let $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}^c$ and $\mathcal{Z} = \mathbb{R}^{c \times b} \times \mathbb{R}^{d \times b} \times \mathbb{R}^b$, and $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be defined as:*

  $$\forall x \in \mathcal{X}, \ \forall \theta = (w, a, b) \in \mathcal{Z}, \ \sigma_*(x; \theta) := w\sigma(a^T x + b)$$

  for $\sigma : \mathbb{R} \to \mathbb{R}$ an activation function that's applied ***pointwise*** *(i.e. $\forall z \in \mathbb{R}^b, \ (\sigma(z))_i = \sigma(z_i) \ \forall i \in \{1. \dots, b\}$).*
  Assume that σ is discriminatory in the sense that $\forall \mu \in \mathcal{M}^S(\mathcal{X})$:

  $$\int_{\mathbb{R}^d} \sigma(a^T x + b) d\mu(x) = 0 \ \forall a \in \mathbb{R}^d, \ \forall b \in \mathbb{R} \Longrightarrow \mu \equiv 0$$

  Then, σ_* is discriminatory in the sense of assumption 1:

  $$\left[\forall z \in \mathcal{Z}, \ \langle g, \sigma_*(\cdot; z) \rangle_{L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})} = 0\right] \Longrightarrow [g = 0 \ \ \pi_{\mathcal{X}} - a.e. \ in \ \ \mathcal{X}]$$

  Proof. See Annex C.1. □

With Assumption 1 in place, we proceed to prove the following results (which are standard in the neural network literature). The proof is esentially the same as in Rotskoff and Vanden-Eijnden [72], only with a few variations accounting for the fact that we work with *Bochner Integrals*. We do include the proofs for completeness:

Proposition 2 *Under assumption 1, the space $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is a linear subspace of $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$.*

Proof. See Annex C.1. □

From there, we get the desired *universality* result:

Theorem 1 (**Universality**) *Under assumption 1, $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is a* ***dense subspace*** *of $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ (in the Bochner L^2-norm topology).*

Proof. See Annex C.1. □

This result is a key part of the current theoretical understanding of NNs and their approximation power: despite the simplicity of the *shallow NN* model, when σ_* and $\mathcal{Z}$ satisfy good properties, any square-integrable function from $\mathcal{X}$ to $\mathcal{Y}$ can be approximated by functions in $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$. This result established with a *general* $\mathcal{Z}$ will be really useful, in particular when considering parameters that constitute **equivariant NNs** (which will soon be introduced).

Now, functions in $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ aren't by themselves the most interesting objects. We want to know whether a given class of *finite-width* neural networks is *universal* or not. That is, consider[5]:

$$\mathcal{N}_{\sigma_*}(\mathcal{Z}) := \left\{ f : \mathcal{X} \to \mathcal{Y} \mid \exists N \in \mathbb{N},\ \exists (c_i, \theta_i)_{i=1}^N \subseteq \mathbb{R} \times \mathcal{Z},\ f(\cdot) = \frac{1}{N}\sum_{i=1}^N c_i \sigma_*(\cdot;\theta_i) \right\}$$

It is clear that $\mathcal{N}_{\sigma_*}(\mathcal{Z}) \subseteq \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ (by considering the *empirical measure* of the parameters: $\frac{1}{N}\sum_{i=1}^N c_i \delta_{\theta_i}$). As noted in Rotskoff and Vanden-Eijnden [72], any function in $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ can be *approximated* (pointwise) with a sequence of *shallow NNs* with *finitely many units*:

Proposition 3 *Under assumption 1, for any $f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$, there exists $(c_n, \theta_n)_{n\in\mathbb{N}} \subseteq \mathbb{R} \times \mathcal{Z}$ such that we can define a finite neural network: $\forall x \in \mathcal{X},\ f_n(x) = \langle \sigma_*(x, \cdot), \frac{1}{n}\sum_{i=1}^n c_i \delta_{\theta_i} \rangle = \frac{1}{n}\sum_{i=1}^n c_i \sigma_*(x;\theta_i)$; that satisfies:*

$$f_n \xrightarrow[\pi_{\mathcal{X}}-a.s.]{n\to\infty} f$$

If we further assume $\pi_{\mathcal{X}}$ to be compactly supported, this convergence holds in $L^p(\pi_{\mathcal{X}})$, $\forall p \geqslant 1$

Proof. The proof is included in Annex C.1 for completeness. □

This result shouldn't be surprising, as it comes naturally from applying the LLN (though making this explicit isn't completely trivial). It also allows us to state a more *down-to-earth* universality result:

Corollary 1 *If assumption 1 holds and $\pi_{\mathcal{X}}$ is compactly supported, then: $\mathcal{N}_{\sigma_*}(\mathcal{Z})$ is **dense** in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$.*

Proof. See Annex C.1. □

2.2.3 Learning Problem

Besides the good approximation properties of the *shallow NN model* from a *functional* perspective, we are interested in its capabilities for solving a given *statistical learning problem.*

[5] Recall, as mentioned when introducing *Barron Spaces*, that the *linear coefficient* is included to allow these models to be *unbounded* even when σ_* might be *bounded.* Further in the thesis, we will limit our scope to *probability measures* over $\mathcal{Z}$, and so our notation might be adapted (this will be made clear later on).

Assume that we have a probability law $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that models our *data distribution* (i.e. our data samples will distribute as $(X, Y) \sim \pi$). Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a **loss function**[6], typically taken to be *convex* (at least in one of its arguments). A really common choice in the literature for such a function, is the **quadratic loss** $\ell(y, \hat{y}) = \frac{1}{2}\|y - \hat{y}\|_{\mathcal{Y}}^2$.

We are interested in obtaining a neural network model with **good generalization power**; i.e. if we consider the *population risk* associated to a given measurable function $f : \mathcal{X} \to \mathcal{Y}$ (a *model*, also denoted $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$) as:

$$R(f) = \mathbb{E}_{(X,Y)\sim\pi}[\ell(f(X), Y)]$$

we are interested in finding a *good set of parameters* $\theta \in \mathcal{Z}^N$ (with $N \in \mathbb{N}$) such that the obtained NN model will minimize the *population risk* $R(\Phi_\theta^N)$.

Now, recall that the problem of finding $\inf_{\theta\in\mathcal{Z}^N} R(\Phi_\theta^N)$ is **highly non-convex** due to the introduction of non-linearities through the *activation function.* Fortunately, in the setting of *shallow NNs* we can reformulate it recalling that for any $\theta \in \mathcal{Z}^N$ and $x \in \mathcal{X}$:

$$\Phi_\theta^N(x) = \langle \sigma_*(x; \cdot), \nu_\theta^N \rangle$$

where $\nu_\theta^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ is the *empirical measure associated with θ*. So, instead of seeing this as an optimization problem over the space of parameters, we can see it as an optimization problem over the infinite-dimensional space of probability measures, by considering the *population risk functional* (with a slight abuse of notation) as $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, given by[7]:

$$R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X; \cdot), \mu\rangle, Y)]$$

We then *rephrase* our problem as that of finding $\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$. The advantage is that, beyond the *more abstract* setting, the functional over probability measures may satisfy good properties. For instance, it is standard to notice that:

Proposition 4 *Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be subsets of separable Hilbert spaces, $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Consider $R : \mathcal{M}(\mathcal{Z}) \to \mathbb{R}$ defined as $R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X; \cdot), \mu\rangle, Y)]$, $\forall \mu \in \mathcal{M}(\mathcal{Z})$. If we suppose that ℓ is **convex on its first argument**, then R is convex (as a function), meaning that: $\forall \lambda \in [0, 1]$, $\forall \nu, \tilde{\nu} \in \mathcal{M}(\mathcal{Z})$*

$$R(\lambda\nu + (1 - \lambda)\tilde{\nu}) \leqslant \lambda R(\nu) + (1 - \lambda)R(\tilde{\nu})$$

Now, beyond the *convexity* of R in this setting, when the loss is chosen to be *quadratic* (even when $\mathcal{Y}$ is a general separable Hilbert space), *solving the learning problem* becomes essentially a problem of *function approximation.* In particular, it is a known fact that whenever $\pi|_{\mathcal{Y}}$ has second order moments (and the *quadratic loss* is considered), the *population risk* allows for the following decomposition:

Lemma 1 *Let $\mathcal{X}, \mathcal{Y}$ be subsets of separable Hilbert spaces, let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be such that $\pi|_{\mathcal{Y}}$ has finite second order moment, and consider the quadratic loss $\ell(y, \hat{y}) = \frac{1}{2}\|y - \hat{y}\|_{\mathcal{Y}}^2$. Then,*

[6] In general, it's simply a function that allows to measure *similarity* between elements of $\mathcal{Y}$.

[7] We shall indistinctly consider it as a functional over the space of all **positive Radon measures** over $\mathcal{Z}$, $\mathcal{M}(\mathcal{Z})$. i.e. as $R : \mathcal{M}(\mathcal{Z}) \to \mathbb{R}$.

for any measurable $f : \mathcal{X} \to \mathcal{Y}$:

$$R(f) = \mathbb{E}_\pi[\|Y - f(X)\|_{\mathcal{Y}}^2] = R_* + \mathbb{E}_\pi[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2]$$

*where $f_*(x) := \mathbb{E}_\pi[Y|X = x]$ and R_* is the Bayes risk of the problem ($R_* = \min_{f\in\mathcal{M}(\mathcal{X},\mathcal{Y})} R(f)$). Furthermore, $f^* \in \mathcal{M}(\mathcal{X},\mathcal{Y})$ is the **unique** ($\pi|_{\mathcal{X}}$-a.e.) minimizer of R over $\mathcal{M}(\mathcal{X},\mathcal{Y})$.*

Proof. We include the proof (which is standard) in section C.2 for completion. □

As noted in Mei et al. [58], this decomposition allows (under suitable universality conditions) for a *dimensionless approximation bound* (coming from Barron [6]) given by:

$$\inf_\theta R(\Phi_\theta^N) \leqslant R_* + \frac{1}{N}\left(2r\int \|\omega\|_2|F^*(\omega)|d\omega\right)^2$$

where F^* is the Fourier transform of f^*, and $r = \sup_{x\in\text{supp}(\pi|_{\mathcal{X}})} \|x\|_2$. In other words, when *universality conditions hold* (such as those of theorem 1), we could (at least theoretically) achieve an arbitrarily small *population risk* (up to R_*). We state this result in the context of *universality of shallow NNs* as defined in this work:

Lemma 2 *Consider the quadratic loss and let $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, s.t. $\pi|_{\mathcal{Y}}$ has finite second order moment. Let assumption 1 hold (in particular, $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is universal by Theorem 1), then:*

$$\inf_{f\in\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))} R(f) = \inf_{\gamma\in\mathcal{M}^S(\mathcal{Z}))} R(\gamma) = R_*$$

Now, if this minimum is attained at some $\mu^* \in \mathcal{M}^S(\mathcal{Z})$, then

$$\forall x \in \mathcal{X}\ \pi_{\mathcal{X}}\text{-a.e.}, \langle\sigma_*(x,\cdot),\mu^*\rangle = f^*(x) = \mathbb{E}_\pi[Y|X = x]$$

*i.e. the optimal model can be **realized** in $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$*

Remark

1. In what will follow, the analysis of the functional $R : \mathcal{M}^S(\mathcal{Z}) \to \mathbb{R}$ will be restricted exclusively to probability measures over $\mathcal{Z}$ (i.e. $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$). This can be done WLOG (as noted in Chizat and Bach [16] in what they refer to as the *partial* 1*-homogeneous case*) by considering $\tilde{\mathcal{Z}} = \mathbb{R}\times\mathcal{Z}$ and $\tilde{\sigma_*} : \mathcal{X}\times\tilde{\mathcal{Z}} \to \mathcal{Y}$ defined by: $\tilde{\sigma_*}(x,(c,\theta)) = c\sigma_*(x,\theta)\ \forall x \in \mathcal{X},\ \forall c \in \mathbb{R},\ \forall\theta \in \mathcal{Z}$. Possible *restrictions* will appear at instances where, for instance, we might assume $\mathcal{Z}$ to be compact (as in assumption 1); however this won't be the standard in the rest of our work. In any case, one shall notice that:

   $$\inf_{\gamma\in\mathcal{M}^S(\mathcal{Z})} R(\gamma) = \inf_{\mu\in\mathcal{P}(\tilde{\mathcal{Z}})} \tilde{R}(\mu)$$

   We will make the distinction in the relevant cases (as is the case here), but we'll drop the $\tilde{(\cdot)}$ in upcoming sections.

2. The fact that the value of the **infimum** $\inf_{\gamma\in\mathcal{M}^S(\mathcal{Z})} R(\gamma)$ might be attained is not at all trivial: despite the fact that the *lower semicontinuity* of R is more or less well known (we'll check it out later), checking some sort of *coercivity* condition for R is quite tricky, and it might require for restrictive assumptions.

In particular, if any *infimizing* sequence $(\gamma_n)_{n\in\mathbb{N}}$ is to be *relatively compact* in $\mathcal{M}^S(\mathcal{Z})$, it must satisfy that $\exists M > 0,\ \exists n_0 \in \mathbb{N},\ \forall n \geqslant n_0, |\gamma_n|_{TV} \leqslant M$. However, *bounding* the sequence in this manner would make it impossible for $[x \mapsto \langle \sigma_*(x,\cdot), \gamma_n\rangle]$ (with bounded σ_*, which is often assumed) to arbitrarily *approximate* f^* in L^2, as f^* might be *unbounded* in principle.

We suggest that for a compactly supported $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (which makes f^* essentially bounded), a potential result might follow by considering $\mathcal{F}_{\sigma_*}(\mathcal{M}^S_M(\mathcal{Z}))$, the space of signed measures over $\mathcal{Z}$ with total variation bounded by some constant $M > 0$ (which is a compact space, seen as a closed subset of $\mathcal{P}([-M, M] \times \mathcal{Z})$)[8]. We were unable to prove this for the time being, and will leave it to future work.

The even more difficult question of whether assuming *universality* (under potentially more general conditions than Assumption 1) allows for this infimum to be attained (for potentially unbounded σ_*); is also still open and to be developed in future work.

3. If we assume $\pi_{\mathcal{X}}$ to be compactly supported, by corollary 1, we can adapt the proof to ensure that the infimum is taken over the shallow neural networks of *finite width*. i.e.:

$$\inf_{\gamma\in\mathcal{M}^S(\mathcal{Z}))} R(\gamma) = \inf_{f\in\mathcal{N}_{\sigma_*}(\mathcal{Z})} R(f) = R_*$$

4. One could reasonable wonder whether some kind of *converse statement* holds true. That is, if $\forall\pi \in \mathcal{P}_2(\mathcal{X}\times\mathcal{Y})$ there exists $\mu \in \mathcal{M}^S(\mathcal{Z})$ such that $\forall x \in \mathcal{X}\ \pi_{\mathcal{X}}$-a.e., $\langle\sigma_*(x,\cdot), \mu^*\rangle = f^*(x) = \mathbb{E}_\pi[Y|X = x]$; then, the class $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is universal. We have tried to tackle such question without significant results; so we will leave its exploration to future work.

As told in the previous observation, from this point onward, we will drift from the *universality* analysis, and turn our focus into the properties of the *convexified* optimization problem over the space of probability measures: $\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) := \mathbb{E}_\pi[\ell(\langle\sigma_*(X;\cdot), \mu\rangle, Y)]$. For this, the context of Optimal Transport Theory (particularly, the idea of Wasserstein Spaces) will be of key relevance. A good review of the general topic may be found in Santambrogio [73], Villani [88] or Ambrosio et al. [1]. We do however include some essential definitions and properties that shall be useful to our work.

2.2.4 Wasserstein Spaces

If $\mathcal{P}(\mathcal{Z})$ is the space of probability measures over a separable Hilbert space $\mathcal{Z}$ (with norm $\|\cdot\|$; we often consider $\mathcal{Z} = \mathbb{R}^D$), let $p \geqslant 1$, and define the space of probability measures with finite p-th moment as:

$$\mathcal{P}_p(\mathcal{Z}) := \left\{\mu \in \mathcal{P}(\mathcal{Z})\ :\ \int_{\mathcal{Z}} \|\theta\|^p \mu(d\theta) < +\infty\right\}$$

This space can be endowed with the *Wasserstein metric*, defined $\forall\mu,\nu \in \mathcal{P}_p(\mathcal{Z})$ as:

$$W_p(\mu,\nu) := \left[\inf_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_\gamma[\|X - Y\|^p]\right]^{\frac{1}{p}}$$

[8] Taking from the previous observation, it is equivalent to assuming $\tilde{\mathcal{Z}}$ to be compact and optimizing $\tilde{R}$ over the compact space $\mathcal{P}(\tilde{\mathcal{Z}})$.

Where $\Pi(\mu, \nu)$ is the space of *couplings between* μ *and* ν. i.e. $\Pi(\mu, \nu) := \{\mathbb{P} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z}) : \mathbb{P}_X = \mu,\ \mathbb{P}_Y = \nu\}$. It is a known result that the infimum involved in the definition of $W_p(\mu, \nu)$ is always attained at some (possibly non-unique) $\gamma \in \Pi(\mu, \nu)$ (see Villani [87], Chapter 1). In $\mathcal{Z} = \mathbb{R}^D$, whenever at least **one** of the measures involved (e.g. $\mu \in \mathcal{P}(\mathcal{Z})$) gives zero mass to Borel sets of Hausdorff dimension (at most) $D - 1$ (e.g. if $\mu \lll \lambda_{\mathbb{R}^D}$), then by **Brénier's Theorem** we know that the optimal coupling is **unique** and given by $(\mathrm{Id}_\times \nabla\varphi)\#\mu$, where φ is some convex function such that $\nabla\varphi\#\mu = \nu$ (see Villani [87], McCann [56]).

Beyond this properties, we have that $\forall \mu, \nu \in \mathcal{P}_p(\mathcal{Z})$, $W_1(\mu, \nu) \leqslant W_p(\mu, \nu)$; and that for $p = 1$, the Kantorovich-Rubinstein dual formulation holds (when μ and ν have bounded support as shown in Ambrosio et al. [1]):

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leqslant 1} \left| \int_{\mathcal{Z}} f(\theta)\mu(d\theta) - \int_{\mathcal{Z}} f(\theta)\nu(d\theta) \right|$$

It is a known fact (see Ambrosio et al. [1]) that for *separable and complete* $\mathcal{Z}$, the space $(\mathcal{P}_p(\mathcal{Z}), W_p)$ is a **complete and separable metric space** (i.e. a Polish space). It is also known that for $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_p(\mathcal{Z})$ and $\mu \in \mathcal{P}_p(\mathcal{Z})$:

$$W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0 \iff \begin{cases} \mu_n \xrightarrow[n\to\infty]{} \mu \\ (\mu_n)_n \textit{ has uniformly integrable p-moments} \end{cases}$$

Having *uniformly integrable p-moments* corresponds to saying that $\forall n \in \mathbb{N}$ it holds that $\lim_{r\to\infty} \int_{\mathcal{Z}\setminus B_r(0)} \|\theta\|^p d\mu_n(\theta) = 0$. In particular, when $\mathcal{Z}$ is a (possibly infinite-dimensional) separable Hilbert Space, this condition simplifies to:

$$W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0 \iff \begin{cases} \mu_n \xrightarrow[n\to\infty]{} \mu \\ \lim_{n\to\infty} \int_{\mathcal{Z}} \|\theta\|^p d\mu_n(\theta) = \int_{\mathcal{Z}} \|\theta\|^p d\mu(\theta) \end{cases}$$

Which is also equivalent to:

$$W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0 \iff \forall f \in C(\mathcal{Z}, \mathbb{R}) \textit{ with p-growth, } \int_{\mathcal{Z}} f d\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{Z}} f d\mu$$

Remark Notice that the space $\mathcal{P}_p(\mathcal{Z})$ is locally compact **if and only if** $\mathcal{Z}$ is compact !

With these elements in place, we can first notice the following *slight generalization* of the characterization of convergence in $(\mathcal{P}_p(\mathcal{Z}), W_p)$:

Lemma 3 *Let* $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_p(\mathcal{Z})$ *and* $\mu \in \mathcal{P}_p(\mathcal{Z})$. *Then:*

$$W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0 \iff \begin{cases} \forall \mathcal{Y} \textit{ (real) separable Hilbert space}, \forall f \in C(\mathcal{Z}, \mathcal{Y}) \textit{ with p-growth,} \\ \qquad \int_{\mathcal{Z}} f d\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{Z}} f d\mu \end{cases}$$

Where p-growth for a function $f : \mathcal{Z} \to \mathcal{Y}$ *is defined as there existing constants* $C, C' > 0$ *such that* $\forall z \in \mathcal{Z}$, $\|f(z)\|_{\mathcal{Y}} \leqslant C + C'\|z\|_{\mathcal{Z}}^p$.

Proof. See Annex C.3. □

From there, a property for the *risk functional* of a learning problem appears immediately:

Proposition 5 *Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be subsets of separable Hilbert Spaces, $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ continuous and positive. Consider $R : \mathcal{M}(\mathcal{Z}) \to \mathbb{R}$ defined as $R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X; \cdot)\mu\rangle, Y)]$, $\forall \mu \in \mathcal{M}(\mathcal{Z})$, and let $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be continuous, with p-growth*[9] *on its second argument ($\forall x \in \mathcal{X}$, $\pi_{\mathcal{X}}$-a.e.) for some $p \in \{0\} \cup [1, \infty)$.*

Then $R : \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ is ***lower semicontinuous*** *(l.s.c.) in the topology of $(\mathcal{P}_p(\mathcal{Z}), W_p)$.*

Proof. The proof follows well known and uses standard arguments, but we include it for completeness in section C.3. □

Further insight into the learning problem can be drawn from understanding the problem as one of Optimal Transport. In particular, interesting properties of NN training might come from the study of **Gradient Flows** in Wasserstein Spaces.

2.2.5 Wasserstein Gradient Flows

This section is loosely based on the necessary elements described in Chizat and Bach [16], Hu et al. [38], Chen et al. [13], Chizat [17]; however for a complete reference on the topic, Carmona and Delarue [12] shall be considered.

Let $\mathcal{Z}$ be an arbitrary separable Hilbert space (results in the literature are actually stated for $\mathcal{Z} = \mathbb{R}^D$, but the aditional generality will be sensibly assumed), and $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a functional over the space of probability measures (in our case, the *population risk*). We consider the following relevant quantities:

Definition 2.3 (Linear Functional Derivative (First Variation)) *For a functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, we can define what we call its linear functional derivative (lfd), as a function:*

$$\frac{\partial R}{\partial \mu} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathbb{R}$$

$$s.t.\ \ \forall \mu, \nu \in \mathcal{P}(\mathcal{Z}), \quad \lim_{h \to 0} \frac{R((1-h)\mu + h\nu) - R(\mu)}{h} = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, \theta) d(\nu - \mu)(\theta)$$

$$\textit{and also (in order to avoid ambiguity on the definition):} \quad \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, \theta) d\mu(\theta) = 0$$

The function $R' : \mu \in \mathcal{P}(\mathcal{Z}) \mapsto \frac{\partial R}{\partial \mu}(\mu, \cdot)$ is also known as the first variation of R at μ.

Definition 2.4 (Intrinsic Derivative) *For a functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ we can also define its intrinsic derivative (or L-differential, as in [12]). Whenever $\frac{\partial R}{\partial \mu} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathbb{R}$ exists and is* ***differentiable*** *on its second argument, the intrinsic derivative of R is defined as:*

$$D_\mu R(\mu, \theta) = \nabla_\theta \left(\frac{\partial R}{\partial \mu}(\mu, \theta) \right)$$

[9] Under the convention that $p = 0$ meaning the function is *bounded*

In some contexts, authors don't give a special name for the *intrinsic derivative*, and just refer to the "*first variation*'s gradient". This comes from the fact that, as stated in Chizat and Bach [16], differentiability of $\frac{\partial R}{\partial \mu}(\mu, \cdot)$ might not be required for the theory to hold: it is enough to be able to define a *subdifferential set.* As a reminder, note that:

Definition 2.5 (Subdifferentials) *Let $f : \mathcal{Z} \to \mathbb{R}$ be any function (potentially non-convex):*

- ***[Sub-gradient]*** *A subgradient of f at $z_0 \in \mathcal{Z}$ is any $p \in \mathcal{Z}$ such that $\forall z \in \mathcal{Z}, \quad f(z) \geqslant f(z_0) + \langle p, z - z_0\rangle + o(z - z_0)$.*
- ***[Subdifferential]*** *For $z \in \mathcal{Z}$, we define the subdifferential of f at z, denoted $\partial f(z)$, as the set of all subgradients of f at z. It is easy to verify that this set is closed and convex.*

i.e. as we'll see in what follows, it will be enough (in some settings) to just consider a *Wasserstein sub-gradient flow*, where it's enough to consider the *subdifferential set of $R'(\mu)$* (which, in this setting, has a particular definition, similar to that of the *usual* subdifferential, but involving the *Wasserstein distance* and transport plans[10]).

To better illustrate the notion of the *linear functional derivative* and the *intrinsic derivative*, consider the following examples:

Example 1. An important example (that will be useful later) is that of the **KL Divergence**. Let $\mu \lll \nu$ and $\frac{d\mu}{d\nu}$ be the corresponding Radon-Nykodym derivative; the KL Divergence between μ and ν is defined as: $D(\mu||\nu) := \int \log(\frac{d\mu}{d\nu}(z))d\mu(z)$. Fixing $\nu \in \mathcal{P}(\mathcal{Z})$ and working with $R(\mu) = D(\mu||\nu)$, we have that (modulo a constant that doesn't depend on z, see [62]):

$$\frac{\partial R}{\partial \mu}(\mu, z) = \log\left(\frac{d\mu}{d\nu}(z)\right) + 1 \quad \textit{and} \quad D_\mu R(\mu, z) = \frac{1}{\frac{d\mu}{d\nu}(z)} \nabla_z \frac{d\mu}{d\nu}(z)$$

2. Whenever $R(\mu) := \int_{\mathcal{Z}} \phi(z)d\mu(z)$ for some *bounded continuously differentiable function* $\phi : \mathcal{Z} \to \mathbb{R}$, it is well known that :

$$\frac{\partial R}{\partial \mu}(\mu, z) = \phi(z) - \int \phi d\mu \quad \textit{and} \quad D_\mu R(\mu, z) = \nabla_z \phi(z)$$

(if ϕ is not differentiable, $\partial R'(\mu) = \partial \phi(\cdot)$)

3. In particular, for the *shallow NN* learning setting, with $\mathcal{Y} \subseteq \mathbb{R}$:

$$\frac{\partial R}{\partial \mu}(\mu, z) = \mathbb{E}_\pi \left[D_1\ell(\langle \sigma_*(X; \cdot), \mu\rangle, Y)\sigma_*(X, z)\right] + (\text{constant not depending on } z)$$

$$D_\mu R(\mu, z) = \mathbb{E}_\pi \left[D_1\ell(\langle \sigma_*(X; \cdot), \mu\rangle, Y)\nabla_z\sigma_*(X, z)\right]$$

[10] The interested reader shall look for a reference in Ambrosio et al. [1]; we won't introduce such a concept since it will have limited usability in our context: we will mostly assume $D_\mu R(\mu, \cdot)$ to be well defined (and in particular, it will *coincide* with the usual subgradient). An extension of our results to the setting of sub-Wasserstein Gradient Flows is definitely of interest for our future work.

This known results can be extended to our setting (where $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ are separable Hilbert spaces), by applying a more *general* argument that passes through Bochner integrals. In particular, as in Chizat and Bach [16], let $\mathcal{H}$ be a Hilbert Space, and consider that $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ can be written as $R(\mu) = L(\langle \Phi, \mu \rangle)$, where $\Phi : \mathcal{Z} \to \mathcal{H}$ is a parametrization of elements in $\mathcal{H}$; $L : \mathcal{H} \to \mathbb{R}$ is some *loss* functional, and the integral $\langle \Phi, \mu \rangle$ is a **Bochner integral** on $\mathcal{H}$. This generalizes our *shallow NN learning* setting, as we might consider the Hilbert Space $\mathcal{H} = L^2(\mathcal{X}, \mathcal{Y}, \pi_{\mathcal{X}})$, $L : \mathcal{H} \to \mathbb{R}$ given by $L(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$ and $\Phi : \mathcal{Z} \to \mathcal{H}$ defined as $\forall \theta \in \mathcal{Z},\ \Phi(\theta) = \sigma_*(\cdot; \theta)$. We recover $R(\mu) := \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y)\right] = L(\langle \Phi, \mu \rangle)$

With this *new* setting in mind, we can prove the following result[11]:

Proposition 6 *Let $\mathcal{H}$ be a separable Hilbert Space and $R(\mu) := L(\langle \Phi, \mu \rangle)$, for some function that's Gateaux-differentiable $L : \mathcal{H} \to \mathbb{R}$ on every direction and of continuous differential; and $\Phi : \mathcal{Z} \to \mathcal{H}$ such that $\forall \mu \in \mathcal{P}(\mathcal{Z}),\ \|\langle \Phi, \mu \rangle\|_{\mathcal{H}} < \infty$.*

Then $\forall \theta \in \mathcal{Z},\ \forall \mu \in \mathcal{P}(\mathcal{Z})$:

$$\begin{aligned} \frac{\partial R}{\partial \mu}(\mu, \theta) &= D_h L(\langle \Phi, \mu \rangle)(\Phi(\theta)) = \langle \nabla_h L(\langle \Phi, \mu \rangle), \Phi(\theta) \rangle_{\mathcal{H}} - C_{R,\mu} \\ D_\mu R(\mu, \theta) &= (D_h L(\langle \Phi, \mu \rangle)(D_\theta \Phi(\theta)))^* = \nabla_\theta \Phi(\theta)(\nabla_h L(\langle \Phi, \mu \rangle)) \end{aligned}$$

Where $C_{R,\mu} := \langle \nabla_h L(\langle \Phi, \mu \rangle), \langle \Phi, \mu \rangle \rangle_{\mathcal{H}}$ is exactly the constant needed to avoid ambiguity in the definition; $(\cdot)^*$ denotes the adjoint operator and, in particular, $\nabla_\theta \Phi(\theta) = (D_\theta \Phi(\theta))^* : \mathcal{H} \to \mathcal{Z}$. When $\mathcal{Z} = \mathbb{R}^D$ this corresponds to the usual definition of the gradient.

Proof. See Annex C.4. □

In particular, for the more general *shallow NN learning* setting we just described (in which the output space might not necessarily be $\mathbb{R}$), we get that:

Corollary 2 *Consider $R(\mu) := \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y)\right]$, which is a risk functional as in proposition 6 when considering:*

- *The Hilbert space: $\mathcal{H} = L^2(\mathcal{X}, \mathcal{Y}, \pi_{\mathcal{X}})$*
- *$L : \mathcal{H} \to \mathbb{R}$ as $L(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$, which is Gateaux-differentiable on every direction in $\mathcal{H}$ if we assume $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ to be continuously differentiable on its first argument, with square-integrable derivative. The differential can be explicitly computed to be (and it is clearly continuous on $\mathcal{H}$):*

  $$D_h L(f)(h) = \mathbb{E}_\pi\left[\langle \nabla_1 \ell\left((f(X), Y\right), h(X) \rangle_{\mathcal{Y}}\right]$$

- *$\Phi : \mathcal{Z} \to \mathcal{H}$ defined as $\forall \theta \in \mathcal{Z},\ \Phi(\theta) = \sigma_*(\cdot; \theta)$, which satisfies $\forall \mu \in \mathcal{P}(\mathcal{Z}),\ \|\langle \Phi, \mu \rangle\|_{\mathcal{H}} < \infty$ under the assumption of σ_* being **bounded** and continuous.*

[11] Recalling that by *Riesz Representation Theorem*, any continuous linear functional $f : \mathcal{H} \to \mathbb{R}$ can be represented by a unique *vector* $h_f \in \mathcal{H}$ such that $\forall x \in \mathcal{H},\ f(x) = \langle h_f, x \rangle_{\mathcal{H}}$. In particular, if f is differentiable, $\forall x \in \mathcal{H},\ D_x f(x) : \mathcal{H} \to \mathbb{R}$ is a continuous linear functional and can be thus be represented by the **gradient** vector: $\nabla_x f(x) := h_{D_x f(x)}$, such that $\forall h \in \mathcal{H},\ D_x f(x)(h) = \langle \nabla_x f(x), h \rangle_{\mathcal{H}}$

$$\frac{\partial R}{\partial \mu}(\mu, \theta) = \mathbb{E}_\pi \left[\langle \nabla_1 \ell \left(\langle \sigma_*(X; \cdot), \mu \rangle, Y \right), \sigma_*(X; \theta) \rangle_{\mathcal{Y}} \right] + (\textit{constant not depending on } z)$$
$$D_\mu R(\mu, \theta) = \mathbb{E}_\pi \left[\nabla_\theta \sigma_*(X; \theta) . \nabla_1 \ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y) \right]$$

Proof. Direct from proposition 6. In any case, the exact same proof used for proposition 6 works under the stated hypothesis. □

The notion of the *linear functional derivative* and the *intrinsic derivative* are useful for defining the notion of WGF in *simple terms* (as we'll see right afterwards). They are also key for exploiting the **convexity** of any functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$:

Definition 2.6 *We say that a functional $R : \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ is of class $\mathcal{C}^1$ if $\frac{\partial R}{\partial \mu}(\mu, \cdot)$ is well defined and bounded for every $\mu \in \mathcal{P}_p(\mathcal{Z})$, and the function $(\mu, z) \in \mathcal{P}_p(\mathcal{Z}) \times \mathcal{Z} \mapsto \frac{\partial R}{\partial \mu}(\mu, z)$ is* ***continuous****.*

Remark Notice that, for bounded σ_* and continuously differentiable ℓ with *square-integrable* derivative, the example of corollary 2 is of class $\mathcal{C}^1$.

Lemma 4 (as in Hu et al. [38], Chizat [17]) *Assume that $R : \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ is* ***convex*** *and of class $\mathcal{C}^1$. Then, for any $\mu, \mu' \in P_p(\mathcal{Z})$, we have:*

$$R(\mu') - R(\mu) \geqslant \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, z) d(\mu' - \mu)(z)$$

Proof. The proof is directly taken from Hu et al. [38], with a few minor details *filled in* for completeness. It shall be found in Annex C.4. □

With these notions in mind, we can define what's known in the literature as a Wasserstein (sub-)Gradient Flow (we take the definition from Chizat and Bach [16], but more depth might be found in Ambrosio et al. [1]). We will (for the moment) not require for $R'(\mu) : \mathcal{Z} \to \mathbb{R}$ to be differentiable, and only its *subdifferential* set $\partial R'(\mu)$ (in this context, known as the *Wasserstein Subdifferential of R*) will be required.

Definition 2.7 *[**Wasserstein (sub-)Gradient Flow**] Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a functional for which $\forall \mu \in \mathcal{P}_2(\mathcal{Z})$, $R'(\mu) = \frac{\partial R}{\partial \mu}(\mu, \cdot)$ is defined and has a defined subdifferential. We define a* ***Wasserstein Gradient Flow (WGF) for*** *R as any absolutely continuous trajectory $(\mu_t)_{t \in [0,T[}$ in $\mathcal{P}_2(\mathcal{Z})$ that satisfies, distributionally on $[0, T[\times \mathcal{Z}$:*

$$\partial_t \mu_t = -\operatorname{div}(v_t \mu_t) \qquad \textit{where} \qquad v_t \in -\partial R'(\mu_t) \ \forall t \geqslant 0 \ \textit{a.e.}$$

The first equation represents a mass conservation (continuity) equation, and the second equation implies that the velocity follows the direction of the subgradient. Figure 2.3 provides a pictorial representation of the idea behind this definition.

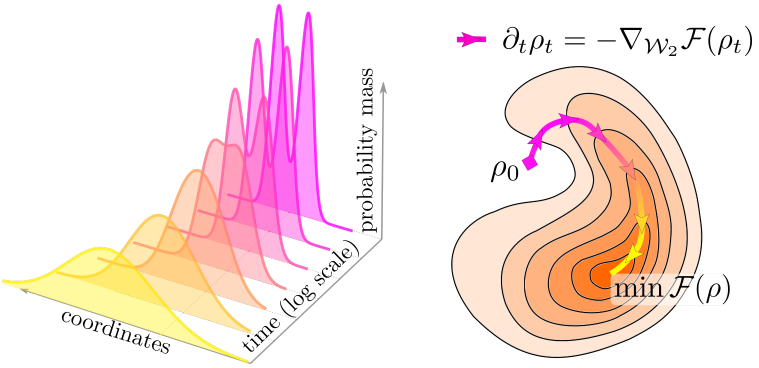


Figure 2.3: Pictorial representation of a Wasserstein gradient flow; taken from the GitHub repository associated to [60]

Whenever R is convex and our intrinsic derivative is properly defined, the ***WGF*** *dynamics is written as*[12]*:*

$$\partial_t \mu_t = \varsigma(t) \operatorname{div} \left(D_\mu R(\mu_t, \cdot) \mu_t \right)$$

where $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ is a scalar that regulates the chosen velocity vector.

Remark Notice that this notion *generalizes* that of a *particle gradient Flow*; as any particle gradient flow $(U(t))_{t \geqslant 0} \in (\mathcal{Z}^m)^{\mathbb{R}_+}$ can be seen as a *Wasserstein Gradient Flow* by considering the empirical measures $\mu_{t,m} = \frac{1}{m} \sum_{i=1}^m \delta_{U_i(t)}$

Now, Chizat and Bach [16] prove that such a *Wasserstein (sub)-Gradient Flow* admits a unique solution, as stated by the following result (the relevant assumptions shall be found in chapter D).

Proposition 7 (**Existence and uniqueness**) *Under assumption 7 and an initial condition $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$ such that $\mu_0(Q_{r_0}) = 1$ for some $Q_{r_0} \subseteq \mathcal{Z}$. Then, there exists a unique $(\mu_t)_{t \geqslant 0}$ WGF for R starting from μ_0, that satisfies the equation for the velocity field:*

$$v_t(u) = \tilde{v}_t(u) - \operatorname{proj}_{\partial V(u)} \left(\tilde{v}_t(u) \right); \quad \tilde{v}_t(u) = - \left[\left\langle R' \left(\int \Phi \mathrm{d}\mu_t \right), \partial_j \Phi(u) \right\rangle \right]_{j=1}^d.$$

In particular, when R has an intrinsic derivative $D_\mu R$, the velocity field corresponds precisely to $D_\mu R(\mu_t, \cdot)$.

As a WGF *follows the negative gradient* of our functional, it is intuitively expected that the *dynamics* of μ_t (the solution of the WGF) will *seek to minimize* R. Unfortunately, even when the functional R is convex, the **stationary points of the dynamics do not necessarily correspond to global minima of** R. For such a thing to *naturally* happen, it is necessary to impose some *regularization* on the functional R (e.g. by adding an *entropy term*). We will further develop this idea in the following sections.

As we'll see, the *training dynamics* of a shallow NN might be seen as a **Wasserstein Gradient Flow** under the right *scaling* limit. To wrap around this point, the following sections describes how the *training* of a NN takes place in practice.

[12] Notice that $v_t = -\varsigma(t) D_\mu R(\mu_t, \cdot) = -\varsigma(t) \nabla_\theta \frac{\partial R}{\partial \mu}(\mu_t, \cdot)$ represents a *velocity vector* that precisely resides in $-\partial R'(\mu_t)$.

2.2.6 Stochastic Gradient Descent Dynamics

Recall the setting of the learning problem, in which we're trying to minimize a functional $R(\mu) = \mathbb{E}_\pi[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)]$ using *shalow NNs* parametrized by $\mathcal{Z}^N$, with $N \in \mathbb{N}$. For notational convenience, we will introduce the following function $L_{x,y} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, for $x \in \mathcal{X},\ y \in \mathcal{Y}$, defined as:

$$L_{x,y}(\mu) := \ell\left(\langle \sigma_*(x,\cdot), \mu\rangle, y\right)$$

This allows us to rewrite our functional as: $R(\mu) := \mathbb{E}_\pi\left[L_{X,Y}(\mu)\right]$. Also, seeing $L_{x,y}$ as a functional from $\mathcal{P}(\mathcal{Z})$ to $\mathbb{R}$ (for each $x \in \mathcal{X},\ y \in \mathcal{Y}$), we can use proposition 6 to write, $\forall x \in \mathcal{X},\ y \in \mathcal{Y}$:

$$\frac{\partial L_{x,y}}{\partial \mu}(\mu,\theta) = \langle \nabla_1 \ell(\langle \sigma_*(x,\cdot), \mu\rangle, y), \sigma_*(x,\theta)\rangle_{\mathcal{Y}}$$

$$D_\mu L_{x,y}(\mu,\theta) = \nabla_\theta \sigma_*(x,\theta) \cdot \nabla_1 \ell(\langle \sigma_*(x,\cdot), \mu\rangle, y)$$

In particular, from corollary 2, we can see that (at least formally): $D_\mu R(\theta,\mu) = \mathbb{E}_\pi\left[D_\mu L_{X,Y}(\theta,\mu)\right]$.

Now, in the ideal case, if we perfectly knew the data distribution π, we could try to achieve the optimization by doing regular **gradient descent iterations** (see Suzuki et al. [82]):

- First, initializing $\forall i \in \{1,\dots,N\},\ \theta_i^0 \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$
- On every iteration $k \in \mathbb{N}$, defining $\forall i \in \{1,\dots,N\}$:

  $$\theta_i^{k+1} = \theta_i^k - s_k^N\, D_\mu R(\nu_\theta^N, \theta_i^k)$$

  Where $(s_k^N)_{k\in\mathbb{N}}$ is a fixed *step-size* (also commonly known as *learning rate*), and we employ $D_\mu R(\nu_\theta^N, \theta_i^k) = \mathbb{E}_\pi[\partial_1 \ell(\Phi_\theta^N(X), Y)\nabla_{\theta_i}(\sigma_*(X;\theta_i^k))]$, the *exact gradient* of the function $\theta \mapsto R(\Phi_\theta^N)$, to update the parameter's values.

However, in practice the law π is generally **unknown**, and only an i.i.d. data sample $\{(X_k, Y_k)\}_{k\in\mathbb{N}}$ (distributed following π) is available. The method for *training the Neural Network* is thus **stochastic gradient descent (SGD)**. Unable to know the exact value of R, we're forced to *approximate it from our data.* For instance, let $N \in \mathbb{N}$ be fixed and consider $\theta \in \mathcal{Z}^N$ our shallow NN's parameter; with the first $B \in \mathbb{N}$ samples we approximate:

$$R(\Phi_\theta^N) \approx \hat{R}^{N,B}(\theta; (X_k, Y_k)_{k=1}^B) = \frac{1}{B}\sum_{k=1}^B \ell(\Phi_\theta^N(X_k), Y_k)$$

The literature usually just considers $B = 1$, and performs the following *training loop*:

- First, consider $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$ and initialize $\forall i \in \{1,\dots,N\},\ \theta_i^0 \overset{i.i.d.}{\sim} \mu_0$.
- On every iteration $k \in \mathbb{N}$, define $\forall i \in \{1,\dots,N\}$:

  $$\theta_i^{k+1} = \theta_i^k - s_k^N\, \partial_1 \ell(\Phi_{\theta^k}^N(X_k), Y_k)\nabla_{\theta_i}(\sigma_*(X_k;\theta_i^k)) \tag{2.1}$$

  where s_k^N is the *learning rate.* Notice that the *exact* gradient from before has been replaced by a *stochastic approximation* (as it corresponds to the gradient of the *empirical loss* evaluated on a single sample). Using our latest notation, we write the iteration as:

  $$\theta_i^{k+1} = \theta_i^k - s_k^N\, D_\mu L_{X_k,Y_k}(\nu_{\theta^k}^N, \theta_i^k)$$

Over the years, as NN models have grown in popularity, many variations of this simple *training loop* have been considered. Most remarkably, in the literature of *mean field limits of shallow NNs*, authors have often considered some of the following variants:

1. **Constant Learning Rate** (see [57, 58, 78, 81, 9] among many others): We consider $\forall k \in \mathbb{N}$, $s_k^N = \frac{\alpha}{N}$ for $\alpha > 0$ a fixed *learning rate* (LR), which is scaled by the number of hidden units.

2. **Regularized Risk** (see [16, 57, 17, 13, 63, 82, 38, 9] among many others): We consider some **regularization term** on the *population risk* we're trying to minimize. We consider the τ-regularized population risk (with $\tau > 0$) as:

$$R^\tau(\mu) = R(\mu) + \tau V(\mu)$$

where V is a **regularizer** which will be assumed to be of the form $V(\mu) = \int_{\mathcal{Z}} r d\mu$ for some potential function $r : \mathcal{Z} \to \mathbb{R}$. Notable examples include the **relative entropy** and the measure's *second moment* $V(\mu) = \int_{\mathcal{Z}} \|\theta\|^2 d\mu(\theta)$. This leads to a *modified update rule* of the form:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\partial_1 \ell(\Phi_{\theta^k}^N(X_k), Y_k)\nabla_{\theta_i}(\sigma_*(X_k;\theta_i^k)) + \tau \nabla_{\theta_i} r(\theta_i^k)\right)$$

Or, in our *neater* notation (defining $L_{x,y}^\tau : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ naturally[13]) :

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(D_\mu L_{X_k,Y_k}^\tau(\nu_{\theta^k}^N, \theta_i^k)\right)$$

3. **Noisy SGD** (see [57, 58, 82, 9] among many others): In order to ensure *convergence* of the training dynamic, noise will have to be introduced in training; this usually takes the form of an i.i.d. sequence $\xi_i^k \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{Id}_{\mathcal{Z}})$ such that the training loop becomes (with the *noise regularization parameter* $\beta > 0$):

$$\theta_i^{k+1} = \theta_i^k - s_k^N \, \partial_1 \ell(\Phi_{\theta^k}^N(X_k), Y_k)\nabla_{\theta_i}(\sigma_*(X_k;\theta_i^k)) + \sqrt{2\beta s_k^N}\xi_i^k$$

Under the **constant LR** regime, with $s_k^N = \frac{\alpha}{N}$, the noise term significantly modifies the training dynamics (even asymptotically, as it introduces a *diffusion* term on the mean-field distributional dynamics). If the noise term was to be further divided by N^δ (with $\delta > 0$) its influence would effectively *vanish* asymptotically; this particular setting corresponds to the so-called **weak-noise** regime.

4. **Random MiniBatch** (see [24]): For every $k \in \mathbb{N}$ consider B_k a random element of $\mathbb{N}^*$, and the corresponding *batch* of data $\{(X_1^k, Y_1^k), \dots, (X_{N_k}^k, Y_{N_k}^k)\}$ (where $(X_j^k, Y_j^k)_{j\in\mathbb{N}}$ is an i.i.d. sample of data considered at each iteration, so that the batches are *independent* over different iterations). This *batch* of data is used to better approximate the population risk's gradient; leading to the following loop:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \frac{1}{B_k}\sum_{j=1}^{B_k} \partial_1 \ell(\Phi_{\theta^k}^N(X_j^k), Y_j^k)\nabla_{\theta_i}(\sigma_*(X_j^k;\theta_i^k))$$

The most common particular case is to take B_k to be constant and equal to $B \in \mathbb{N}^*$.

[13] i.e. we define it as $L_{x,y}^\tau(\mu) = L_{x,y}(\mu) + \tau V(\mu)$, $\forall \mu \in \mathcal{P}(\mathcal{Z})$ so that $R^\tau(\mu) = \mathbb{E}_\pi[L_{X,Y}^\tau(\mu)]$

5. **Different Learning Rates** (see [9] and [57]): Consider the parameters $\zeta \in [0,1)$, $\upsilon \in [0,1]$ and $\alpha > 0$; and define $\alpha^N_{\zeta,\upsilon} = (\alpha N^{(\upsilon-1)})^{1/(1-\zeta)}$. Under this regime, we consider the **learning rate** to be:
$$\forall k \in \mathbb{N},\ s_k^N = \frac{\alpha N^{\upsilon-1}}{\left(k + \frac{1}{\alpha^N_{\zeta,\upsilon}}\right)^{\zeta}}$$
As noted in Bortoli et al. [9], such a learning rate appears *naturally* when studying the dynamic (most notably, $\alpha^N_{\zeta,\upsilon}$). Most remarkably, depending on the value of υ, the limiting mean-field dynamic is different (in particular, for $0 \leqslant \upsilon < 1$ it stays *the same* as usual; but for $\upsilon = 1$ and aditional *diffusion term* appears in the limiting PDE). In general, for the theory of Mei et al. [57] to work, it suffices to assume that the LR can be written as $s_k^N = \varepsilon\varsigma(k\varepsilon)$ for some fixed $\varepsilon > 0$ and a regular function $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$.

Considering all of these elements at once, we shall study the most *general* version of SGD (noting that it suffices to make the corresponding parameters *trivial* in order to recover the original training loop); i.e. $\forall k \in \mathbb{N}$:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\frac{1}{B_k} \sum_{j=1}^{B_k} \partial_1 \ell(\Phi^N_{\theta^k}(X_j^k), Y_j^k) \nabla_{\theta_i}(\sigma_*(X_j^k;\theta_i^k)) + \tau \nabla_{\theta_i} r(\theta_i^k) \right) + \sqrt{2\beta s_k^N}\xi_i^k \quad (2.2)$$

With $\tau, \beta, \varepsilon > 0$, B_k random in $\mathbb{N}^*$, $\xi_i^k \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{Id}_{\mathcal{Z}})$ and $s_k^N = \varepsilon\varsigma(k\varepsilon)$ with $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ a sufficiently regular function. In our *neater* notation, this can be stated (simply) as:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\frac{1}{B_k} \sum_{j=1}^{B_k} D_\mu L^\tau_{X_j^k, Y_j^k}(\nu^N_{\theta^k}, \theta_i^k) \right) + \sqrt{2\beta s_k^N}\xi_i^k \quad (2.3)$$

As mentioned earlier, understanding how the parameter $\theta^k := (\theta_i^k)_{i=1}^N \in \mathcal{Z}^N$ evolves by following equation (2.2) directly, can be an exceedingly challenging problem to solve (specially in the case of NNs). Fortunately, under the lens of *shallow NNs*, recasting the problem as one of minimization over $\mathcal{P}(\mathcal{Z})$ and taking the limit (under the right scaling) as $N \to \infty$, allows for stronger guarantees (as the limiting object has nicer behaviour). For instance, as previously mentioned, we shall see how this SGD training dynamics can be understood as a Wasserstein Gradient Flow in $(\mathcal{P}_2(\mathcal{Z}), W_2)$.

2.2.7 SGD as a WGF: a Law of Large Numbers

Recall that, for a given $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$, the associated network Φ^N_θ can be seen as an *integral* against the **empirical measure** associated with θ, $\nu^N_\theta = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$. More specifically:
$$\forall x \in \mathcal{X},\ \Phi^N_\theta(x) = \langle \sigma_*(x;\cdot), \nu^N_\theta \rangle$$
And thus, our highly non-convex (and hard to solve) optimization problem:

$$\inf_{\theta \in \mathcal{Z}^N} R(\Phi^N_\theta) = \mathbb{E}_\pi[\ell(\langle \Phi^N_\theta(X), Y)]$$

is recast as

$$\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\nu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y)]$$

which is (whenever $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is chosen convex on its first argument) a **convex optimization problem** on the space $\mathcal{P}(\mathcal{Z})$.

Furthermore, a classic result (as can be found in Mei et al. [57], or originally in Barron [6]) directly relates the optimum of the original problem against that of the *convexified* problem over $\mathcal{P}(\mathcal{Z})$. A general version is established in Hu et al. [38]:

Theorem 2 (from Hu et al. [38]) *Assume that the* $\frac{\delta^2 R}{\delta \mu^2}$ *exists*[14]*, is jointly continuous in both variables, and there is* $L > 0$ *such that for any random variables* η_1, η_2 *such that* $\mathbb{E}[|\eta_i|^2] < \infty$, $i = 1, 2$*, it holds that*

$$\mathbb{E}\left[\sup_{\nu \in \mathcal{P}_2(\mathcal{Z})} \left| \frac{\delta R}{\delta \mu}(\nu, \eta_1) \right| \right] + \mathbb{E}\left[\sup_{\nu \in \mathcal{P}_2(\mathcal{Z})} \left| \frac{\delta^2 R}{\delta \mu^2}(\nu, \eta_1, \eta_2) \right| \right] \leqslant L.$$

Then:

$$\left| \inf_{\theta \in \mathcal{Z}^N} R\left(\nu_\theta^N\right) - \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu) \right| \leqslant \frac{2L}{N}$$

Proof. This theorem is entirely proven in [38], but they assume that the infimum is attained. Luckily, no such assumption is needed, and for completeness we include the variant of the proof in section C.5. □

A quite similar result is stated in Mei et al. [57] (Proposition 1) considering the quadratic loss and assumptions directly related to that case; however, Hu et al. [38] state that theorem 2 generalizes the result from [57]. Now, recall that the general training dynamics is driven by equation (2.2), i.e. $\forall k \in \mathbb{N}$:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\frac{1}{B_k} \sum_{j=1}^{B_k} \partial_1 \ell(\Phi_{\theta^k}^N(X_j^k), Y_j^k) \nabla_{\theta_i}(\sigma_*(X_j^k; \theta_i^k)) + \tau \nabla_{\theta_i} r(\theta_i^k) \right) + \sqrt{2\beta s_k^N} \xi_i^k \quad (2.2)$$

As parameters influence the *population risk* only via their *empirical measure*, it will be interesting to see how it evolves along succesive SGD iterations. Slightly abusing notation, denote the empirical measure of the parameters after each SGD iteration as: $\nu_k^N := \nu_{\theta^k}^N = \frac{1}{N} \sum_{i=1}^N \delta_{\theta_i^k}$.

One of the main results of the Mean Field Theory of shallow NNs is the **propagation of chaos** of the *particle system*. This means that, as individual NN parameters (particles) are exchangeable (and thus, characterized by their empirical measure), each of their *trajectories* following SGD should eventually (as $N \to \infty$) become *statistically independent*, and their law shall tend to that of a fixed limiting process (corresponding to the WGF of the population risk).

[14] Where $\frac{\delta^2 R}{\delta \mu^2} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is defined as the lfd of $\frac{\partial R}{\partial \mu}$ seen as a function over $\mathcal{P}(\mathcal{Z})$.

To get to this *limiting process*, some authors (such as Bortoli et al. [9]) pass through an *intermediate step*, in which we could understand the SGD iterations as part of a *continuous gradient flow* in the finite space $\mathcal{Z}^N$. For this, recall our notation $L^{\tau}_{x,y}(\mu) := \ell\left(\langle \sigma_*(x,\cdot), \mu\rangle, y\right) + \int_{\mathcal{Z}} r d\mu$ such that the SGD iteration can be written as:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\frac{1}{B_k} \sum_{j=1}^{B_k} D_\mu L^{\tau}_{X_j^k, Y_j^k}(\nu^N_{\theta^k}, \theta_i^k) \right) + \sqrt{2\beta s_k^N} \xi_i^k \tag{2.3}$$

Define $\Sigma(\mu,\theta) := \mathbb{E}_\pi \left[\left(D_\mu L^{\tau}_{X,Y}(\mu,\theta) - D_\mu R^{\tau}(\mu,\theta) \right) \otimes \left(D_\mu L^{\tau}_{X,Y}(\mu,\theta) - D_\mu R^{\tau}(\mu,\theta) \right) \right]$, where $\otimes$ represents the *outer product* (or generally, the *tensor product*) between vectors in $\mathcal{Z}$; this is the *covariance matrix* of the vector $D_\mu L^{\tau}_{X,Y}(\mu,\theta)$ (with $(X,Y) \sim \pi$). Now, we can consider the following *continuous time SGD iteration* for a set of parameters $(\theta^t)_{t\geqslant 0} := ((\theta_i^t)_{i=1}^N)_{t\geqslant 0}$, given by:

$$d\theta_i^t = \varsigma(t) \left[D_\mu R^{\tau}(\nu_t^N, \theta_i^t) dt + \sqrt{\frac{\varepsilon}{B}} \sqrt{\Sigma}(\theta_i^t, \nu_t^N) d\tilde{B}_t^i + \sqrt{2\beta} dB_t^i \right] \tag{2.4}$$

where $((B_t^i)_{t\geqslant 0})_{i\in\mathbb{N}}$ and $((\tilde{B}_t^i)_{t\geqslant 0})_{i\in\mathbb{N}}$ are independent families of independent Brownian motions on $\mathcal{Z}$. The derivation of this dynamic is rather heuristic and follows the ideas of Bortoli et al. [9]; a more formal derivation (as well as some results comparing the continuous time approximation to the original process) shall be sought in Fontaine et al. [33] (in the specific case of the learning rate from point *5.* in section 2.2.6[15]). In many papers, such as Bortoli et al. [9], the properties of this continuous time dynamic are studied as a proxy to the original particle system, but without the issues that appear under *discrete-time* iterations. It is also interesting to note that the *covariance term* dissappears in the mean field limit under some standard conditions[16]. Though some interesting insights can be obtained from studying such a dynamic, we won't dive into it in much detail.

We will now focus on the following limiting **distributional dynamics (DD)**:

$$\partial_t \mu_t = \varsigma(t) \left[\operatorname{div}\left(\left(D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r \right) \mu_t \right) + \beta \Delta \mu_t \right] \tag{2.5}$$

It's not hard to notice that this distributional dynamics correspond to the **Wasserstein Gradient Flow** minimizing the *entropy-regularized* convex functional:

$$R^{\tau.\beta}(\mu) := R(\mu) + \tau \int r d\mu + \beta H_\lambda(\mu)$$

where $H_\nu(\mu) := D(\mu||\nu) = \int \log(\frac{d\mu}{d\nu}(z)) d\mu(z)$ is the *relative entropy* (also known as the KL divergence) between μ and ν (see section 2.2.5)[17] (with $\mu \lll \nu$ and $\frac{d\mu}{d\nu}$ being the corresponding Radon-Nykodym derivative). Notice that by setting $\tau, \beta = 0$ we recover the WGF for R. Whenever $\beta = 0$, this equation is often known to have a (unique) solution *distributionally*; for $\beta > 0$ the solutions to this equation are actually **strong**.

[15] i.e. For $\zeta \in [0,1)$, $\upsilon \in [0,1]$ and $\alpha > 0$; define $\alpha^N_{\zeta,\upsilon} = (\alpha N^{(\upsilon-1)})^{1/(1-\zeta)}$ and consider: $\varepsilon^N_{\zeta,\upsilon,\alpha} = \alpha^N_{\zeta,\upsilon}$, $\varsigma_\zeta(t) = (1+t)^{-\zeta}$, such that $\forall k \in \mathbb{N}$, $s_k^N = \varepsilon^N_{\zeta,\upsilon,\alpha} \varsigma_\zeta(k\varepsilon^N_{\zeta,\upsilon,\alpha})$

[16] As noted in Bortoli et al. [9], for $\upsilon \in [0,1)$, this *covariance term* vanishes in the MF limit; but this doesn't happen in the $\upsilon = 1$ regime, making the MF dynamic behave differently.

[17] While doing this literature review, we noticed that Suzuki et al. [82] writes $R^{\tau,\beta}$ using the *negative entropy* instead of the usual *relative entropy*, which we believe could be a typo. However, as most other works from the literature agree on the use the relative entropy, this choice won't affect our results at all.

Equation (2.5) corresponds to what is known (in the literature) as a Fokker-Planck equation. This is in direct correspondance with what is referred to as a **McKean-Vlasov Equation**; a **nonlinear SDE** that describes the evolution of a *type of parameter* under the training dynamics[18]:

$$dZ_t = \varsigma(t)\left[-\left(D_\mu R(\mu_t, Z_t) + \tau \nabla_\theta r(Z_t)\right)dt + \sqrt{2\beta}dB_t\right] \text{ with } \mu_t = \mathbf{Law}(Z_t) \tag{2.6}$$

Where $(B_t)_{t\geqslant 0}$ is a D-dimensional standard Brownian Motion. This equivalent formulation (See Sznitman [83], Theorem 1.1 for a reference result) shall prove useful to characterize some of the relevant results. Whenever $\beta = 0$, there is no longer a *Langevin* component, and we shall simply refer to the SDE as the Mean Field Dynamics (MFD) or the McKean-Vlasov Equation.

Finally, there's also a way of seeing the DD of equation (2.5) through an ODE of *characteristics* (which is detailed in papers such as Rotskoff and Vanden-Eijnden [72] and Chen et al. [15]). Despite the inherent interest coming from such a description, we won't delve much into its details.

With all of these elements in mind, the standard **Propagation of Chaos** result from the literature (also known as a ***Law of Large Numbers***) is stated as follows:

Theorem 3 (**Propagation of Chaos; sketch**) *Let $s_k^N = \varepsilon\varsigma(k\varepsilon)$ for $\varepsilon > 0$ and $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ a sufficiently regular function. Let $T > 0$ and let $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$ be an initial condition.*

Consider the sequence of parameters, $(\theta^k)_{k\in\mathbb{N}}$, obtained from following the SGD Dynamics (equation (2.2)) starting from μ_0; and let $(\nu_k^N)_{k\in\mathbb{N}}$ be the associated empirical measure. Similarly, consider $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ to be the unique solution of the distributional dynamics (equation (2.5)) starting from μ_0.

Consider the Skorokhod space of càdlàg processes $D_E([0,T])$ (where $E = \mathcal{P}(\mathcal{Z})$). Notice that $\mu := (\mu_t)_{t\geqslant 0} \in D_E([0,T])$ and define $\mu^{N,\varepsilon} := (\nu^N_{[t/\varepsilon]})_{t\geqslant 0} \in D_E([0,T])$

*Under relevant **technical assumptions** involving **regularity** of σ_* and similar others; and under the **right scaling** (of ε with respect to N), we have that:*

$$\mu^{N,\varepsilon} \xRightarrow[\substack{N\to\infty \\ \varepsilon\to 0}]{} \mu$$

*where $\Rightarrow$ denotes weak convergence in $D_E([0,T])$[19]. Under the right conditions, this convergence might be stronger (e.g. in **Wasserstein-2 metric**: $W_2(\mu^{N,\varepsilon}, \mu) \to 0$)*

The previous result vaguely states the idea behind the usual **Propagation of Chaos** result for the training dynamics of shallow NNs. Generally, the proof for such a result essentially involves establishing the relative compactness of the sequence $(\mu^N)_{N\in\mathbb{N}}$ (via *tightness*, using the Prokhorov theorem) and identifying its limit in a *unique* way (independent of the subsequence chosen). More technical insight might be found in Annex D, where some of the

[18] In this context, some authors also call it *the **Mean Field Langevin Dynamics (MFLD)***. We will refer to it indistinctively

[19] Note that weak convergence to a constant implies convergence in probability, so a stronger result holds.

key technical assumptions for these kinds of results are included (notably those from Chizat and Bach [16], Mei et al. [57], Sirignano and Spiliopoulos [78], Descours et al. [24] and Bortoli et al. [9]).

For sake of completeness, we include some of the most frequently found versions of this theorem. In particular, consider a quadratic loss function ($\ell(y, \hat{y}) = \|y - \hat{y}\|_{\mathcal{Y}}^2$), and following a simple calculation (analogue to that of Mei et al. [57] or Rotskoff and Vanden-Eijnden [72]) we may see the population risk $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ can be rewritten as:

$$R(\mu) = R_{\#} + 2\int_{\mathcal{Z}} F(\theta_1)d\mu(\theta_1) + \int_{\mathcal{Z}\times\mathcal{Z}} K(\theta_1, \theta_2)d\mu^2(\theta_1, \theta_2), \quad \forall\mu \in \mathcal{P}(\mathcal{Z})$$

Where we define $F(\theta) = -\mathbb{E}[\langle Y, \sigma_*(X, \theta)\rangle_{\mathcal{Y}}]$, $K(\theta, \theta') = \mathbb{E}[\langle\sigma_*(X, \theta), \sigma_*(X, \theta')\rangle_{\mathcal{Y}}]$, and $R_{\#} = \mathbb{E}[\|Y\|_{\mathcal{Y}}^2]$. This is *quadratic* on μ, making the problem *conceptually simpler*. Also, it can be shown that K defines a **positive definite kernel**[20]. Under this form, we can define what they refer to as the **potential function** $\Psi : \mathcal{Z} \times \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, given by: $\Psi(\theta, \mu) = F(\theta) + \int_{\mathcal{Z}} K(\theta, \theta')d\mu(\theta')$, $\forall\theta \in \mathcal{Z}$, $\forall\mu \in \mathcal{P}(\mathcal{Z})$. This is nothing else than the linear functional derivative (halved), $\Psi(\theta, \mu) = \frac{1}{2}\frac{\partial R}{\partial \mu}(\mu; \theta)$.

With these commonly used elements from the literature, we can state this **Propagation of Chaos** result more precisely (as before, the relevant technical assumptions shall be found in Chapter D):

Theorem 4 ((**Propagation of Chaos**) as in [16, 57, 72, 78, 9, 24]) *Under different settings we have:*

1. *Consider the quadratic loss $\ell(y, \hat{y}) = |y - \hat{y}|^2$; let $\varepsilon_N = \frac{1}{N}$ and $\varsigma \equiv \alpha > 0$ (i.e. $s_k^N = \frac{\alpha}{N}$ is the simple learning rate).*

   (a) ***[Regular SGD]** Let $(\theta^k)_{k\in\mathbb{N}}$ be obtained from following the **simple** SGD Dynamics (equation (2.1)) starting from μ_0; and let $\mu := (\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ be the unique solution[21] of the simple distributional dynamics (equation (2.5) with $\tau = 0, \beta = 0$) starting from μ_0. i.e. the DD given by (the Wasserstein gradient flow for R in $(\mathcal{P}_2(\mathbb{R}^D), \mathcal{W}_2)$):*

   $$\partial_t\mu_t = 2\alpha\,\mathrm{div}_\theta(\mu_t\nabla_\theta\Psi(\theta; \mu_t)) \tag{2.7}$$

   Consider that assumption 8 (1. or 2.) holds.

   (b) ***[Noisy SGD]** Let $(\theta^k)_{k\in\mathbb{N}}$ be obtained from following the **noisy and regularized** SGD Dynamics (equation (2.2) with $B_k \equiv 1$ and $r(\theta) = \frac{1}{2}\|\theta\|^2$) starting from μ_0; and let $\mu := (\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ be the unique solution[22] of the distributional dynamics (equation (2.5)) starting from μ_0. i.e. the DD given by (the Wasserstein gradient flow for $R^{\tau,\beta}$ in $(\mathcal{P}_2(\mathbb{R}^D), \mathcal{W}_2)$):*

   $$\partial_t\mu_t = 2\alpha\left[\mathrm{div}_\theta\left((\nabla_\theta\Psi(\theta; \rho_t) + \tau\nabla_\theta r)\mu_t\right) + \beta\Delta_\theta\mu_t\right]$$

   Let assumption 8 (2.) hold.

[20] Indeed, from the fact that $\forall\gamma \in \mathcal{M}^S(\mathcal{Z})$ $\|\langle\sigma_*, \gamma\rangle\|^2 = \int_{\mathcal{Z}^2} K(\theta, \theta')d\gamma(\theta)d\gamma(\theta')$ (as seen in the proof of proposition 2); the result follows from taking $\gamma = \sum_{i=1}^N c_i\delta_{z_i}$ for arbitrary $(c_i, z_i)_{i=1}^N \in (\mathbb{R} \times \mathcal{Z})^N$.

[21] The equation is satisfied in the weak sense, but whenever μ_0 admits density u_0, then it holds in the **strong** sense).

[22] In this case, the equation is satisfied in the **strong** sense.

(c) *[**MiniBatch SGD**] (as in Descours et al. [24]). Let $(\theta^k)_{k\in\mathbb{N}}$ be obtained from following the **minibatch** SGD Dynamics (equation (2.2) with $\tau = 0$ and the noise term divided by N^δ with $\delta > 0$) starting from μ_0; and let $\mu := (\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ be the unique solution of the corresponding distributional dynamics (equation (2.5)) starting from μ_0 (i.e. equation (2.7), the WGF for R in $(\mathcal{P}_2(\mathbb{R}^D), \mathcal{W}_2)$). Further assume that assumption 8 (3.) holds.*

Assuming the sufficient technical conditions, the rescaled empirical training process $\mu^N := (\nu^N_{\lfloor Nt \rfloor})_{t\geqslant 0} \in D_E([0,T])$ converges weakly (as $N \to \infty$) to μ.

2. *Consider ℓ to be any loss function that's convex (on its first argument). Let $\zeta \in [0,1)$, $\upsilon \in [0,1]$ and $\alpha > 0$; define $\alpha^N_{\zeta,\upsilon} = (\alpha N^{(\upsilon-1)})^{1/(1-\zeta)}$ and consider: $\varepsilon_N = \alpha^N_{\zeta,\upsilon}$, $\varsigma(t) = (1+t)^{-\zeta}$, such that $\forall k \in \mathbb{N}$, $s^N_k = \varepsilon_N \varsigma(k\varepsilon_N)$. Consider the SGD training dynamic with fixed batchsize $B \in \mathbb{N}^*$ starting from μ_0, let $\mu^N := (\mu^N_t)_{t\geqslant 0}$ be the law of the solution to the continuous time N-particle-system dynamic as in equation (2.4) (initialized i.i.d.) and let $\mu := (\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ be the unique solution to the limiting distributional dynamics (equation (2.5) with $\tau = 1, \beta = 0$) starting from μ_0. Under assumption 8 (4.), we have that for any fixed $m \in \mathbb{N}$, we have $\lim_{N\to+\infty} W_2(\mu^{1:m,N}, (\mu)^{\otimes m}) = 0$, where $\mu^{1:m,N}$ is the law of the first m particles in the N-particle-system continuous time dynamic.*

For more references, see Theorem 3 of Mei et al. [57], Theorem 1.2 of Sirignano and Spiliopoulos [78], Theorem 2.6 of Chizat and Bach [16] or Proposition 3.2 Rotskoff and Vanden-Eijnden [72]. Also: Theorem 1 of Bortoli et al. [9], Theorem 3 of Suzuki et al. [82], Theorem 1 of Descours et al. [24] among many others.

Remark This kind of *propagation of chaos* result tells us that, in the asymptotic limit, each *unit* (neuron) in the neural network *loses* its dependence on the rest of the neurons and behaves "independently", following the fixed law given by the process μ (the solution to the DD equation (2.5)). This is what naturally gives it the name of the *Law of Large Numbers* in this setting: asymptotically, the evolution of our parameters evolves as a fixed process (the *mean field* process).

Remark In the work of Mei et al. [57], it is mentioned that when the data has a law that is *invariant under the action of a certain group* (e.g. π invariant under left rotations)[23], the limiting dynamic also *benefits* from the same *symmetry.* In particular, this should allow us to *reduce* the *dimensionality of the problem*, as we might only seek solutions of the DD, μ_t, that are invariant under the same group of symmetries. This assertion is not proven in any way within Mei et al. [57], and it does not rely on any *known* results from the literature; it seems to be more of a *practical indication* for solving the PDE in the limit. The goal of this work is precisely to *understand how the presence of **symmetries of the data under the action of groups** can influence the training of NN models, particularly in the MF limit.*

This classical *propagation of chaos* result can be also understood from the optic of *statistical independence* (from where the name of *propagation of chaos* appears). The last variant described in Theorem 4 focuses on this approach: as N grows, the evolution of a

[23] This notion will be clarified later in the document.

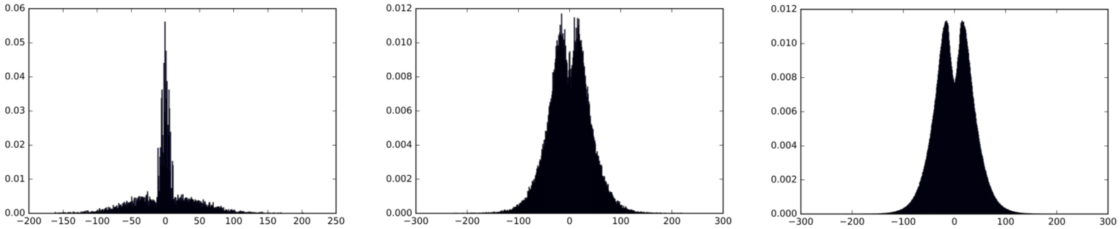


Figure 2.4: Illustration of the *propagation of chaos* result: As N grows large (and the SGD iterations as well), all the network parameters tend to align with a fixed limiting distribution (as with the LLN). Taken from Sirignano and Spiliopoulos [78].

fixed amount of particles m along the *N-particle-system* continuous time dynamic, becomes really close to that of a set of m independent *typical* particles that all follow the *mean field* process.

Analogously, the first variant from Theorem 4 can be stated as follows (see Theorem 1.6 of Sirignano and Spiliopoulos [78]):

Proposition 8 *[Propagation of Chaos] Under the assumptions of Theorem 4, for $T < \infty$ and $t \in [0, T]$, define the law of the first N particles along the rescaled SGD training dynamics as $\rho_t^N := \boldsymbol{Law}(\theta_1^{\lfloor Nt \rfloor}, \ldots, \theta_N^{\lfloor Nt \rfloor})$. Then, the sequence $\{\rho_\cdot^N\}_{N \in \mathbb{N}}$ is μ-chaotic* [24].

Figure 2.4 can be useful to illustrate this idea of *statistical independence.*

As the propagation of chaos result *mimics* the LLN, one can follow the same spirit to establish some kind of *Central Limit Theorem* for the training dynamics. More specifically, as in the usual probabilistic setting, one desires to *approximate* the *training process*' behaviour, for large N, as:

$$\nu_{\lfloor Nt \rfloor}^N \approx \mu_t + \frac{1}{\sqrt{N}} \eta_t$$

where the so-called *fluctuation process* $(\eta_t)_{t \geqslant 0}$ is a Gaussian process with a specific variance-covariance structure. The main results along these lines have been established in papers such as Rotskoff and Vanden-Eijnden [72], Sirignano and Spiliopoulos [81], Chen et al. [15]; the interested reader can find some reference results from the literature describing the CLT result for the (rescaled) training process of shallow NNs in Section D.1.2.

Despite the interest such a result generates, the understanding of this aspect of the training dynamic is still being developped in the literature (notably, Descours et al. [24] is quite recent); notably there's no clear references for expressing this result beyond the *quadratic-loss case.* For this current work, we won't delve into the details of how this CLT is established nor how the *symmetries of the data* translate into symmetries of the *fluctuation process.* We will leave all these interesting questions for future work, and focus on the *global convergence* results that characterize the mean field dynamics.

[24] This means that $\forall t \in [0, T]$, $\forall f_1, \ldots, f_k \in C_b^2(\mathcal{Z})$, $\lim_{N \to \infty} \langle f_1 \times \cdots \times f_k, \rho_t^N \rangle = \Pi_{i=1}^k \langle f_i, \mu_t \rangle$. i.e. When observing a fixed number of particles, in the limit their joint law *behaves* like an i.i.d. sample from the fixed limiting process μ_t

2.2.8 Convergence to a Global Optimizer

Despite the good properties of the Mean Field Dynamics on the *unregularized case*, we fail to get proper **global convergence** results: though the loss function *decays* along the dynamic of μ_t, this is not necessarily *strict*, and *stationary points* of the dynamics might not correspond to global minima of the problem. What is known in the **noiseless case** is that **whenever μ_t converges (as $t \to \infty$) in W_2 distance, it does so to a global minimum of the loss function R**. This is stated in the following result (whose technical assumptions are given in chapter D):

Theorem 5 *[**Global Convergence**] (noiseless case) Consider $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$ and suppose that assumption 9 holds. Let $(\mu_t)_{t \geqslant 0}$ be a WGF of R initialized at μ_0.*

If *$(\mu_t)_t$ converges to some $\mu_\infty \in \mathcal{P}_2(\mathcal{Z})$ in W_2, then μ_∞ is a global minimizer of R over $\mathcal{M}(\mathcal{Z})$.*

In particular, if $(u_m(t))_{m \in \mathbb{N}, t \geqslant 0}$ is a sequence of classical gradient flows initialized in supp(μ_0) such that $\mu_{m,0}$ converges weakly to μ_0 then (limits can be interchanged):

$$\lim_{t,m \to \infty} R(\mu_{m,t}) = \min_{\mu \in \mathcal{M}(\mathcal{Z})} R(\mu).$$

*i.e. the **SGD Training dynamics converge (in long time) to the global minimum of the learning problem**.*

As noted in Descours et al. [24], studying this *noiseless* problem is quite hard, and not many results are known to hold. The introduction of *noise* and *regularization* becomes fundamental in order to ensure that an optimum will be achieved through the dynamic. Thus, a good framework for understanding the problem of **global convergence** could be the one introduced at the beginning of section 2.2.7: the **Mean Field Langevin Dynamics** (see Hu et al. [38], Chen et al. [13], Nitanda et al. [63] and Suzuki et al. [82] for some good references).

Recall the setting of our **regularized problem**, consider the following standard assumption:

Assumption 2 *The functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is **convex**, bounded from below (e.g. by 0) and it is of class $\mathcal{C}^1$ (as in definition 2.6; in particular, it is also l.s.c.).*

Given a **convex** functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ (e.g. for the *learning problem*: $R(\mu) = \mathbb{E}_\pi[\ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y)]$), we define its **regularized version** (with parameters $\tau, \beta > 0$ and $\nu \in \mathcal{P}(\mathcal{Z})$ a.c. wrt to λ, the lebesgue measure on $\mathcal{Z}$) as:

$$R_\nu^{\tau,\beta}(\mu) := R(\mu) + \tau \int r d\mu + \beta H_\nu(\mu)$$

Where $r : \mathcal{Z} \to \mathbb{R}$ is a *regularization* term and $H_\nu(\mu) := D(\mu || \nu) = \int \log(\frac{d\mu}{d\nu}(z)) d\mu(z)$ is the *relative entropy* between μ and ν (with $\mu \lll \nu$). When context is clear, we might simply write it as $R^{\tau,\beta}$ (usually when considering $\nu = \lambda$). An example in which ν might not necessarily be λ comes from the setting of Hu et al. [38], in which ν is chosen as the **Gibbs**

Measure in $\mathcal{Z}$. That is, the measure ν whose density (wrt λ) is given by:

$$g(x) = e^{-U(x)} \quad \text{with} \quad U : \mathcal{Z} \to \mathbb{R} \quad \text{s.t.} \quad \int_{\mathcal{Z}} e^{-U(x)}\, d\lambda(x) = 1,$$

Remark

- To unify our notation, we'll consider the case of $\nu = \lambda$ (Lebesgue measure) as one in which U is chosen as $U \equiv 0$ (even though strictly speaking this wouldn't satisfy the *integrability* condition).
- Notice that if we set $\tau = \beta$, choose $U(z) = r(z)$ and consider $R_\nu^{0,\beta}$ (without a regularization term; i.e. $R_\nu^{0,\beta} = R(\mu) + \beta H_\nu(\mu)$); it is the same as considering $R_\lambda^{\beta,\beta}$ with the corresponding regularization term (i.e. $R_\lambda^{\beta,\beta} = R(\mu) + \beta \int r d\mu + \beta H_\lambda(\mu)$). With this in mind, we can consider this setting in which *Gibbs Measures* are used as a particular case of our previously described framework. We will however retain both r and U for sake of *completeness.*

Now, consider some standard assumptions that might be taken about the *Gibbs measure's potential* (and also, equivalently, the risk regularizer):

Assumption 3 (As in Hu et al. [38]) *$U : \mathcal{Z} \to \mathbb{R}$ is assumed to be $\mathcal{C}^\infty$, with ∇U **Lipschitz** continuous, and such that $\exists C_U > 0$, $\exists C'_U \in \mathbb{R}$ such that*[25] *$\forall x \in \mathcal{Z}:\ \nabla U(x) \cdot x \geqslant C_U \|x\|^2 + C'_U$. When required, we will also assume that $r : \mathcal{Z} \to \mathbb{R}$ satisfies these conditions.*

The advantage of the *regularized problem* is that $R_\nu^{\tau,\beta}$ includes an **entropy term**, which guarantees **strict convexity, weak lower semicontinuity and compact sublevel sets**[26] for $R_\nu^{\tau,\beta}$. In particular, it will admit a **unique minimizer** $\mu^{*,\tau,\beta,\nu}$ (or, for simplicity, just μ^* when context is clear), as shown by the following proposition:

Proposition 9 (Existence and Uniqueness of the minimizer (regularized case)) *Let R satisfy assumption 2, and let ν be the Gibbs measure with potential U. Then, $R_\nu^{\tau,\beta}$ has a **unique minimizer**, $\mu^{*,\tau,\beta,\nu} \in \mathcal{P}(\mathcal{Z})$, absolutely continuous with respect to Lebesgue measure λ. When U satisfies assumption 3, it also belongs to $\mathcal{P}_2(\mathcal{Z})$.*

Proof. This result is taken directly from Hu et al. [38]. We include the proof in section C.6 for completeness. □

Under our goal of minimizing R over $\mathcal{P}(\mathcal{Z})$ through the training dynamic, we will be forced to pass through the *regularized version of the problem* if we want to achieve any sort of **global convergence guarantee**. Unfortunately, even though we will gain *global convergence* of the training dynamic, this will be to the *global minimum* of $R_\nu^{\tau,\beta}$, which could in principle be radically different to the minimizers of R in the original problem. Luckily, some *proximity* might be expected between the *regularized* and *unregularized* problem values, at least when ν is taken to be the Gibbs measure. This result is presented in a *slightly modified* version from its original formulation in Hu et al. [38]; we make it *more general* in order for it to fit in the *general framework of the regularized problem* we just presented:

[25] Note that these conditions imply that $\exists 0 \leqslant C' \leqslant C$ s.t. $\forall x \in \mathcal{Z}$, $C'\|x\|^2 - C \leqslant U(x) \leqslant C(1 + \|x\|^2)$ (i.e. U has quadratic growth) and $|\Delta U(x)| \leqslant C$

[26] See Hu et al. [38] and Lynch et al. [54]

Proposition 10 (**Γ-convergence**, as in Hu et al. [38]) *Let $\mathcal{Z} = \mathbb{R}^D$. If R is W_p-continuous, ν is a Gibbs measure whose potential U satisfies assumption 3 and the regularizer r also satisfies assumption 3, then $R_\nu^{\tau,\beta}$ Γ-converges to R when $\tau, \beta \downarrow 0$. Particularly, given $\mu^{*,\tau,\beta,\nu}$ the minimizer of $R_\nu^{\tau,\beta}$, we have*

$$\overline{\lim_{\tau,\beta \to 0}} R(\mu^{*,\tau,\beta,\nu}) = \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu).$$

In particular, every cluster point of $(\mu^{*,\tau,\beta,\nu})_{\tau,\beta}$ is a minimizer of R.

Proof. To *generalize* the result from Hu et al. [38] (to the simultaneous limit of $\tau, \beta \to 0$) we employ essentially their same techniques and follow their exact same proof structure. In any case, for completeness, we include it in Section C.6. □

Better still, as shown in Hu et al. [38], the minimizers of $R_\nu^{\tau,\beta}$ can be characterized via the following proposition:

Proposition 11 (from Hu et al. [38]) *Let R satisfy assumption 2, and let ν be the Gibbs measure with potential U; let both U and r satisfy assumption 3. Then, the following are equivalent:*

- $\mu^* = \arg\min_{\mu \in \mathcal{P}(\mathcal{Z})} R_\nu^{\tau,\beta}(\mu)$
- *μ^* is equivalent to λ (Lebesgue measure on $\mathcal{Z}$) and*

  $$\frac{\delta R}{\delta \mu}(\mu^*, z) + \tau r(z) + \beta \log(u^*(z)) + \beta U(z) \quad \textit{is a constant} \quad \lambda \text{ - a.s. } \forall z \in \mathcal{Z}$$

  where u^* denotes the density of μ^* wrt λ.

Proof. It can be done straightfowardly from Hu et al. [38] as it's a direct adaptation of their result (Proposition 2.5). □

In particular, in this setting it appears convenient to define, for any measure $\mu \in \mathcal{P}(\mathcal{Z})$ the probability measure $\hat{\mu}$ defined by its density wrt Lebesgue (slightly abusing notation):

$$\hat{\mu}(z) \propto \exp\left(-\frac{1}{\beta}\frac{\partial R}{\partial \mu}(\mu, z) - \frac{\tau}{\beta} r(z) - U(z)\right)$$

As noted in Nitanda et al. [63], Chen et al. [13], Proposition 11 tells us that the global minimum μ^* satisfies a *self-consistency* condition: $\mu^* = \widehat{\mu^*}$

In a similar spirit to that of section 2.2.7, we may recall that the **WGF** in $(\mathcal{P}_2(\mathcal{Z}), W_2)$ for $R_\nu^{\tau,\beta}$ (with ν the Gibbs measure of $\mathcal{Z}$) corresponds to:

$$\partial_t \mu_t = \varsigma(t) \left[\text{div}\left(\left(D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r + \beta \nabla_\theta U\right) \mu_t\right) + \beta \Delta \mu_t\right] \tag{2.8}$$

From classic works on the *mean field* literature (e.g. from Sznitman [83]) it is known that such an equation has a **unique strong solution** (under the right technical assumptions, see assumption 10 in Chapter D); in particular, it *regularizes* the measure solutions, making them have a *density*[27]. This also holds for the corresponding McKean-Vlasov Equation, which is written as:

$$dZ_t = \varsigma(t)\left[-\left(D_\mu R(\mu, Z_t) + \tau\nabla_\theta r(Z_t) + \beta\nabla_\theta U(Z_t)\right)dt + \sqrt{2\beta}dB_t\right] \text{ with } \mu_t = \mathbf{Law}(Z_t) \tag{2.9}$$

Where $(B_t)_{t\geqslant 0}$ is a D-dimensional standard Brownian Motion. As previously mentioned, this **McKean-Vlasov** equation is what's also referred to (in the literature) as the **Mean Field Langevin Dynamics (MFLD)**. Furthermore, as seen in theorem 3, it is expected that the SGD training procedure will converge to this *mean field training dynamic*. It will thus be desirable to relate the *marginals* of the MFLD to the global optimum of the regularized problem.

A first remarkable result in the *regularized case* is that the following *free energy dissipation* formula holds (and it can be proven using Itô Calculus; the technical assumptions, once again, shall be found in Chapter D).

Theorem 6 (from Hu et al. [38] and Chen et al. [13]) *Let $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$, and let assumption 3 and 10 hold; then:*

$$\forall t > 0,\ \frac{d}{dt}(R_\nu^{\tau,\beta}(\mu_t)) = -\varsigma(t)\int_{\mathcal{Z}}\left|D_\mu R(\mu_t, z) + \tau\nabla r(z) + \beta\frac{\nabla u_t}{u_t}(z) + \beta\nabla U(z)\right|^2 d\mu_t(z)$$

where u_t denotes the density of $\mu_t := \boldsymbol{Law}(X_t)$, the solution to equation (2.9). i.e. following the MFLD makes the regularized risk decrease at a known rate. This is known as the energy dissipation equation.

Remark Notice that this equation can be rewritten using the **Fisher divergence** (or *relative Fisher Information*) between two measures. This quantity is defined as:

$$I(\mu||\nu) := \int_{\mathcal{Z}}\left\|\nabla\log(\frac{d\mu}{d\nu}(z))\right\|^2 d\mu(z)$$

Then, almost by definition, we get:

$$\frac{d}{dt}(R_\nu^{\tau,\beta}(\mu_t)) = -\beta^2\varsigma(t)I(\mu_t||\hat{\mu}_t)$$

From this we could infer that the *stationary points for the dynamic* (i.e. those for which the left hand side becomes null), must be fixed points for the $\hat{(\cdot)}$ operator on $\mathcal{P}_2(\mathcal{Z})$. In particular, this can also serve as an alternative proof for proposition 11.

From the remark, we can see that theorem 6 implies that the MFLD converges to the unique global optimizer of the regularized problem:

[27] Moreover, assuming assumption 10 (see Hu et al. [38]), the solution is **stable** with respect to the initial law; i.e. $\forall\mu_0, \mu_0' \in \mathcal{P}_2(\mathcal{Z}),\ \forall t > 0,\ \exists C_t > 0 :\ W_2(\mu_t, \mu_t') \leqslant C_t W_2(\mu_0, \mu_0')$.

Theorem 7 (from Hu et al. [38]) *Let R satisfy assumption 2; also assume that assumption 3 and 10 hold. Consider $\mu_0 \in \cup_{p>2}\mathcal{P}_p(\mathcal{Z})$ and let $(\mu_t)_{t\geqslant 0}$ be the solution to the **MFLD** starting from μ_0. Then, the equation has an **invariant measure**[28], μ_∞, that satisfies:*

$$\mu_\infty := \arg\min_{\mu\in\mathcal{P}(\mathcal{Z})} R_\nu^{\tau,\beta}(\mu) \quad \textit{and} \quad \lim_{t\to\infty} W_2(\mu_t, \mu_\infty) = 0$$

Remark Global Convergence Results such as theorem 6 or theorem 7 have been established as early as in Mei et al. [57] (for the quadratic loss and under the simplest SGD dynamics). However, settings such as those of [13, 38, 63, 82, 17] are of incredible interest to establish essentially the same results under fundamentaly more general assumptions.

Making further *technical* assumptions on our regularized functionals leads to better *convergence results* as well as a *uniform-in-time* propagation of chaos result (established in Chen et al. [13]). Consider the following definition (which is prevalent in the MFLD literature, see [17, 13, 63, 82]):

Definition 2.8 *We say $\mu \in \mathcal{P}(\mathcal{Z})$ satisfies the Log-Sobolev Inequality with constant $\vartheta > 0$ (in short, LSI(ϑ)), if for any smooth function $\phi : \mathcal{Z} \to \mathbb{R}$ with $\mathbb{E}_\mu[\phi^2] < \infty$, we have:*

$$\mathbb{E}_\mu[\phi^2 \log(\phi^2)] - \mathbb{E}_\mu[\phi^2]\log(\mathbb{E}_\mu[\phi^2]) \leqslant \frac{2}{\vartheta}\mathbb{E}_\mu[\|\nabla\phi\|_2^2].$$

This is equivalent to saying that, $\forall \nu \in \mathcal{P}(\mathcal{Z})$ s.t. $\nu \lll \mu$:

$$D(\nu||\mu) := \int_{\mathcal{Z}} \log(\frac{d\nu}{d\mu}(z))d\nu(z) \leqslant \frac{1}{2\vartheta}\int_{\mathcal{Z}} \left\|\nabla\log(\frac{d\nu}{d\mu}(z))\right\|^2 d\nu(z) =: \frac{1}{2\vartheta}I(\nu||\mu)$$

*where $D(\nu||\mu)$ is the KL divergence and $I(\mu||\nu)$ is the **Fisher divergence** (or relative Fisher Information).*

Remark Written like that, this inequality serves to ensure *exponential convergence to minimizers* whenever $\mu = e^{-V}$ (for V some potential function) satisfies a LSI. In our setting, as done by most authors in recent years, we need to assume it *uniformly* over $\mathcal{P}_2(\mathcal{Z})$ to get the global convergence results we desire.

Assumption 4 (**Uniform LSI** from [17, 13, 63, 82]) *There exists $\vartheta > 0$ such that $\forall\mu \in \mathcal{P}_2(\mathcal{Z})$, $\hat{\mu}$ satisfies LSI(ϑ).*

Remark This **LSI** is a recurrent element in the literature of WGF and Optimal Transport in general (see, for instance, Otto and Villani [66]). In particular, it implies the *Poincaré Inequality*:

$$\forall\phi \in \mathcal{C}_b^1(\mathcal{Z}),\ \mathrm{Var}_{\hat{\mu}}(\phi) \leqslant \frac{1}{2\vartheta}\mathbb{E}_{\hat{\mu}}[|\nabla\phi|^2]$$

Talagrand's T_2-transport inequality follows as well:

$$\forall\nu \in \mathcal{P}_2(\mathcal{Z}),\ \vartheta W_2^2(\nu,\hat{\mu}) \leqslant D(\nu||\hat{\mu})$$

Moreover, all these inequalities are *stable under tensorization*, which allows for properly studying the Propagation of Chaos problem (as we'll see right after).

[28] A measure μ is said to be **invariant** for the equation if $\mathbf{Law}(X_t) = \mu$ for all $t \geqslant 0$

Beyond the *characterization* of the decay provided by Hu et al. [38], Chen et al. [13] provide the following guarantee:

Theorem 8 (from Chen et al. [13] and Chizat [17]) *Let assumptions 2, 4 and 10 hold. Then, if for some $t_0 \geqslant 0$, μ_{t_0} has* ***finite entropy*** *and* ***finite second moment****; then $\forall t \geqslant t_0$,*

$$D(\mu_t||\mu_\infty) \leqslant R_\nu^{\tau,\beta}(\mu_t) - R_\nu^{\tau,\beta}(\mu_\infty) \leqslant (R_\nu^{\tau,\beta}(\mu_{t_0}) - R_\nu^{\tau,\beta}(\mu_\infty))e^{-2\beta\vartheta\int_{t_0}^t \varsigma(s)ds}$$

where $\mu_\infty = \mu^{\tau,\beta,\nu} = \arg\min_{\mu\in\mathcal{P}(\mathcal{Z})} R_\nu^{\tau,\beta}(\mu)$. By Talagrand's inequality this also amounts to exponential W_2 convergence of μ_t to μ_∞ [29]*.*

i.e. The value function following the MFLD converges exponentially fast to the optimum value of the problem. This also implies an exponential convergence in relative entropy.

Proof. The result in Chen et al. [13] is established in the setting with $\tau = 0, \beta = 1$ and $\varsigma \equiv 1$; however, from the insight taken from Chen et al. [13], Chizat [17] (among others) one can show that, in its most general form, the result holds as stated. □

All the results we've stated speak to the ability of the limiting Mean Field process to converge in large time to the unique minimizer of the regularized problem (the so-called *global convergence* results). Now, could we directly relate this global convergence to the original *discrete* SGD dynamics with *finitely many particles*? i.e. can we quantify how the *the regularized SGD training dynamics* approaches the global optimum of the regularized population risk $R_\nu^{\tau,\beta}$? Works such as Chizat [17], Chen et al. [13], Nitanda et al. [63] provide interesting results (such as a *uniform-in-time* propagation of chaos), but centered in the **non-stochastic** Gradient Descent Training. SGD results mainly come from the recent developments by Suzuki et al. [82].

In any case, we will now revisit some *Propagation of Chaos* results, under the lens of the regularized dynamics. We know that a distribution over N particles (i.e. a random variable $Z = (Z^i)_{i=1}^N$) can be expressed by its law $\mu^{(N)} \in \mathcal{P}(\mathcal{Z}^N)$. As we've already seen, the idea behind the *propagation of chaos* results basically states that, as the number of particles N increases, particles behave as if they are independent; in some sense, the joint distribution of the N particles ($\mu^{(N)} \in \mathcal{P}(\mathcal{Z}^N)$) *approaches* a product measure ($\overline{\mu}^{\otimes N}$ for some $\overline{\mu} \in \mathcal{P}(\mathcal{Z})$). This is exactly what the last variant of theorem 4 demonstrates for the *continuous-time* SGD training process (which approaches *independent* realizations of the mean field dynamics). In a similar way, Chen et al. [13] establish a *uniform-in-time* propagation of chaos result for the (continuous-time) particule system[30] that follows the MFLD equation (2.6). Though they have stronger results, the following *corollary* illustrates the point pretty well:

Corollary 3 (from Chen et al. [13]) *Assume R satisfies assumptions 2 and 10 (2.) and assumption 4. Suppose $m_0 \in \mathcal{P}_6(\mathcal{Z})$, m_0 has finite entropy, and $m_0^N = m_0^{\otimes N}$ (initialization is*

[29] Thus, under the right technical assumptions, it also amounts to Theorem 4 of Mei et al. [57]

[30] i.e. the particles $Z = (Z^i)_{i=1}^N$ are initialized i.i.d. and they follow, for every $i \in \{1, \ldots, N\}$

$$dZ_t^i = \varsigma(t)\left[-\left(D_\mu R(\nu_Z^N, Z_t^i) + \tau\nabla_\theta r(Z_t^i)\right)dt + \sqrt{2\beta}dB_t\right]$$

i.i.d.). Then there exist constants $C, \kappa, N_0 > 0$, depending on $\vartheta, M^R_{mm}, M^R_{mx}, m_0$, and D (as in $\mathcal{Z} = \mathbb{R}^D$), such that:

$$\sup_{t\in[0,\infty)} \frac{1}{N} W_2^2(m_t^N, m_t^{\otimes N}) \leqslant \frac{C}{N^\kappa} \quad \textit{for every } N \geqslant N_0$$

If additionally R is such that $\sup_{\mu\in\mathcal{P}_2(\mathcal{Z})} \sup_{x\in\mathcal{Z}} |\nabla^k D_\mu R(\mu, x)| < +\infty$ for $k = 2, 3$, we also have:

$$\sup_{t\in[0,\infty)} \frac{1}{N} D(m_t^N || m_t^{\otimes N}) \leqslant \frac{C}{N^\kappa} \quad \textit{for every } N \geqslant N_0$$

upon redefining the constants $C, \kappa, N_0 > 0$.

Beyond the continuous time Gradient Descent Dynamics, we can get bounds (as in Suzuki et al. [82]) relating the noisy *discrete time* SGD dynamics (as given by equation (2.2)) to the optimum of the regularized problem. First, we might desire to adapt our *risk* functional $R^{\tau,\beta} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ in order to evaluate this *N-particle system.* Under this lens, it's natural to define the *risk* associated to the whole particle system as $R^{N,\tau,\beta} : \mathcal{P}(\mathcal{Z}^N) \to \mathbb{R}$ such that $\forall \mu^{(N)} \in \mathcal{P}(\mathcal{Z}^N)$:

$$R^{N,\tau,\beta}(\mu^{(N)}) = N\mathbb{E}_{\theta\sim\mu^{(N)}}[R^\tau(\nu_\theta^N)] + \beta H_\lambda(\mu^{(N)})$$

Knowing that in the *regularized* case there is a unique minimum to which we can converge, we might wonder if $R^{N,\tau,\beta}(\mu^{(N)})$ will be close (or not) to $NR^{\tau,\beta}(\mu^*)$. Indeed, from theorem 2, one could get the following bound:

$$0 \leqslant \inf_{\mu^{(N)}\in\mathcal{P}(\mathcal{Z}^N)} \frac{1}{N} R^{N,\tau,\beta}(\mu^{(N)}) - R^{\tau,\beta}(\mu^*) \leqslant \frac{C_{\tau,\beta}}{N}$$

for some constant $C_{\tau,\beta} > 0$. Thanks to the following result, we could understand that, propagation of chaos might be achieved as long as we're able to control the difference in *risk*:

Lemma 5 (from Suzuki et al. [82]) *Assume assumption 2, assumption 3, assumption 4 and that $\exists \lambda_1, \lambda_2 > 0$ and $c_r > 0$ such that $\forall x \in \mathcal{Z}$ $\lambda_1 \mathrm{Id}_\mathcal{Z} \leqslant \nabla\nabla^\top r(x) \leqslant \lambda_2 \mathrm{Id}_\mathcal{Z}$ (in the matrix order), $x^\top \nabla r(x) \geqslant \lambda_1 \|x\|^2$, and $0 \leqslant r(x) \leqslant \lambda_2(c_r + \|x\|^2)$, and $\nabla r(0) = 0$. Then:*

$$W_2^2(\mu^{(N)}, \mu^{*\,N}) \leqslant \frac{2}{\beta\vartheta}(R^{N,\tau,\beta}(\mu^{(N)}) - NR^{\tau,\beta}(\mu^*)).$$

In particular, results along the following lines have been established:

Theorem 9 (Sketch, from Suzuki et al. [82]) *Let μ^* be the optimum of the regularized problem and $(\theta_k)_{k\in\mathbb{N}}$ be the parameters trained using equation (2.2) (with $\alpha > 0$ a constant learning rate). Assume assumptions 11 to 13 and $\beta\alpha\vartheta \leqslant \frac{1}{4}$, $\alpha \leqslant \frac{\lambda_1}{4\lambda_2}$. Then: $\frac{1}{N}\mathbb{E}\left[R^{N,\tau,\beta}(\mu_k^{(N)})\right] - R^{\tau,\beta}(\mu^*) \leqslant e^{-C_{\tau,\beta,\vartheta}k}\left[\frac{1}{N}\mathbb{E}\left[R^{N,\tau,\beta}(\mu_k^{(N)})\right] - R^{\tau,\beta}(\mu^*))\right]$ + constants*

We won't dive into excessive details for such results, as they have been recently developped, and escape a bit from the main focus of the current work.

Many extensions of the setting presented here have been studied in the literature (notably, the *annealed* dynamics from [17], among many others). We will, however, not dive any further into them in this review, as sufficient detail of the key elements of the MF theory of *shallow NNs* have been provided; further insight shall be sought in the original material.

2.2.9 Applicability to the NN setting

Among the sea of different frameworks and results provided in the previous section, one may wonder whether the relevant **hypothesis** for some of the key results (which were given in their most *general* forms) hold in the setting of the *shallow* NNs that appear in practice.

The first restriction most of these results face, is that they will require σ_* to be **bounded** in order to work. This will, unfortunately, go against properties such as the *approximation power* of NNs (in particular, their *universality*). The standard assumption will be the following:

Assumption 5 (Standard Assumptions on *shallow NNs*) *A standard NN setting will be the following:*

1. *Let* $\mathcal{X} = \mathbb{R}^d$ *and* $\mathcal{Y} = \mathbb{R}^c$ *and* $\mathcal{Z} = \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} \times \mathbb{R}^b$.
2. *Let* $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ *be of the form:*

$$\forall \theta = (W, A, B) \in \mathcal{Z};\ \forall x \in \mathcal{X},\ \sigma_*(x;\theta) = \varphi(W)\sigma(A^T x + B)$$

   where $\sigma : \mathbb{R} \to \mathbb{R}$ *is an activation function (which is applied pointwise) and* $\varphi : \mathbb{R} \to [-M, M]$ *is a* ***truncation function*** *(also applied pointwise), with* $M < +\infty$. σ *and* φ *are assumed to be at least* ***continuously differentiable***.

3. *Let* $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ *be the data distribution, with finite second moment*[31].
4. *Let* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *be a* ***convex*** *and* ***smooth*** *loss function, such that* $\ell \geqslant 0$.
5. *With all of these elements, consider* $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ *given by,* $\forall \mu \in \mathcal{P}(\mathcal{Z})$, $R(\mu) = \mathbb{E}_\pi[\ell(\langle \sigma_*(X,\cdot), \mu\rangle, Y)]$.

These assumptions ensure (together with stronger impositions) that:

Corollary 4 *Suppose assumption 5 holds:*

- *Further assume* $\forall \theta \in \mathcal{Z}$, $\sigma_*(\cdot,\theta) \in L^2(\pi|_\mathcal{X})$ *and* $\exists C > 0$, $\forall \theta \in \mathcal{Z}$, $\|\sigma_*(\cdot,\theta)\|_{L^2(\pi|_\mathcal{X})} \leqslant C(1 + |\theta|^2)$. *Then* $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ *is* ***convex***, $\mathcal{C}^1$ *and* ***bounded from below*** *(i.e. assumption 2 holds).*
- *Further assume* $\ell(y,\hat{y}) = \|y - \hat{y}\|_2^2$ *(quadratic loss),* $r(\theta) = \|\theta\|^2$ *(quadratic regularization) and also that* σ *and* φ *have bounded derivatives to up to fourth order. Also suppose that* σ *is sigmoidal (i.e. continuous, non decreasing,* 0 *at* $-\infty$ *and* 1 *at* $+\infty$*). Then assumption 10 (2.) hold as well:*

[31] An alternative, stronger, but standard hypothesis is to suppose π compactly supported

- *$D^2_\mu R^\tau$ has* ***bounded 2-norm****, making $D_\mu R^\tau(\cdot, z)$ W_1-Lipschitz with constant $M^{R^\tau}_{mm} = \left(\|\varphi'\|_\infty + \|\varphi\|_\infty \|\sigma'\|_\infty (1 + \int |x|^2 \pi_{\mathcal{X}}(dx)\right)^{1/2}$.*
- *By boundedness of the derivatives of σ and φ, the rest of the bounds follow as well.*

Furthermore, by the boundedness of φ and the fact that $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$, assumption 4 holds with $\vartheta = \tau \exp\left(-2(E_\pi[\|Y\|] + \|\varphi\|_\infty)\|\varphi\|_\infty\right)$.

Remark The truncation function $\varphi : \mathbb{R} \to [-M, M]$, introduced to make σ_* bounded (in order to ensure minimal results), hinders on the ability of shallow NNs to approximate arbitrary functions. As noted previously, if $M < +\infty$, any $h \in \mathcal{F}_{\sigma_*}(\mathcal{P}_2(\mathcal{Z}))$ will satisfy that $\|h\|_\infty \leqslant M$. Thus, we will never be able to *perfectly* approximate unbounded functions[32].

Despite this, Chen et al. [13] observe that Barron's theorem (see Barron [6]) might be used to validate the results *even under truncation.* Indeed, if $f^* = \mathbb{E}_{\pi[Y|X=\cdot]}$ is such that $\exists F^*$ complex-valued measure (which we call its *fourier transform*), $\exists K_r > 0$:

$$\forall x \in B(0, K_r), \;\; f^*(x) = f^*(0) + \int_{\mathbb{R}^d} (e^{i\omega \cdot x} - 1) F^*(d\omega)$$

If $M \geqslant K_r \int_{\mathcal{Z}} |\omega||F(d\omega)| + |f(0)|$, and $\exists c_+, c_- \in \mathbb{R}$ such that $\varphi(c_+) = M$ and $\varphi(c_-) = -M$; then, for every $\pi|_{\mathcal{X}} \in \mathcal{P}(B(0, K_r))$, the best approximation error is zero; i.e.

$$\inf_{\Phi \in \mathcal{F}_{\sigma,\varphi}(\mathcal{P}_2(\mathcal{Z}))} \|f - \Phi\|_{L^2(\pi|_{\mathcal{X}})} = 0$$

2.3 Mean Field Limit in *Deep Neural Networks*

From the previous section, it becomes clear that the understanding of the Mean Field limit of single-layer NNs is in a quite advanced stage. The next natural step in this context is to dive into how these results and behaviours change when considering *deep* neural networks (i.e. with more than one hidden layer).

As we've previously signaled, *some* of the *deep neural network architectures* can be expressed in the form: $\frac{1}{N}\sum_{i=1}^N \sigma_*(\cdot, \theta_i)$ for *some* descriptor $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ (see Rotskoff and Vanden-Eijnden [72] for further reference). However, a true *mean field* understanding of deep neural networks, to their full extent, is part of the literature's ongoing work.

In particular, the works of Araújo et al. [2] and Sirignano and Spiliopoulos [79] began the analysis in that direction, managing to prove some relevant results, but with multiple technical difficulties that each overcame in different ways.

In the case of Sirignano and Spiliopoulos [79], they study 2-hidden-layer networks (WLOG as it's sufficient to understand the general case), of the form:

$$\Phi_{N_1,N_2}(x;\theta) := \frac{1}{N_2}\sum_{i=1}^{N_2} C^i \sigma\left(\frac{1}{N_1}\sum_{j=1}^{N_1} W^{2,i,j}\sigma(W^{1,j} \cdot x)\right)$$

[32] This contrasts the case of $M = +\infty$ and $\pi|_{\mathcal{X}}$ compactly supported, when we know, by corollary 1, that the *infimum of the problem is zero*

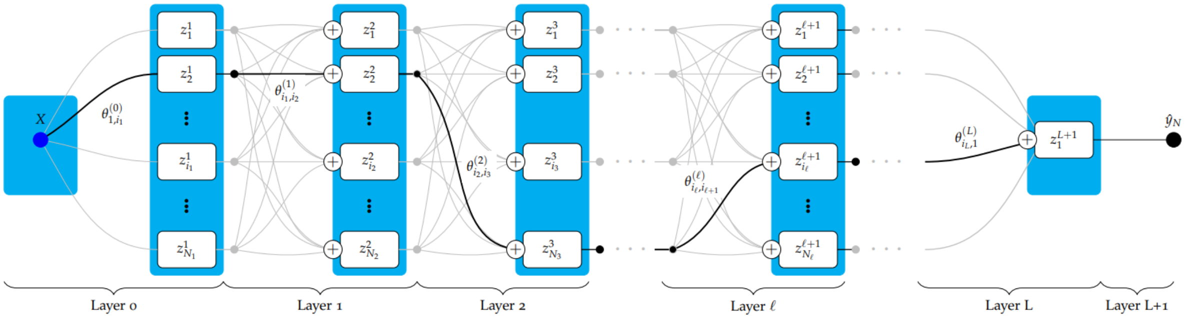


Figure 2.5: Illustration of the *Paths of weights* in a multilayer neural network: these are the basic units that are studied in the Mean Field limit of deep neural networks. Taken from Araújo et al. [2].

where the parameters are $\theta = (C^1, \dots, C^{N_2}, W^{2,1,1}, \dots, W^{2,N_1,N_2}, W^{1,1}, \dots, W^{1,N_1})$, and they are updated following the usual SGD. Unfortunately, for technical reasons, they are forced to consider *rescaled learning rates* in the form: $\alpha_C^{N_1,N_2} = \frac{N_2}{N_1}$, $\alpha_{W,1}^{N_1,N_2} = 1$, $\alpha_{W,2}^{N_1,N_2} = N_2$ (all of which are very rarely used values in practice). For the same reason, the limit of *infinite width* that they consider must be taken one layer at a time (i.e. $\lim_{N_2 \to \infty} \lim_{N_1 \to \infty} \Phi_{N_1,N_2}(x; \theta_t)$). This introduces its own technical difficulties and, above all, is very unintuitive from the perspective of applications.

On the other hand, Araújo et al. [2] manage to relax the assumptions of Sirignano and Spiliopoulos [78], at the cost of *freezing* the **first and last layers** of the network (i.e. having them *not* be trained in the SGD iterations; similar to *random features*) and having all layers scale according to the same value of N. This is clearly restrictive in terms of *applications*, but it avoids the problem of having *'2 different scalings'* between the *external* layers (according to N) and the *internal* layers (according to N^2) within multilayer networks. Unfortunately, this introduces other technical problems related to the continuity of a conditional probability distribution.

Both papers try to attack the problem in a similar fashion as in the *shallow NN* case: they seek for a *limiting mean field process* that could potentially approximate the SGD training dynamic, establishing their well-posedness, as well as relevant *propagation of chaos* (and eventually *global convergence*) results. Unfortunately, in the setting of *deep NNs*, the **weights cannot become statistically independent in the large N limit** (due to the *interconnection between layers*). Therefore, a proper analysis of how *propagation of chaos* takes place in the multilayer case must be done with the so-called *paths of weights* in the network. This idea is illustrated in Figure 2.5.

A more *modern* understanding of the *mean field limit* of *deep* neural networks has been recently established in Nguyen and Pham [61]. They make use of what they refer to as *neuronal embeddings* and manage to establish both a relevant *mean field regime* (characterized by a system of ODEs) as well as some *global convergence* guarantees; all while not having to assume *convexity* of the loss function (which is crucial in all the theory described in section 2.2). Despite the undeniable interest of such an approach, the techniques employed in Nguyen and Pham [61] are rather *heuristical* and they also escape from the realm of our current research on the topic; we will therefore refrain from diving further into its details.

More generally, the increased complexity and the limited literature related to the Mean Field limit of multilayer neural networks have kept us from delving deeper into these ideas. Without a doubt, one of our objectives for future work is to better understand this MF limit of deep NNs, particularly under the lens of *symmetries*. We advise the interested reader to seek further details in the relevant literature.

Chapter 3

Exploiting Symmetries with Neural Networks

In recent times, *Deep Learning* models have revolutionized the technological industry. However, the *deep* models that have seen the most success, don't exactly correspond to the *fully connected 'shallow' NNs* considered in the previous section. The most *successful* NN architectures in practice (such as CNNs, RNNs, Transformers and GNNs) are those that *leverage* the intrinsic properties of data to reduce their complexity.

Example [**Invariant and Equivariant Tasks**] To illustrate the *idea* behind our depiction of symmetries in the *learning* framework, we consider two examples of *tasks* which inolve some kind of *symmetry* that we might want to exploit:

- If we wanted to detect the presence (or absence) of a *dog* in an image, it shouldn't matter to us in *which orientation* the image arrives (see Figure 3.1 for *some orientations*): the underlying *classification function* we want to discover (i.e. the *way* of associating an image to a label) *does not depend on the orientation of the photograph*; for all possible orientations, a dog should be detected. This is what we call an *invariant task*.
- Analogously, if instead of *classifying* whether a dog is present (or not) in an image, we were interested in detecting the *position* of the dog's nose in the image, *simmetry* also plays a role. When *rotating* the image by a certain amount (e.g. 90°) the *detected nose position* should rotate by an equal amount. That is, our *underlying "nose-detection" function* should *commute* with the symmetric transformation. This is what we call an *equivariant task*.

With this in mind, it seems natural that a *good NN model* should be capable of *understanding* and *exploiting* these *symmetries* present in the data to achieve better results on the given task. More precisely, the NN architecture considered for the problem should take the symmetry into account.

In this work, we will focus on understanding the different *symmetry-leveraging techniques* that are commonly used in the literature to *exploit* a problem's symmetry. These include ideas

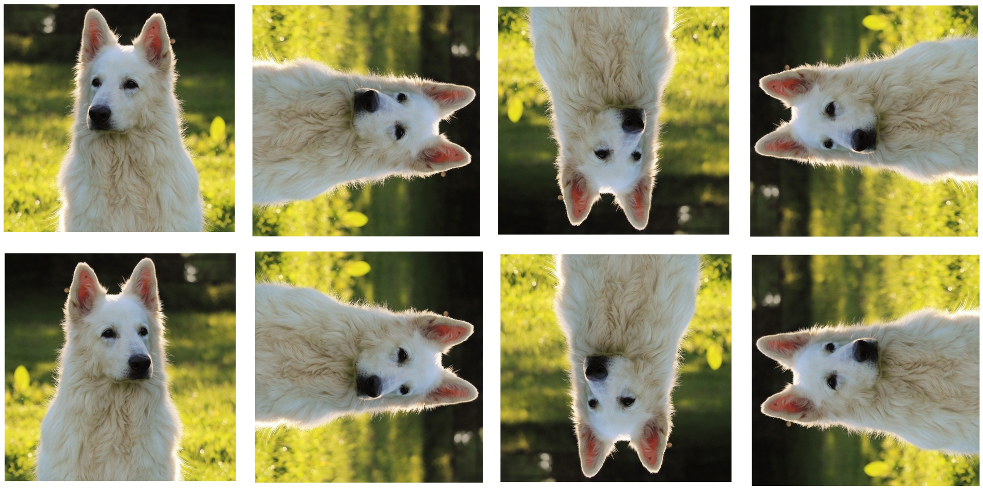


Figure 3.1: Illustration of possible orientations of a photograph under the action of group $G = D_4$.

such as Data Augmentation (**DA**), Feature Averaging (**FA**), and the recently popularized so-called *Equivariant Architectures* (**EA**) of neural networks. This all falls within the context of **Geometric Deep Learning** (GDL) [10] and, in particular, equivariant architectures account for a significant portion of the most popular architectures today (CNNs [34], Transformers [86], GraphNNs [74], among many others).

In what follows, a brief overview of results on the theory of *Geometric Deep Learning* will be given. Particularly, ideas from Bronstein et al. [10],Kondor and Trivedi [47],Cohen et al. [19], Finzi et al. [30], Finzi et al. [31],Elesedy and Zaidi [28] and Flinth and Ohlsson [32] will be discussed. Deep Learning Literature is rapidly and constantly moving, so possibly some interesting elements from the literature are being left out: a more complete view might be found by going through the original material.

3.1 Symmetries as Group Actions

A first thing we notice in our example, is that the **symmetries** of the data are being encoded as an action of a **group** G over the *features* and *labels*.

Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be general topological spaces endowed with their Borel σ-fields (often they will be *separable* **Hilbert spaces** or even, as for the practical implementation of NNs, just $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^c$ and $\mathcal{Z} = \mathbb{R}^D$). Consider a **topological group** $(G, \mathcal{T}_G)$ [1] endowed with its **Borel σ-field**.

We say that G acts on $\mathcal{Z}$ (on the *left*), which we denote $G \circlearrowright \mathcal{Z}$, whenever there exists a

[1] Recall that this means that, under the topological structure given by $\mathcal{T}_G$, the group operations (multiplication and inversion) are **continuous**

map:

$$\begin{aligned} T : G{\times}\mathcal{Z} &\to \mathcal{Z} \\ (g,z) &\mapsto T(g,z) \end{aligned}$$

that satisfies $T(e_G, z) = z$ and $T(g_1, T(g_2, z)) = T(g_1.g_2, z),\ \forall g_1, g_2 \in G,\ \forall z \in \mathcal{Z}$. Equivalently, this amounts to having a **group homomorphism** between G and $\mathrm{Sym}(\mathcal{Z})$ (this is why we usually denote $T(g, \cdot)$ by T_g). The action is said to be *faithful* when this homomorphism is injective; it is said to be *free* whenever $\forall z \in \mathcal{Z}$ the *stabilizer*[2] of z is trivial, i.e. $G_z = \{e_G\}$. We denote by Gz the **orbit** of $z \in \mathcal{Z}$ by G (i.e. $\{gz \ : \ g \in G\}$). The **orbit space**, denoted $G\backslash\mathcal{Z}$ is the set of all orbits of points in $\mathcal{Z}$. There exists a **canonical projection** map $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ that associates, to every $z \in \mathcal{Z}$ its associated orbit $Gz \in G\backslash\mathcal{Z}$. $G\backslash\mathcal{Z}$ is usually endowed with the **quotient topology** ($\mathcal{T}_{G\backslash\mathcal{Z}} := \{A \subset G\backslash\mathcal{Z} \ : \ p^{-1}(A) \in \mathcal{T}_{\mathcal{Z}}\}$, the smallest one making p continuous[3]).

We say that G **acts continuously** on $\mathcal{Z}$ whenever T is **continuous** (with respect to the product topology) [4]. It is clear that, a **continuous action**, $G \circlearrowright \mathcal{Z}$, acts on $\mathcal{Z}$ via **homeomorphisms** ($\forall g \in G,\ T_g : \mathcal{Z} \to \mathcal{Z}$ is an homeomorphism). We might further assume, when $\mathcal{Z}$ is metric, that G acts on $\mathcal{Z}$ via **isometries** (i.e. $\forall g \in G,\ T_g : \mathcal{Z} \to \mathcal{Z}$ is an isommetry of $\mathcal{Z}$). Finally, we say that the action $G \circlearrowright \mathcal{Z}$ is **proper** if for **every** pair of compact sets $K_1, K_2 \subseteq \mathcal{Z}$ the set $G_{K_1,K_2} = \{g \in G \ : \ T_g(K_1) \cap K_2 \neq \varnothing\}$ is **compact** (in G)[5].

We hereby state some remarkable properties of group actions:

Proposition 12 (Properties of Continuous Group Actions (see [25, 68, 26, 8])) *Let $G \circlearrowright \mathcal{Z}$ be a continuous group action; then:*

1. *$p : \mathcal{Z} \to G\backslash\mathcal{Z}$ is **continuous** but also an **open** map (i.e. $\forall U \in \mathcal{T}_{\mathcal{Z}},\ p(U) \in \mathcal{T}_{G\backslash\mathcal{Z}}$)*
2. *If $\mathcal{Z}$ and G are Hausdorff and $G \circlearrowright \mathcal{Z}$ properly, then $\forall z \in \mathcal{Z}, Gz \subseteq \mathcal{Z}$ is closed and also $G\backslash\mathcal{Z}$ is Hausdorff. If $\mathcal{Z}$ is also **locally compact**[6], then $G\backslash\mathcal{Z}$ is **locally compact** as well.*
3. *If G is **compact**, then it acts **properly** on any Hausdorff $\mathcal{Z}$.*
4. *The following are holds when both G and $\mathcal{Z}$ are Polish spaces (i.e. separable and metrizable by a complete metric); and G is locally compact while $\mathcal{Z}$ is Hausdorff:*

$$[G\backslash\mathcal{Z} \textit{ is } T_0] \iff \mathcal{T}_{G\backslash\mathcal{Z}} \textit{ generates } \mathcal{B}_{G\backslash\mathcal{Z}} \iff \textit{Each orbit is } G_\delta \textit{ in } \mathcal{Z}$$

If further G and $\mathcal{Z}$ are first countable [7], these are also equivalent to:

$$\iff \exists s : G\backslash\mathcal{Z} \to \mathcal{Z} \textit{ a Borel cross-section of } p \iff p \textit{ admits a Borel transversal}$$

[2] Recall that the stabilizer subgroup of an element $z \in \mathcal{Z}$ is given by $G_z := \{g \in G \ : \ T_g.z = z\}$

[3] Similarly, one may define the natural *Borel σ-field* on $G\backslash\mathcal{Z}$ as $\mathcal{B}_{G\backslash\mathcal{Z}} := \{A \subset G\backslash\mathcal{Z} \ : \ p^{-1}(A) \in \mathcal{B}_{\mathcal{Z}}\}$

[4] Similarly, we say G **acts measurably** on $\mathcal{Z}$ whenever T is measurable (wrt the product σ-field)

[5] Equivalently, if the **graph** application $(g, z) \in G \times \mathcal{Z} \mapsto (z, T_g.z) \in \mathcal{Z} \times \mathcal{Z}$ is **proper**.

[6] A potential problem with assuming local compactness would be that if $\mathcal{Z}$ was a t.v.s., this would force it to be finite-dimensional.

[7] i.e. each point has a countable basis of open sets

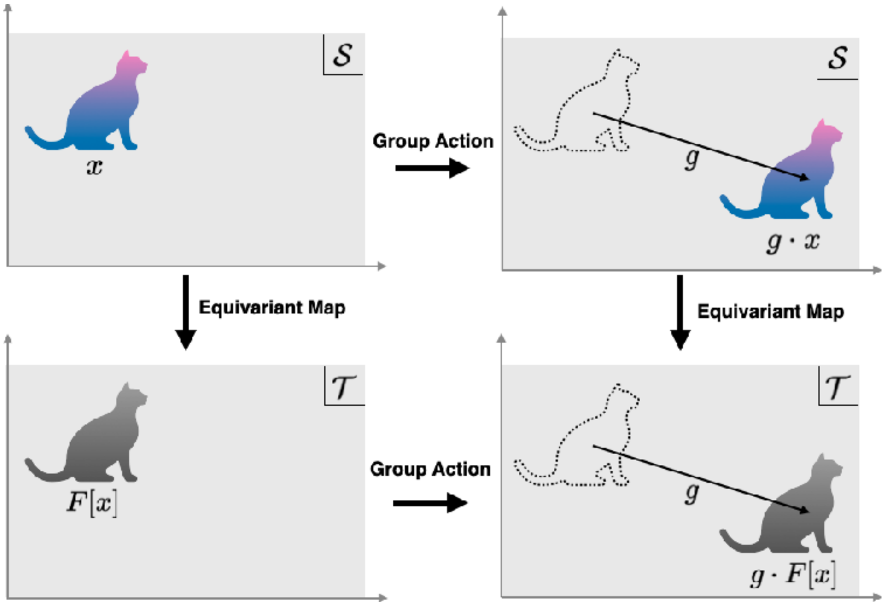


Figure 3.2: Illustration of an Equivariant Function. Taken from Kumagai and Sannai [48]

The standard assumption we will make is that G is locally compact, second countable and Hausdorff (denoted lcsH). Under these conditions, it is well known (see Druţu and Kapovich [25]) that G admits a (left)-**Haar measure** $\lambda_G \in \mathcal{M}(G)$ (i.e. a **left**-invariant measure on G that's finite on every compact set, outer regular on Borel sets and inner regular on open sets). In particular, whenever G is **compact**, we know that λ_G can be normalized (and it is unique and also **right G-invariant**); in this case, it is to be interpreted as the *uniform distribution* on G.

For our study of the **symmetries** of NNs, the following notion of an Invariant (or Equivariant) map will be key:

Definition 3.1 ***[Invariant/Equivariant Functions]*** *Let $G \circlearrowright \mathcal{X}$ (via $T^{\mathcal{X}}$) and $G \circlearrowright \mathcal{Y}$ (via $T^{\mathcal{Y}}$). We say that $f : \mathcal{X} \to \mathcal{Y}$ is G-**equivariant** if*

$$\forall g \in G: \ f \circ T_g^{\mathcal{X}} = T_g^{\mathcal{Y}} \circ f$$

*If $T^{\mathcal{Y}} \equiv \mathrm{Id}_{\mathcal{Y}}$ (trivial action), we say that f is G-**invariant**.*

The idea behind equivariance is well represented in Figure 3.2: an equivariant function, under transformations in its input, undergoes the *same* transformations in its output.

An interesting particular case of the last definition appears for *multivariable functions*:

Definition 3.2 ***[Jointly Invariant/Equivariant Functions]*** *Let $G \circlearrowright \mathcal{X}_1$ (via $T^{\mathcal{X}_1}$), $G \circlearrowright \mathcal{X}_2$ (via $T^{\mathcal{X}_2}$) and $G \circlearrowright \mathcal{Y}$ (via $T^{\mathcal{Y}}$). We say that $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$ is **jointly** G-**equivariant** if*

$$\forall g \in G, \ \forall x_1 \in \mathcal{X}_1, \ \forall x_2 \in \mathcal{X}_2: \ f(T_g^{\mathcal{X}_1}.x_1, T_g^{\mathcal{X}_2}.x_2) = T_g^{\mathcal{Y}}.f(x_1, x_2)$$

*If $T^{\mathcal{Y}} \equiv \mathrm{Id}_{\mathcal{Y}}$ (trivial action), we say that f is **jointly** G-**invariant**.*

*We may also call such functions G-equivariant/invariant on **both arguments**, or even just plainly G-equivariant/invariant*[8]*. An analogous definition still holds for an arbitrary (finite) amount of input arguments.*

Beyond the general definitions, we will now restrict ourselves to the setting of **compact groups** that act over (separable) **Hilbert spaces** via **representations** (i.e. linear maps)[9].

Definition 3.3 ***[Group Action via Representations]*** *Let G be a **compact group** with normalized Haar measure λ_G, and let $\mathcal{Z}$ be a (separable) **Hilbert space**.*

*We say that G acts on $\mathcal{Z}$ **linearly** via the **representation** ρ (denoted $G \circlearrowright_\rho \mathcal{Z}$) when ρ is a **group homomorphism***

$$\rho : G \to \mathrm{GL}(\mathcal{Z})$$

i.e. it associates each $g \in G$ to a linear and bounded invertible operator $\rho(g) \in \mathrm{GL}(\mathcal{Z})$, also satisfying the relation $\forall g, h \in G,\ \rho(gh) = \rho(g)\rho(h)$. This is esentially the definition of a continuous group action, but $\rho_g : \mathcal{Z} \to \mathcal{Z}$ is not only a homeomorphism: it is also linear.

*One may further assume that the considered representation is **orthogonal** (or **unitary**). That is, for all $g \in G$, $\rho(g)$ is a **unitary** operator ($\rho(g)\rho(g)^* = \mathrm{Id}_\mathcal{Z}$).*

Example Despite assuming a **compact** group and **orthogonal** representations might seem somewhat restrictive, many well known examples can be placed in this setting:

- The *trivial representation* (where $\forall g \in G,\ \rho(g) = \mathrm{Id}_\mathcal{Z}$) is the simplest (yet most common) example.
- $\mathcal{S}_n$ acting on $\mathbb{R}^n$ by permutation of the coordinates. This also extends to an action on $\mathbb{R}^{n^k}$ via *simultaneous* permutation of coordinates.
- $\mathbb{Z}_n^2$ acting on images in $\mathbb{R}^{n\times n}$ via *translations* (cyclic rotation of coordinates).
- C_4 acting on images in $\mathbb{R}^{n\times n}$ via 90° rotations.
- An infinite-dimensional example (which appears widely in the literature) is that of the action of any compact G over $L^\infty(G)$. It is given, for all $f \in L^\infty(G),\ g, h \in G$, by $(\rho(g).f)(h) = f(g^{-1}h)$; which is indeed a linear and bounded map.

As an illustration, consider figure 3.1, where the group $G = D_4$ is acting on the space of *images* $\mathbb{R}^2$ by 90° rotations and *vertical flips*. This can be understood via the representation given by $\left\langle \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\rangle$.

Remark Notice that, as long as G is compact, any representation $G \circlearrowright_\rho \mathcal{Z}$ can be cast as an orthogonal representation. To achieve this one may consider an **equivalent inner product**

[8] Being *G-jointly equivariant* is essentially the same as being G-equivariant with respect to the actions: $G \circlearrowright (\mathcal{X}_1 \times \mathcal{X}_2)$ (via $T^{\mathcal{X}_1} \times T^{\mathcal{X}_2}$) and $G \circlearrowright \mathcal{Y}$ (via $T^{\mathcal{Y}}$)

[9] This shall be enough, as it's the most used setting in the literature. In particular, in the NN setting, usually the *features* and *labels* are respectively $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^c$. A potential extension to *group actions via isometries* (on simply *metric* $\mathcal{Z}$) could potentially be pursued in future works.

on $\mathcal{Z}$ given by $\langle a, b\rangle_G = \int_G \langle \rho_g a, \rho_g b\rangle_{\mathcal{Z}} d\lambda_G(g), \ \forall a, b \in \mathcal{Z}$. This makes ρ unitary and it is known (see Elesedy and Zaidi [28]) as the *Weyl trick*.

Remark After having introduced the main framework that will be considered to model *symmetries*, one could naturally notice the several *limitations* involved. Having *symmetries* encoded as *group actions* immediately implies that every transformation *must have an inverse*. In particular, transformations such as "introducing *noise* into an image" or "blocking parts of the image" (among many others), fall outside our current theoretical framework. In order to tackle these limitations without losing too many of the advantageous properties of group theory, some *variants* of the framework have emerged in the literature:

- **Partial Symmetries**: It accounts for the possibility of having the objects NOT being invariant for **every** group transformation, but rather just for *subset* of them.

  e.g. the detection of a *digit* on an image is not a *rotationally invariant* task, as 6 and 9 may be confused under a 180° rotation. However, it is *partially* rotationally invariant, as we might only consider rotations of $< 90°$ to avoid confusions (and still profit from the symmetric properties).

- **Approximate Symmetries**: It accounts for the possibility of having the objects not be *perfectly* invariant for every group transformation. This means that the objects might be *invariant* up to a small *error*: e.g. *compressing and reconstructing* an image will leave it *invariant*, except for potentially a small error due to a lossy reconstruction (under *approximate* invariance, these differences can be deemed *not relevant*). We will develop this idea further in the following sections.

Despite the inherent interest of these variants of the framework, we will not consider them into any more detail, leaving their exploration (particularly in our setting) as future work. The interested reader might look at Petrache and Trivedi [69] for some deeper analysis.

3.2 Theory of Invariant Measures

To develop our theory, we will assume that our data is **symmetric** in some sense. It is therefore highly relevant to understand how exactly the **symmetries** with respect to a group can be encoded into probability laws. Most of the notions described here are extracted from Kallenberg [44, 45] and Bloem-Reddy and Teh [7].

3.2.1 Base Results

Consider G to be a *lcsH* group, with (left)-Haar measure λ_G, that acts *measurably*[10] on a space $\mathcal{Z}$. We start by recalling some key definitions; such as that of a kernel (as described in Kallenberg [45]):

[10] Recall that this corresponds to the map degining the action, $(g, x) \in G \times \mathcal{Z} \mapsto g.x \in \mathcal{Z}$ being measurable.

Definition 3.4 (Kernel) *Let $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$ be two measurable spaces. A* ***kernel*** *φ from E_1 to E_2 (usually denoted $\varphi : E_1 \to E_2$) is a function $\varphi : E_1 \times \mathcal{E}_2 \to [0, \infty]$ such that:*

- $\forall A \in \mathcal{E}_2,\ x \in E_1 \mapsto \varphi(x, A)$ *is a* ***measurable*** *function.*
- $\forall x \in E_1,\ \varphi(x, \cdot)$ *is a* ***measure*** *on $(E_2, \mathcal{E}_2)$.*

Equivalently, one may define a kernel as being a measurable function $\varphi : E_1 \to \mathcal{M}(E_2)$ (i.e. such that it associates every point in E_1 with a measure over $(E_2, \mathcal{E}_2)$).

Also recall the definition of the **pushforward** of a measure:

Definition 3.5 *Let $(E_1, \mathcal{E}_1, \mu)$ be a probability space and $(E_2, \mathcal{E}_2)$ be a measurable space. Consider a measurable function $T : (E_1, \mathcal{E}_1, \mu) \to (E_2, \mathcal{E}_2)$; then, the* ***pushforward measure*** *of μ by T, $T_{\#}\mu \in \mathcal{P}(E_2, \mathcal{E}_2)$, is defined as:*

$$(T_{\#}\mu)(C) = \mu(T^{-1}(C)),\ \forall C \in \mathcal{E}_2$$

Following from this definition, we get (extending from simple functions to integrable functions) the well-known *change of variables formula*:

Lemma 6 *For any $f : E_2 \to \mathbb{R}$, $f \in L^1(E_2, \mathcal{E}_2, T\#\mu) \iff f \circ T \in L^1(E_1, \mathcal{E}_1, \mu)$ In that case, we have:*

$$\int_{E_2} f(y) d(T\#\mu)(y) = \int_{E_1} f(T(x)) d\mu(x)$$

A direct corollary following from this is that:

Lemma 7 *Let $\mu \in \mathcal{P}(\mathcal{Z})$ and let Z be a random variable with distribution μ. Consider $T : \mathcal{Z} \to \tilde{\mathcal{Z}}$ a measurable map, and consider the random variable $\tilde{Z} := T(Z)$, whose law we denote ν. Then, $\nu = T\#\mu$*

Considering the group action $G \circlearrowright_M \mathcal{Z}$, we define (by slightly abusing notation):

- $\forall g \in G,\ M_g : z \in \mathcal{Z} \mapsto M_g.z \in \mathcal{Z}$
- $\forall z \in \mathcal{Z};\ T_z : g \in G \mapsto M_g.z \in \mathcal{Z}$

We say a set $A \in \mathcal{B}_{\mathcal{Z}}$ is G-invariant if $\forall g \in G,\ M_g^{-1}A = A$; and we denote the σ-field of G-invariant sets as $\mathcal{I}_{\mathcal{X}}^G$. We say a **measure** $\mu \in \mathcal{M}(\mathcal{Z})$ is G-invariant if $\forall g \in G,\ M_g\#\mu = \mu$; and we denote the set of all G-invariant measures over $\mathcal{Z}$ as $\mathcal{M}^G(\mathcal{Z})$. i.e.

$$\mathcal{M}^G(\mathcal{Z}) := \{\mu \in \mathcal{M}(\mathcal{Z})\ :\ \forall g \in G,\ M_g\#\mu = \mu\}$$

Analogously, we say that a kernel $\varphi : \mathcal{Z} \to \tilde{\mathcal{Z}}$ is G-**invariant** if

$$\forall g \in G,\ \forall z \in \mathcal{Z},\ \varphi_{M_g.z} = \varphi_z \circ \tilde{M}_g^{-1}$$

For readability, we will consider G to be a **compact** group as we present the following results. More generality may be found in Kallenberg [45].

The following definition is key for understanding how any G-invariant measure is characterized.

Definition 3.6 (Orbit Measure) *Given a **compact** group G (with normalized Haar measure λ_G) acting on a space $\mathcal{Z}$, we define the **orbit measure kernel** $\varphi : \mathcal{Z} \to \mathcal{Z}$ as:*

$$\forall z \in \mathcal{Z}, \quad \varphi_z := \lambda_G \circ T_z^{-1}$$

Remark Notice that this **orbit measure kernel** satisfies $\forall z, \tilde{z} \in \mathcal{Z}$:

- φ_z is a σ-finite **probability** measure.
- Whenever $Gz = G\tilde{z}$ (they are in the same orbit) $\varphi_z = \varphi_{\tilde{z}}$
- Whenever $Gz \neq G\tilde{z}$ (different orbits) $\varphi_z \perp \varphi_{\tilde{z}}$
- φ_z *concentrates* on Gz.
- $\varphi : \mathcal{Z} \to \mathcal{Z}$ is a G-invariant kernel, as it satisfies:

$$\forall z \in \mathcal{Z}, \ \forall g \in G : \varphi_z \circ M_g^{-1} = \varphi_z = \varphi_{M_g . z}$$

Intuitively, this means that for any $z \in \mathcal{Z}$, φ_z is a **uniform** probability distribution on the **orbit** Gz (therefore the name of **orbit measure**).

These **orbit measures** allow us to establish (following Kallenberg [45]) the renowned **ergodic decomposition theorem** for invariant measures:

Theorem 10 (Theorem 7.3 from Kallenberg [45]) *Let G be a **compact** group acting on $\mathcal{Z}$ and φ the corresponding orbit measure kernel. Let $\nu \in \mathcal{M}(\mathcal{Z})$ be a σ-finite measure, then:*

$$\nu \in \mathcal{M}^G(\mathcal{Z}) \iff \exists \mu \in \mathcal{M}(\varphi(\mathcal{Z})), \sigma\text{-finite s.t. } \nu = \int_{\varphi(\mathcal{Z})} m d\mu(m)$$

*In this case, the measure μ is **unique** and it satisfies:*

$$\forall f : \mathcal{M}(\mathcal{Z}) \to \mathbb{R}_+ \text{ measurable,} \int f d\mu = \int f(\varphi_x) d\nu(x)$$

$$\text{In particular, } \nu = \int_{\mathcal{Z}} \varphi_x(\cdot) d\nu(x) = \int_{\varphi(\mathcal{Z})} m d\mu(m)$$

In short, this theorem tells us that **any invariant measure over $\mathcal{Z}$ can be seen as a *mix* of orbit measures**. Other key base results that will be relevant in what follows are related to Invariant Disintegrations and Radon-Nikodym derivatives:

First, the following *disintegration* theorem:

Theorem 11 (Theorem 7.6 from Kallenberg [45]) *Let G be a measurable group with Haar measure λ_G acting measurably on $\mathcal{Z}$ and $\mathcal{Y}$, where $\mathcal{Y}$ is Borel. If $\eta \in \mathcal{M}^G(\mathcal{Z} \times \mathcal{Y})$ and $\nu \in \mathcal{M}^G(\mathcal{Z})$ are two σ-finite measures such that $\eta(\cdot \times \mathcal{Y}) \lll \nu$; then there exists a G-invariant kernel $\varphi : \mathcal{Z} \to \mathcal{Y}$ such that: $\eta = \nu \otimes \varphi$. i.e. $\forall f : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}_+$ measurable, we have:*

$$\int_{\mathcal{Z}\times\mathcal{Y}} f(x,y) d\eta(x,y) = \int_{\mathcal{Z}} \int_{\mathcal{Y}} f(x,y) d\varphi_x(y) d\nu(x)$$

On the other hand, we can characterize G-invariant measures that are absolutely continuous with respect to a G-invariant measure as thos that have a **G-invariant density function**.

Theorem 12 (Theorem 7.8 from Kallenberg [45]) *Let $G \circlearrowright \mathcal{Z}$ measurably and let $\mu, \nu \in \mathcal{M}^G(\mathcal{Z})$ be σ-finite. Then:*

$$\mu \lll \nu \text{ on } \mathcal{B}_{\mathcal{Z}} \iff \mu \lll \nu \text{ on } \mathcal{I}_{\mathcal{X}}^G \iff \exists h : \mathcal{Z} \to \mathbb{R}_+ \text{ } G\text{-invariant and measurable} : \ h = \frac{d\mu}{d\nu}$$

3.2.2 Applicability in the Learning Framework

Consider that our data arrives following a fixed distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Our underlying assumption will be that the data **is symmetric** in some sense. We will encode this through the G-invariance of the joint law.

Definition 3.7 ***[Invariant Laws]*** *Let G be a compact group acting on $\mathcal{X}$ and $\mathcal{Y}$ via representations ($\rho^{\mathcal{X}}$ and $\rho^{\mathcal{Y}}$ respectively)*

- *Recall that, given a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we say that μ is **G-invariant** if $\forall g \in G, \ \mu = \rho_g^{\mathcal{X}} \# \mu$.*
  In particular, if X is a random variable with law μ, this can also be written as:

  $$X \overset{(d)}{=} \rho_g^{\mathcal{X}}.X \ \forall g \in G$$

  *and X is said to be **G-invariant in law**.*

- *Similarly, a probability measure $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is said to be (jointly) G-invariant when it is so with respect to the joint action $G \circlearrowright (\mathcal{X} \times \mathcal{Y})$ via $\rho^{\mathcal{X}} \times \rho^{\mathcal{Y}}$.*
  In particular, if a r.v. (X, Y) has law $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, then

  $$(X, Y) \overset{(d)}{=} (\rho_g^{\mathcal{X}}.X, \rho_g^{\mathcal{Y}}.Y) \ \forall g \in G$$

  *which means that the pair (X, Y) is **jointly G-invariant in law** (or also, G-**equivariant in law**). As a consequence, each marginal ($\pi|_{\mathcal{X}}$ and $\pi|_{\mathcal{Y}}$) is G-invariant as well.*

Remark The work by Bloem-Reddy and Teh [7] extensively characterizes the notions of *equivariance in law* and other similar concepts that are fundamental in the context of GDL. In particular, the following result is shown:

Theorem 13 (from Bloem-Reddy and Teh [7]) *If G is a compact group acting measurably on $\mathcal{X}$ and $\mathcal{Y}$ (Borel spaces); and also, there exists a measurable* ***representative equivariant*** *$\tau : \mathcal{X} \to G$ (i.e. measurable G-equivariant function, $\forall g \in G$, $\tau(g.x) = g.\tau(x)$). Let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and suppose $\pi|_{\mathcal{X}}$ is G-invariant. Then: π is (jointly) G-invariant* ***if and only if***

$$\exists f : [0,1]\times\mathcal{X} \to \mathcal{Y} \ \text{ measurable } G\text{-equivariant s.t. } (X,Y) \overset{a.s.}{=} (X, f(\eta, X)),\ \eta \sim \mathcal{U}([0,1]) \perp X$$

In other words, understanding the law of the pair π can always be *reduced* to approximating a G-equivariant function, modulo uniform and independent noise.

They also delve into the relevance of **maximal invariant** statistics[11] for determining good properties of *equivariant* distributions. Remarkably, for τ a measurable **representative equivariant**, $M_\tau : \mathcal{X} \to \mathcal{X}$ defined by $M_\tau(x) = \tau(x)^{-1}.x$ is a **maximal invariant** (see Lemma 8 in [7][12]).

As we're working over vector spaces, we would like to adapt this *noise outsourcing* characterization of theorem 13 into an *additive* version. Fortunately, we can always do that in the setting where $\pi \in \mathcal{P}_2^G(\mathcal{X} \times \mathcal{Y})$:

Proposition 13 *Let G be a compact group acting measurably on $\mathcal{X}$ (Borel space) and $\mathcal{Y}$ (separable Hilbert space). Let $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ be (jointly) G-invariant and such that $\mathbb{E}_\pi[\|Y\|^2] < \infty$. Then:*

$$\exists f^* : \mathcal{X} \to \mathcal{Y} \ \text{ measurable } G\text{-equivariant s.t. } (X,Y) \overset{a.s.}{=} (X, f^*(X) + \xi)$$

where ξ is a centered r.v. with finite variance, and such that $\forall h : \mathcal{X} \to \mathcal{Y}$ measurable, $\mathbb{E}[\langle \xi, h(X)\rangle_{\mathcal{Y}}] = 0$.

Proof. This proof is based on rather standard arguments. It is however original (as far as we know), so we include it entirely in section C.7. □

Remark Actually, the core of proposition 13 is proving that $f^*(x) = \mathbb{E}[Y|X = x]$ must be a ($\pi_{\mathcal{X}}$-a.e.) G-equivariant function.

With these ideas in mind, when dealing with data that is jointly G-invariant in law, a natural idea would be to seek a way to *leverage* the symmetries in network training. The following section contains the key idea of *model symmetrization* that appears vastly in the literature.

3.3 Symmetrization of Models

Under the setting of **group representations** of a compact group G, any map can be turned into a G-invariant/equivariant map via the following *symmetrization* operator:

[11] Which are statistics $M : \mathcal{X} \to \mathcal{S}$ (with $\mathcal{S}$ some borel space) such that $M(g.x) = M(x)$ $\forall x \in \mathcal{X}$, $\forall g \in G$ and $\forall x, \tilde{x} \in \mathcal{X}$, whenever $Gx = G\tilde{x}$, $M(x) = M(\tilde{x})$

[12] They also show a way of *easily* constructing G-equivariant maps from a representative equivariant: for any $f : \mathcal{X} \to \mathcal{Y}$ just define $\tau(x).f(\tau(x)^{-1}.x)$

Definition 3.8 (**Symmetrization (Orbit Averaging)**) *Let G be a* ***compact*** *group of (normalized) Haar measure λ_G, such that $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{X}$ and $G \circlearrowright_{\rho^{\mathcal{Y}}} \mathcal{Y}$. Let $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ be the set of all* ***measurable*** *maps from $\mathcal{X}$ to $\mathcal{Y}$; and $\mathcal{M}^G(\mathcal{X}, \mathcal{Y}) \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ the set of those that are G-equivariant.*

We define the following symmetrization operator $\mathcal{Q} : \mathcal{M}(\mathcal{X}, \mathcal{Y}) \to \mathcal{M}^G(\mathcal{X}, \mathcal{Y})$ by its action over any measurable $f : \mathcal{X} \to \mathcal{Y}$:

$$(\mathcal{Q}f)(x) = \int_G \rho^{\mathcal{Y}}_{g^{-1}}.f(\rho^{\mathcal{X}}_g.x)d\lambda_G(g)$$

By the G-invariance of λ_G, the resulting function $(\mathcal{Q}f)$ is exactly G-equivariant (with respect to the corresponding actions).

In particular, for the invariant case ($\rho^{\mathcal{Y}} \equiv \mathrm{Id}_{\mathcal{Y}}$), the same operator is denoted by $\mathcal{S}$ (and defined as: $(\mathcal{S}f)(x) = \int_G f(\rho^{\mathcal{X}}_g.x)d\lambda_G(g)$, which always yields a G-invariant map)[13].

Proposition 14 (Properties of the Symmetrization Operator (from Elesedy and Zaidi [28])) *Consider a* ***compact*** *group G that acts on $\mathcal{X}$ and $\mathcal{Y}$ (via $\rho^{\mathcal{X}}$ and $\rho^{\mathcal{Y}}$ respectively). Let $\pi_{\mathcal{X}}$ be a measure on $\mathcal{X}$.*

1. *$\forall f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ we have that: f is G-equivariant $\iff$ $\mathcal{Q}f = f$.*
2. *$\mathcal{Q}$ has two eigenvalues: 0 and 1.*

Further assume that $\rho^{\mathcal{Y}}$ is a ***unitary*** *representation and $\pi_{\mathcal{X}}$ is G-invariant. Then:*

3. *Whenever $f \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$, we have that $\mathcal{Q}f \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$.*
4. *$\mathcal{Q}$ is self-adjoint.*
   i.e. $\mathcal{Q}$ is the ***orthogonal projection from*** *$L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$* ***onto*** *$L^2_G(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$*

These properties of $\mathcal{Q}$ allow for the following decomposition lemma (from Elesedy and Zaidi [28]):

Lemma 8 (Equivariant Decomposition from Elesedy and Zaidi [28]) *Let G be a* ***compact*** *group acting on $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{X}$ and $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{Y}$ (with $\rho^{\mathcal{Y}}$* ***unitary****); and let $\pi_{\mathcal{X}}$ be a G-invariant measure on $\mathcal{X}$.*

For any $U \leqslant L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ closed under $\mathcal{Q}$, U admits an ***orthogonal decomposition into symmetric and antisymmetric parts****: $U = \mathcal{I}_U \bigoplus \mathcal{I}_U^{\perp}$, where $\mathcal{I}_U := \{f \in U \ : \ f \text{ is } G\text{-equivariant}\}$ and $\mathcal{I}_U^{\perp} := \{f \in U \ : \ \mathcal{Q}f = 0\}$*

Having noticed that the symmetrization operator $\mathcal{Q}$ acts as an orthogonal projection; it is natural to ask **how much one might win from such symmetrization**. As in Elesedy and Zaidi [28], we can define:

[13] Strictly speaking, the action of G over $\mathcal{X}$ **doesn't need to be linear**: the operator shall be well defined (i.e. yielding a G-equivariant measurable function) as long as at least the action on $\mathcal{Y}$ is kept linear.

Definition 3.9 (Symmetrization Gap (from Elesedy and Zaidi [28])) *Let G be a **compact** group acting on $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{X}$ and $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{Y}$ (with $\rho^{\mathcal{Y}}$ **unitary**); and let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Define, for any $f \in L^2(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})$, the **generalization gap** as:*

$$\Delta(f, \mathcal{Q}f) := \mathbb{E}_\pi[\ell(f(X), Y)] - \mathbb{E}_\pi[\ell((\mathcal{Q}f)(X), Y)]$$

It quantifies the difference in population risk between a model and its symmetrized version.

From this definition, Elesedy and Zaidi [28] prove the following result (under the quadratic loss):

Lemma 9 (Symmetrization Gap Characterization (from Elesedy and Zaidi [28])) *Let G be a **compact** group acting on $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{X}$ and $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{Y}$ (with $\rho^{\mathcal{Y}}$ **unitary**); consider the **quadratic loss** and let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be such that $\pi|_{\mathcal{X}}$ is G-invariant. Also, assume that there exists $f^* : \mathcal{X} \to \mathcal{Y}$ G-equivariant such that $(X, Y) \sim \pi$ satisfies: $Y = f^*(X) + \xi$ with ξ centered of finite variance and independent of X. Then, the generalization gap satisfies:*

$$\Delta(f, \mathcal{Q}f) := \mathbb{E}_\pi[\|Y - f(X)\|_{\mathcal{Y}}^2] - \mathbb{E}_\pi[\|Y - (\mathcal{Q}f)(X)\|_{\mathcal{Y}}^2] = \|f^\perp\|_{\pi|_{\mathcal{X}}}^2$$

We can actually *improve* this result and make it more general by using proposition 13 and drawing inspiration from a recent paper by Huang et al. [40]:

Lemma 10 (Symmetrization Gap Characterization) *Let G be a **compact** group acting on $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{X}$ and $G \circlearrowright_{\rho^{\mathcal{X}}} \mathcal{Y}$ (with $\rho^{\mathcal{Y}}$ **unitary**); consider the **quadratic loss** and let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be such that $\mathbb{E}_\pi[\|Y\|^2] < \infty$.*

Assume that $\pi|_{\mathcal{X}}$ is G-invariant, but π is only H-invariant with respect to some $H \leqslant G$ (closed). Then, the generalization gap satisfies:

$$\Delta(f, \mathcal{Q}_G f) := \mathbb{E}_\pi[\|Y - f(X)\|_{\mathcal{Y}}^2] - \mathbb{E}_\pi[\|Y - (\mathcal{Q}_G f)(X)\|_{\mathcal{Y}}^2] = -2\langle f^*, f_G^\perp \rangle_{L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})} + \|f_G^\perp\|_{L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})}^2$$

where $f^*(x) = \mathbb{E}_\pi[Y | X = x]$ is the conditional expectation function of Y wrt X, and we denote $f_G^\perp := f - \mathcal{Q}_G f$.

In particular, if π is G-invariant as well, we get $\Delta(f, \mathcal{Q}_G f) = \|f_G^\perp\|_{L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})}^2$

Proof. As far as we know, the result (as is stated, in its most general form) hasn't been proven in the literature. The proof is, however, really similar to that employed in Elesedy and Zaidi [28] and Huang et al. [40] for their corresponding statements. We do, however, draw from proposition 13 in order to state the result for an arbitrary π that's square integrable. Our proof of the statement shall be found in section C.8. □

Remark What Lemma 10 esentially tells us is that, if we *try to symmetrize* a model with respect to a group that has *"more symmetries"* than what are actually observable in our data (i.e. π in itself is only H-invariant, but we symmetrize with respect to $G \geqslant H$); we can either *win* or *lose* generalization power according to the interplay between the two presented terms. In particular, if π is G invariant, there's a *strict generalization benefit* from choosing a *symmetric model* to tackle our *learning problem* (which gives the name to the paper Elesedy and Zaidi [28]). In particular, whenever $f_G^\perp$ is non-zero (on a strictly positive $\pi|_\mathcal{X}$-measure set) there's a **strict gain** in generalization power from using the symmetrized version of the model.

Beyond these theoretical elements, the focus of the related literature has been, in general, on *leveraging* the symmetries of the data in some way during training, in order to get better models. Model *symmetrization* is one key alternative, but many others exist.

3.4 Techniques for Leveraging Symmetries

There are three main objects of study in the literature when it comes to leveraging the *G-invariance* of data: **Data Augmentation** (**DA**), **Feature Averaging** (**FA**), and more recently, the use of **Equivariant Architectures** (**EA**). The first two techniques have been extensively studied in the context of neural network learning (see Lyle et al. [53], Chen et al. [14], Huang et al. [39] for a comparison between both, and Mei et al. [59], Li et al. [52], Dao et al. [21] for applications in the context of NTK). On the other hand, **equivariant architectures** of NNs have been studied on their own right (see Kondor and Trivedi [47], Cohen et al. [19], Maron et al. [55], Yarotsky [92], Zaheer et al. [93], Weiler and Cesa [90], Wood and Shawe-Taylor [91], Shawe-Taylor [76, 77] among many others). Comparisons between all three methods are scarce and their appearence in the setting of the Mean Field Limit of NNs is practically non-existent. This work, in part, aims to discover how these different techniques can be compared and how they might influence the MF limit of NNs.

Recall that, in our learning setting, as described in section 2.1, we have i.i.d. data coming with a distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and we try to find a model $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ such that $R(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$ will be minimized (where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a **loss** function that's often assumed **convex**). Let $\Phi_\theta^{N,\sigma} : \mathcal{X} \to \mathcal{Y}$ be a NN model with architecture (N, σ) and parameter $\theta \in \Theta$ (as described in section 2.1.1). The different techniques considered in the literature amount to different approaches to this problem:

3.4.1 Data Augmentation (DA)

Data augmentation is one of the most used *regularization techniques* when training NN models. It is, in part, due to the ease of implementation, but also due to its flexibility (the applied transformations may not have any kind of *group* structure).

Under this setting we do not modify the architecture of the NN. Instead, we **penalize it** during training, in order to *teach* it the symmetries of the problem. It is thus a *trained* equivariance (NOT necessarily enforced). In this setting we talk about a *generic* model $f : \mathcal{X} \to \mathcal{Y}$ which could, in particular, be one of NNs as described above.

Essentially, we train the NN to **minimize an *averaged* version of the risk**:

$$R^G(f) = \int_G \mathbb{E}_{(X,Y)\sim\pi}[\ell(f(\rho_g^\mathcal{X}.X), \rho_g^\mathcal{Y}.Y)]\mathrm{d}\lambda_G(g)$$

where λ_G is the normalized Haar measure of G. Actually, when $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is assumed (jointly) G-invariant $R^G(f) = R(f)$ so there's *no difference* between both problems.

In practice, **DA** consists in *symmetrizing* the **empirical risk** (that's what we actually optimize during training); and so, for a given sample of i.i.d. data $(X_k, Y_k)_{k=1}^B$, instead of optimizing $\hat{R}(f) = \frac{1}{B}\sum_{k=1}^B \ell(f(X_k), Y_k)$, we optimize the **augmented** empirical risk:

$$\hat{R}^G(f) = \frac{1}{B}\sum_{k=1}^{B}\int_G \ell(f(\rho_g^{\mathcal{X}}.X_k), \rho_g^{\mathcal{Y}}.Y_k)d\lambda_G(g)$$

In practice, as most of the time integrating over G is intractable, we also *approximate this* using an i.i.d. sample of group elements $((g_i^k)_{i=1}^{B_G})_{k=1}^B \sim \lambda_G$; we optimize:

$$\hat{R}^G_{B_G}(f) = \frac{1}{B\,B_G}\sum_{k=1}^{B}\sum_{i=1}^{B_G}\ell(f(\rho_{g_i^k}^{\mathcal{X}}.X_k), \rho_{g_i^k}^{\mathcal{Y}}.Y_k)$$

Despite the fact that this technique has *no guarantee* of obtaining an *equivariant* model at the end of training (it only *penalizes* models intelligently) it is still highly relevant in practical applications.

Regarding *theoretical results* about this method, Chen et al. [14] dives deeply into establishing a group theoretic framework in which the advantages of **DA** with respect to other techniques can be studied (Lyle et al. [53] provides complementary results that shall also be considered).

The main result from these papers is that the average over G in the augmented risk serves to reduce the variance of the ERM estimator compared to a NN without such augmentation. Consider the following lemma from Chen et al. [14]:

Lemma 11 (Exact Invariance Lemma from Chen et al. [14]) *Let G be a compact group with Haar measure λ_G and $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ a G-invariant law (with $\rho^{\mathcal{Y}}$ the trivial action). Let $h : \mathcal{X} \to \mathcal{Y}$ be an arbitrary measurable function, such that $[(x, g) \mapsto (h(g.x))] \in L^2(\pi|_{\mathcal{X}} \times \mu_{\mathcal{G}})$. Define $h^G := (\mathcal{S}h)$ as the orbit average of h. Then:*

1. *$\forall x \in \mathcal{X}$, $h^G(x) = \mathbb{E}[h(Z)|Z \in p(x)]$ i.e. the symmetrized model is given by the conditional expectation of h given the corresponding orbit*[14].
2. *By the law of total expectations, we get that: $\mathbb{E}_{X\sim\pi|_{\mathcal{X}}}[h(X)] = \mathbb{E}_{X\sim\pi|_{\mathcal{X}}}[h^G(X)]$*
3. *By the law of total variance, the covariance of $h(X)$ is as follows:*

$$\boldsymbol{Var}_{X\sim\pi|_{\mathcal{X}}}[h(X)] = \boldsymbol{Var}_{X\sim\pi|_{\mathcal{X}}}[h^G(X)] + \mathbb{E}_{X\sim\pi|_{\mathcal{X}}}\left[\boldsymbol{Var}_{g\sim\lambda_G}[h(g.X)]\right]$$

4. *For any convex function $\varphi : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}_{X\sim\pi|_{\mathcal{X}}}[\varphi(h(X))] \geqslant \mathbb{E}_{X\sim\pi|_{\mathcal{X}}}[\varphi(h^G(X))]$*

From here, a simple, yet insightful, proposition follows, depicting how the *augmented risk* is (in general) a better approximator of the population risk than the *non-augmented* version:

[14] This is what, in the literature, is usually referred to as the *Rao-Blackwellization* of estimators in this context.

Proposition 15 (Improvement of Invariance, from Lyle et al. [53]) *If π is jointly G-invariant and $f : \mathcal{X} \to \mathcal{Y}$ is such that $\ell(f(\cdot), \cdot) \in L^2(\pi)$ (also, suppose that ℓ is jointly G-invariant); then:*

- *The augmented empirical risk is unbiased with respect to the normal empirical risk: $R^G(f) = \mathbb{E}_{(X_k,Y_k)_{k=1}^B}[\hat{R}^G(f)] = \mathbb{E}_{(X_k,Y_k)_{k=1}^B}[\hat{R}(f)] = R(f)$ (i.e. it is unbiased)*
- *The variance of the augmented risk is lower than the non-augmented one.*

$$\boldsymbol{Var}_{(X_k,Y_k)_{k=1}^B}[\hat{R}^G(f)] \leqslant \boldsymbol{Var}_{(X_k,Y_k)_{k=1}^B}[\hat{R}(f)]$$

All in all, this means that the augmented empirical risk gives a better estimate of the real population risk than the non-augmented one (so, one should expect a more effective minimization).

Deeper results concerning **DA** have been found over the years. For example, some PAC-Bayes bounds have been established in this context (see Lyle et al. [53]). Many other interesting results concerning **DA** exist, and we won't be able to cover them all in this work: the original material shall be sought for deeper insights.

3.4.2 Feature Averaging (FA)

Feature Averaging is another quite popular technique for *leveraging symmetries* in our NN models. In short, **FA** consists simply on **symmetrizing** the original model $f : \mathcal{X} \to \mathcal{Y}$ in order to obtain an explicitly equivariant model.

In the case of a NN, $\Phi_\theta^{N,\sigma}$, the **feature-averaged** model corresponds to considering

$$\Phi_\theta^{N,\sigma,\mathbf{FA}}(x) := \int_G \rho_{g^{-1}}^{\mathcal{Y}}.\Phi_\theta^{N,\sigma}(\rho_g^{\mathcal{X}}.x) d\lambda_G(g) = (\mathcal{Q}\Phi_\theta^{N,\sigma})(x)$$

where λ_G is the normalized Haar measure of G.

On the one hand, **FA** *ensures* that the obtained model will be *equivariant* (as shown in proposition 14). Unfortunately, this comes at the cost of being **higly inefficient** in terms of the number of parameters/computations to be performed. No *reduction in the number of parameters* is being performed nor is the symmetry being *encoded* in the architecture in any way. It is rather an arbitrary model that is being *forced* into being G-equivariant by averaging over all possible transformations of the input.

Despite this observation, the following result by Elesedy and Zaidi [28] shows that **FA** is, in some sense, the *best possible way* of *forcing* the model to be G-equivariant:

Proposition 16 (**FA** as Least Squares, from Elesedy and Zaidi [28]) *For any $f \in L^2(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})$, **FA** with $\mathcal{Q}$ maps f to $\overline{f}$, the ($\pi|_{\mathcal{X}}$-a.e.) unique solution to the **least squares** problem:*

$$\overline{f} = \arg \min_{s \in \mathcal{I}_{L^2(\mathcal{X},\mathcal{Y};\pi|_{\mathcal{X}})}} \|f - s\|_{L^2(\mathcal{X},\mathcal{Y};\pi|_{\mathcal{X}})}^2$$

On the other hand, similar to the case of **DA**, *variance reduction* results have been established:

Proposition 17 (Prop.5 from Lyle et al. [53]) *If π is G-invariant and $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function in its first argument that's also jointly G-invariant. Then, for all $f : \mathcal{X} \to \mathcal{Y}$,*

$$\hat{R}_\ell(\mathcal{Q}f) = \hat{R}^G(\mathcal{Q}f) \leqslant \hat{R}_G(f)$$

i.e. The risk of any symmetrized model is lower than that of the original one.

Moreover, if $\ell(f(\cdot), \cdot) \in L^2(\pi)$; then: $\boldsymbol{Var}_{(X_k,Y_k)_{k=1}^B}[\hat{R}(\mathcal{Q}f)] \leqslant \boldsymbol{Var}_{(X_k,Y_k)_{k=1}^B}[\hat{R}^G(f)]$

*i.e. the variance of the **FA** risk estimator is lower than that of the **DA** one. This amounts simply to stating that through **FA** a better estimate of the population risk shall be found (and hopefully this will help the minimization process).*

Beyond this, Lyle et al. [53] establish a quantitative *advantage* (in terms of the KL divergence of their PAC-Bayes bound) as well as quantify the *symmetrization gap* between **FA** and non **FA** models. A similar result is established in Elesedy and Zaidi [28], in which they prove that the *Rademacher complexity*[15] of a class of models $\mathcal{F}$ is *larger* than the *complexity* of the class containing only the *symmetrized* version of the models in $\mathcal{F}$; furthermore, this *advantage* is quantifiable in terms of the *complexity* of the class of *antisymmetrized* versions of models in $\mathcal{F}$.

In some specific cases, Elesedy and Zaidi [28] are able to quantitatively calculate the **generalisation gap** from considering a G-equivariant model. For example, in the case of a **linear regression** model, where the predictors are of the form $f_W(x) = W^T x$ with $W \in \mathbb{R}^{d \times c}$, they are able to establish that:

$$\Delta(f_W, \mathcal{Q}f_W) = \left\| \sqrt{\mathbb{E}_{X\sim\pi|\mathcal{X}}[XX^\top]} W^\perp \right\|_F^2$$

where $W^\perp$ is the *orthogonal* component of W with respect to the *intertwining average*[16]. We notice from this particular example that, in the linear setting, doing **FA** amounts, esentially, to considering an *equivariant "architecture"* for the linear model (as we'll see in the following section, many elements of the theory resemble this particular example).

In general, many other results exist for **FA** models, and further insights might be sought in the extense literature.

Remark What's particularly been a recurrent interesting question in the literature is about the *comparison* between **DA** and **FA** as "Symmetry-Leveraging mechanisms". Works like Chen et al. [14] and Lyle et al. [53] delve into this comparison and establish that, at least theoretically, **FA** has an advantage over **DA** when the loss function ℓ is convex; also, both methods are better than *not using anything* other than the original NN model.

[15] For deeper insight into its definition and the formal statement of the theorem, we advise to look into Elesedy and Zaidi [28]

[16] The *intertwining average* is defined as $W^G := \int_G \rho_g^{\mathcal{X}} W \rho_{g^{-1}}^{\mathcal{Y}} d\lambda_G(g)$, so that $W^\perp = W - W^G$

However, these results rely *merely* on the convexity of the loss function, and better approximations to the problem have been made over the years. For instance, Dao et al. [21] prove that **DA** corresponds, in a first order approximation, to doing **FA**; and in a second order approximation, to minimizing a *variance-regularized* version of the objective. In a similar fashion, Li et al. [52] prove that, in the context of **kernel classifiers**, optimizing a kernel model using an *augmented dataset* (i.e. **DA**) will yield the exact same result as optimizing a *symmetrized* model on the original dataset (i.e. **FA**). In particular, in the context of kernel classifiers, an explicitly invariant model is achieved through **DA**, despite there being no *explicit constraint* to ensure this. This fact is also noted by Lyle et al. [53] when studying linear regression models under **DA**. All in all, **DA** and **FA** aren't so different, at least in the context of *linear* (or *linearized*) regressions.

Finally, the *symmetry-leveraging technique* that has been most popularized in recent years (following the ideas of Bronstein et al. [10]) is that of **equivariant architectures** (**EA**) which we'll review next.

3.4.3 Equivariant Architectures (EA)

As mentioned in previous chapters, *Equivariant Architectures (EA)* such as CNNs, Transformers, GNNs, PointNets, among many others; have been gaining an *increasing popularity* over recent years, specially due to their great success at *introducing an inductive bias* on the network architecture, allowing the model to *profit* from the data's symmetry (e.g. CNNs on *traslation-invariant* images; GNNs on *permutation-invariant* graphs, etc.). In short, *equivariant architectures* assume that a certain set of symmetries to hold, and they are *built* in such a way that allows to *simplify* the model (usually, through imposing *parameter sharing* within each layer) withouth damaging the model's generalization power.

Following the formalism of Flinth and Ohlsson [32], let's say we can define a multilayer NN model as $\Phi_A^L : \mathcal{X}_0 := \mathcal{X} \to \mathcal{X}_1 \to \mathcal{X}_2 \to \cdots \to \mathcal{X}_L =: \mathcal{Y}$ such that:

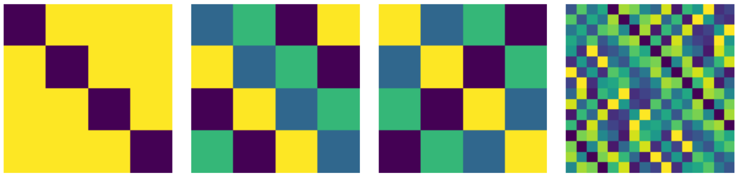
$$x_0 = x,\ x_{i+1} = \sigma_i(A_i x_i)\ \forall i \in \{0, \dots, L-1\},\ \Phi_A(x) = x_L$$

Where $\forall i \in \{0, \dots, L-1\}$, $A_i : \mathcal{X}_i \to \mathcal{X}_{i+1}$ are the linear maps between the hidden vector spaces $(\mathcal{X}_i)_{i=0}^L$, and $\sigma_i : \mathcal{X}_{i+1} \to \mathcal{X}_{i+1}$ are the (non-linear) activation functions.

Let $\mathbf{Hom}(\mathcal{X}_i, \mathcal{X}_{i+1})$ be the set of *linear maps* between $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ and consider $\mathcal{L} = \Pi_{i=0}^{L-1}\mathbf{Hom}(\mathcal{X}_i, \mathcal{X}_{i+1})$. Then, the network is *exactly parameterized* by $A = (A_i)_{i=0}^{L-1} \in \mathcal{L}$; that is, knowing the parameters of each linear layer (i.e. the *matrix's parameters*) is enough to reconstruct the action of the network.

Now consider a compact group G such that $G \circlearrowright_{\rho_i} \mathcal{X}_i$ for all $i \in \{0, \dots, L\}$ (such that $\rho_0 = \rho_{\mathcal{X}}$ and $\rho_L = \rho_{\mathcal{Y}}$). We say that an NN model has an *equivariant architecture* (**EA**; or simply, that it is an *equivariant NN*) whenever *each intermediate layer is an equivariant map* $\mathcal{X}_i \to \mathcal{X}_{i+1}$. In other words, we impose that

$$\rho_{i+1}(g).x_{i+1} = \sigma_i(A_i.\rho_i(g).x_i),\ \forall g \in G,\ \forall i \in \{0, \dots, L-1\}$$



(a) S_4 (b) $\mathbb{Z}_4$ (c) $\mathbb{Z}_2^2$ (d) $\mathbb{Z}_4 \ltimes \mathbb{Z}_2^2$

Figure 3.3: Illustration of a *basis* in the matrix space $\mathbf{Hom}_G$ for different groups. Cells of the same color are essentially *shared* parameters. This figure was taken from Finzi et al. [31].

Consider the space $\mathbf{Hom}_G(\mathcal{X}_i, \mathcal{X}_{i+1})$, which corresponds to all *G-equivariant linear maps* between $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ (i.e. $A_i : \mathcal{X}_i \to \mathcal{X}_{i+1}$ such that $\rho_{i+1}(g).A_i x = A_i \rho_i(g).x,\ \forall g \in G,\ \forall x \in \mathcal{X}_i$). Clearly, if all activation functions σ_i are *G-equivariant*[17]; and each linear layer A_i lives in $\mathbf{Hom}_G(\mathcal{X}_i, \mathcal{X}_{i+1})$; the resulting model will be a NN with **EA** (i.e. each layer is an equivariant map). In particular, we can consider what we refer to as the *equivariant parameter space*, given by:

$$\mathcal{E}^G = \Pi_{i=0}^{L-1} \mathbf{Hom}_G(\mathcal{X}_i, \mathcal{X}_{i+1})$$

This allows to simplify our description, since whenever the parameter vector A lives in $\mathcal{E}^G$ (and the activations are all G-equivariant), the resulting NN will be *equivariant*. The *space of equivariant parameters* has been widely studied in the Geometric Deep Learning literature, seeking to understand its *universality* properties (see Yarotsky [92], Zaheer et al. [93], Maron et al. [55]) and simpler ways to characterize it (see Cohen et al. [19], Kondor and Trivedi [47], Weiler and Cesa [90], Aronsson [3], Lang and Weiler [49]).

In particular, it is shown that $\mathcal{E}^G$ corresponds to the space of all mappings that can be written as a *group convolution* against kernels with good equivariance properties (see theorems 3.2, 3.3, and 3.4 of [19], theorem 1 of [47], or theorem 4.1 of [49])[18].

Another way of understanding **EA**s is by seeing the equivariant linear maps as simple *matrices* where some of their entries are shared (see Wood and Shawe-Taylor [91], Ravanbakhsh et al. [71], Finzi et al. [31]). That is, the "*group convolution*" that takes place from one layer to the next, is achieved with a single matrix multiplication (where various of the matrix parameters are repeated). As we're working with *group representations*; solving the constraint $\rho_{i+1}(g).A_i\rho_i(g^{-1}) = A_i\ \forall g \in G,\ \forall i \in \{0, \ldots, L-1\}$ amounts to solving a simple linear system (as described in Finzi et al. [31]). Figure 3.3 illustrates how this *parameter sharing* looks like for matrices under different symmetry groups. This characterization of G-equivariant linear maps explicitly shows that equivariant networks have *fewer* parameters than their *fully-connected* counterparts (as multiple parameters are *shared* within the same layer). On the other hand, this description allows us to *embed* an equivariant NN into an *ambient space*

[17] A simple (and widely used) example of such activation functions are the *coordinate-wise* application of some *real* activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

[18] e.g. Cohen et al. [19] essentially says that $\mathbf{Hom}_G(V_1, V_2)$ (for V_1 and V_2 two vector spaces) is *isomorphic* (under the right homogeneity conditions) to the space of *bi-equivariant kernels of G* (which build the *convolution*):

$$\mathcal{K}_G = \{\kappa : G \to \mathrm{Hom}(V_1, V_2)\ :\ \forall g \in G,\ \forall h_1 \in H_1,\ \forall h_2 \in H_2 :\ \kappa(h_2 g h_1) = \rho_2(h_2)\kappa(h)\rho_1(h_1)\}$$

corresponding to that of a fully-connected NN: the equivariant architecture lives in a closed linear subspace ($\mathcal{E}^G$) of the whole *parameter space* ($\mathcal{L}$). This idea will ease the "translation" of the setting of **EA** to the MF study of NNs.

As to how equivariant architectures are implemented in practice (in a *flexible* way, allowing or many potentially different groups), Finzi et al. [30] and Finzi et al. [31] shed some light into how we can characterize $\mathcal{E}^G$ by a linear system (which can be solved using tensor algebra) and also how an **EA** for a possibly infinite group might be approximated (e.g. implementing *group convolutions* through Monte Carlo sampling [30]). Recent works such as Puny et al. [70] also try to *simplify* the *group averaging* procedure by only applying it over a G-equivariant *frame* (subset of all group elements).

The idea of finding the parameters of $\mathcal{E}^G$ by solving a linear system is also tackled in Flinth and Ohlsson [32]. Particularly, from our formulation of **EA** networks we can define, from the action of G on each $\mathcal{X}_i$, an action of G on $\mathbf{Hom}(\mathcal{X}_i, \mathcal{X}_{i+1})$. Indeed, for all $A_i \in \mathbf{Hom}(\mathcal{X}_i, \mathcal{X}_{i+1})$ and $g \in G$, we consider:

$$\bar{\rho}_i(g).A_i := \rho_{i+1}(g)A_i\rho_i(g)^{-1}$$

From here, we can construct a G-action on the entire parameter space, $\mathcal{L}$, by simply defining:

$$\forall A \in \mathcal{L} : \ \bar{\rho}(g).A := (\bar{\rho}_i(g).A_i)_{i=0}^L$$

In other words, from actions on the *latent spaces* of the network, we can generate an action on *the network itself* (via its parameter $A = (A_i)_{i=0}^L \in \mathcal{L}$). Furthermore,

Proposition 18 (Characterization of $\mathcal{E}^G$ (from Finzi et al. [31] and Flinth and Ohlsson [32])) *Let $G \circlearrowright_{\rho_i} \mathcal{X}_i$, $\forall i \in \{0, \dots, L\}$ and define $G \circlearrowright_{\bar{\rho}} \mathcal{L}$ as in the previous paragraph. Then:*

$$A \in \mathcal{E}^G \iff \forall g \in G; \ \bar{\rho}(g).A = A$$

In other words: $\mathcal{E}^G = \{A \in \mathcal{L} \ : \ \forall g \in G; \ \bar{\rho}(g).A = A\}$.

This simple fact will be crucial for our study of **EA** networks in the following sections (specially related to how we introduce them in the MF limit of shallow NNs). One can immediately notice (as the *linear* representations of G acting on each underlying latent space are continuous; and thus $\bar{\rho}$ is continuous as well), that $\mathcal{E}^G = \bigcap_{g \in G} \mathrm{Ker}(\bar{\rho}(g) - \mathrm{Id}_{\mathcal{L}})$ is a *closed linear subspace* of $\mathcal{L}$ (as it is an arbitrary intersection of *closed linear subspaces*). We can then define *the orthogonal projection $P_{\mathcal{E}^G}$* onto $\mathcal{E}^G$: consider λ_G the unique *normalized* Haar measure on G (a *compact group*), and notice that the projection has the form:

$$\begin{aligned} P_{\mathcal{E}^G} : \mathcal{L} &\to \mathcal{E}^G \\ A &\mapsto \int_G \bar{\rho}(g).A d\lambda_G(g) \end{aligned}$$

Proof. See Flinth and Ohlsson [32]. For completeness, we include the proof in Annex C.9. □

One might notice the resemblance with the *intertwining average* described in the *linear regression* example from section 3.4.2. Actually, Elesedy and Zaidi [28] go beyond such an example, and try to tackle the case of *shallow NNs* using similar techniques. In particular, consider a *single-layer NN* ($\Phi_W(x) := \sigma(Wx)$) with σ being C-Lipschitz and applied *pointwise* such that $\rho^{\mathcal{Y}}$ commutes with σ. They can prove (for $\mu = \mathcal{N}(0, \mathrm{Id}_d)$) that:

$$\inf_{s \in \mathcal{I}_{L^2(\mathbb{R}^d, \mathbb{R}^c; \mu)}} \mathbb{E}_{X \sim \mu}[\|\Phi_W(X) - s(X)\|_2^2] \leqslant 2C^2 \|W^\perp\|_F^2$$

i.e. the closest possible G-equivariant model from the original Φ_W is at a distance bounded by the *anti-symmetric* component of the weights (suggesting that $\|W^\perp\|_F^2$ might an interesting *penalization term* to include during the *training* of neural networks).

On a different branch, Lawrence et al. [50] study the training dynamics of *linear*[19] G-equivariant models; analogously, Mei et al. [59] bring certain ideas from *invariant* networks into the context of random features and NTK (where they are able to extract some quantitative guarantees). In general, new advancements in the theory of **EA**s are constantly being made by researchers; however, despite the heavy interest in the literature for such models, the comparison between **DA, FA**, and the use of **EA** has not been too extensively developed.

Very recently, Flinth and Ohlsson [32] have studied how the (continuous-time) training dynamics of network models with equivariant architectures behave, comparing stationary points and local/global minima for different training strategies (in particular, **DA**, **EA** and the *vanilla* training). Notably, beyond the **DA** and **FA** *risk functionals* that people usually try to minimize (R^G and R^{FA} respectively), they introduce the *equivariant risk* of a model, as (for $R : \mathcal{L} \to \mathbb{R}$ the *original* risk functional)

$$R^{EA}(A) = R(P_{\mathcal{E}^G}.A) \;\; \forall A \in \mathcal{L}$$

which would be the *risk* associated to a model if its parameters were to be *projected* onto the space of explicitly equivariant parameters. In their paper, Flinth and Ohlsson [32] prove that (under reasonable assumptions) $\forall A \in \mathcal{E}^G$ we have $\nabla R^G(A) = P_{\mathcal{E}^G} \nabla R(A) = \nabla R(P_{\mathcal{E}^G}.A)$; implying that the space $\mathcal{E}^G$ is invariant under the *gradient flow* of R^G. Similarly, it allows them to relate the stationary points of the dynamic of R^{aug} to those of R^{eqv}.

As we'll see in upcoming sections, similar results will arise in the study of the MF limit of shallow NNs. In general, the goal of our work is to bridge the ideas from the world of *equivariance* in neural networks to the Mean Field context, aiming to understand the impacts of the different techniques considered (**DA**, **FA**, or **EA**) on the properties of the MF limit.

Remark Many other *symmetry-leveraging* techniques are continuously being proposed and exploited in the deep learning literature, far beyond the scope of our review. A remarkable example that recently appeared in the literature is the idea of *canonicalization* introduced in Kaba et al. [42]. It esentially establishes an architecture divided in *two steps*: first, a *canonicalization* function, that tries to place the input image into a *canonical orientation* (similar to how the *representative equivariant* of Bloem-Reddy and Teh [7] worked); then a simple unconstrained NN able to perform the task over *canonically-oriented* input data; and finally the *canonicalization's "inverse"* in output space (to make the model *equivariant*). This idea has a really interesting potential (as it has not been extensively studied in the literature), so it shall be looked upon in future work. Kim et al. [46] also provide interesting

[19] i.e. without activation functions, a really commonly used proxy for studying NNs.

variants of the same idea; and many other techniques shall be found in the ever-growing literature on NNs.

Chapter 4

Symmetry in the study of “shallow” Neural Networks

In this section, we will address some of the specifics of truly *studying symmetries for the MF Limit* of *shallow* NNs.

Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Hilbert Spaces (most commonly, for real-world NNs, we will take $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^c$; we'll however define the ideas in general). Let's assume that some *compact* group G acts on our *feature and label spaces*; i.e. $G \circlearrowright_{\rho} \mathcal{X}$ and $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$. We can suppose these representations to be *orthogonal*[1] (as in Mei et al. [59], Flinth and Ohlsson [32]; as already discussed, given that G is compact, this assumption is not truly restrictive). When no confusion is possible, for ease of notation we will write $\rho_g := \rho(g)$ for all relevant group representations.

Suppose we have data $(X_k, Y_k)_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \pi$, and further assume that its law, π, is *jointly G-invariant*; i.e. for $(X, Y) \sim \pi$: $(X, Y) \overset{(d)}{=} (\rho_g.X, \hat{\rho}_g.Y)\ \ \forall g \in G$

Based on this action of G on the data space $\mathcal{X} \times \mathcal{Y}$, we want to be able to understand some properties of the associated Mean Field limit (compared to the cases without such symmetry). As the most widely studied case in the literature is that of *shallow* NNs, we will begin by analyzing this type of model.

Recall that, *archetypically*, a *shallow* NN can be written as follows: let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$, and define the *shallow* NN to be:

$$\begin{aligned} \Phi_\theta^N : \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x; \theta_i) \end{aligned}$$

where $\sigma_* : \mathcal{X} \times \mathbb{R}^D \to \mathcal{Y}$ is the so-called *activation function* of the network, considered to be $\sigma_*(x; \theta_i) = w_i \sigma(A_i \cdot x + b_i)$, with $\theta_i = (w_i, A_i, b_i) \in \mathbb{R}^D := \mathbb{R}^{d+2}$ and $\theta := (\theta_i)_{i=1}^N \in (\mathbb{R}^D)^N$.

[1] As mentioned in section 2.2 many *interesting examples admit a unitary representation* ($G = \mathcal{S}_n$ on $\mathcal{X} = \mathbb{R}^n$, $G = \mathcal{S}_n$ on $\mathcal{X} = \mathbb{R}^{n^k}$, $G = \mathbb{Z}_n^2$ or $G = C_4$ (rotations) on $\mathcal{X} = \mathbb{R}^{n\times n}$, among many others).

4.1 Defining Symmetries in the Parameter Space

In terms of *equivariant NNs*, the *archetypical shallow NN* doesn't allow for very *interesting* **EA**. Such a network is esentially a function $\Phi_\theta^N : \mathbb{R}^d \to \mathbb{R}^N \to \mathbb{R}$, and so, as not many orthogonal representations are available on $\mathbb{R}$, this won't lead to very interesting equivariant models. On the other hand, as the *exchangeability* of the N hidden units of the NN is a *crucial* condition when taking the MF limit (or else, there wouldn't be a *propagation of chaos*), the *only meaningful action of G on $\mathbb{R}^N$* has to be the *trivial one* ($\rho_{\mathbb{R}^N} \equiv \mathrm{Id}_N$)[2]. In part due to this, **EA** networks in this setting get easily reduced to being trivial. As we will see later, if we take $G = O(d)$ acting on $\mathbb{R}^d$ via matrix multiplication and trivially on $\mathbb{R}$, then any **EA** will have $A_i = 0,\ \forall i \in \{1, \ldots, N\}$ and thus *the output of the NN won't depend on the input at all.* We are therefore in need of a more *interesting* definition of a *shallow* neural network in order to make the MF limit also interesting from the viewpoint of **EA**s.

4.1.1 Making the Parameter Space Interesting

An idea for making the group action on the space of parameters *more interesting* is to consider a *higher dimensional label space* and a *more complex* intermediate (or *hidden*) space. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^c$. Let $b \in \mathbb{N}$ and $G \circlearrowright_\eta \mathbb{R}^b$ (i.e. we also have a representation of G on $\mathbb{R}^b$ given by η). Given $N \in \mathbb{N}$, we will now have $N \cdot b$ hidden units, in a NN that will be of the form (using the notation of Flinth and Ohlsson [32]): $\Phi_B : \mathbb{R}^d \to \mathbb{R}^{bN} \to \mathbb{R}^c$ with $B = (W, A) \in \mathcal{L} := \mathcal{M}_{c\times(bN)}(\mathbb{R}) \times \mathcal{M}_{(bN)\times d}(\mathbb{R})$, such that[3]:

$$\Phi_B^N(x) := \frac{1}{N} W\sigma(AX)$$

Where we can write this expression *by blocks*

$$W = (W_1, \ldots, W_N), \ \text{ and } \ A = \left(\begin{array}{c} A_1 \\ \hline \vdots \\ \hline A_N \end{array}\right)$$

with $(W_k, A_k) \in \mathbb{R}^{c\times b} \times \mathbb{R}^{b\times d}$. This translates to:

$$\Phi_B^N(x) := \frac{1}{N}\sum_{k=1}^N W_k \sigma(A_k X)$$

To simplify the notation of the parameters, we will consider $W_k = w_k \in \mathbb{R}^{c\times b}$ and $A_k = a_k^T$, with $a_k \in \mathbb{R}^{d\times b}$; so that we can define a *parameter* $\theta_k := (w_k, a_k) \in \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} = \mathbb{R}^{(c+d)\times b} \cong \mathbb{R}^{(c+d)b} =: \mathbb{R}^D$, we will denote the *vectorized* version of parameters as $(\vec{w}_k, \vec{a}_k) \in \mathbb{R}^{cb} \times \mathbb{R}^{db}$.

[2] Otherwise, parameters building an equivariant network under such an action would be *shared* between hidden units (making them *not exchangeable* and yielding propagation of chaos impossible).

[3] Assuming, for simplicity, that there is no bias and that the second layer is *normalized* by the number of units N

We define $\theta = (\theta_k)_{k=1}^N$, and this allows us to write the network (using the typical *shallow NN* notation) as:

$$\Phi_\theta^N(x) = \frac{1}{N}\sum_{k=1}^N w_k\sigma(a_k^T x) := \frac{1}{N}\sum_{i=1}^N \sigma_*(x;\theta_i)$$

where we have defined an *activation function* $\sigma_* : \mathcal{X} \times \mathbb{R}^D \to \mathcal{Y}$ as $\sigma_*(x;\theta_k) = w_k\sigma(a_k^T x)$, with $\theta_k := (w_k, a_k) \in \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} \cong \mathbb{R}^D$. The advantage of writing the network in this format is that the *transition to the mean field limit* should happen without major issues.

4.1.2 Making a (somewhat) interesting group representation

In the MF limit of the network, the evolution of a *typical parameter* $\tilde{\theta} \in \mathbb{R}^D$ is studied. Given that π is (jointly) G-invariant, we would like to draw conclusions about this *typical* parameter; for instance, whether its distribution μ *inherits* from π's G-invariance in some way. To answer this, it becomes necessary to extrapolate, from the actions on $\mathcal{X}$ and $\mathcal{Y}$, an *action* $G \circlearrowright \mathbb{R}^D$ (or equivalently, on $\mathcal{L}$) that is consistent with the behaviour of the data.

The natural definition for such an action $G \circlearrowright \mathcal{L}$ was given in the previous chapter (see proposition 18; based on the ideas of [32, 31]); but it requires having some action of G on every hidden layer as well (i.e. on $\mathbb{R}^{bN}$ in our case).

The solution that will (for the time being) allow us to obtain a result *without significantly modifying the ideas from the MF context* will be to set the action $G \circlearrowright_{\tilde{\rho}} \mathbb{R}^{Nb}$ by the (orthogonal) representation $\tilde{\rho}_\cdot := \mathrm{Id}_N \otimes \eta_\cdot$ (where $\otimes$ is the Kronecker product and Id_N is the identity matrix of $\mathbb{R}^N$). The tensorization by Id_N is there to ensure the required *exchangeability* between the NN hidden units in order to take the MF limit. In other words, it means that $\tilde{\rho}$ will act independently on each separate hidden *unit* (which will now be a *block* of size b in $\mathbb{R}^{Nb}$ thanks to our new definition). i.e. if we have $Z = (Z_1, \ldots, Z_N) \in \mathbb{R}^{b\times N}$ and we consider it's vectorization $\vec{Z} = \left(\begin{array}{c} Z_1 \\ \hline \vdots \\ \hline Z_N \end{array}\right) \in \mathbb{R}^{Nb}$; then, by properties of the Kronecker product:

$$\tilde{\rho}_g\vec{Z} = (\mathrm{Id}_N \otimes \eta_g)\vec{Z} = \mathrm{vec}(\eta_g Z \mathrm{Id}_N) = \mathrm{vec}(\eta_g Z) = \mathrm{vec}((\eta_g Z_1, \ldots, \eta_g Z_N)) = \left(\begin{array}{c} \eta_g Z_1 \\ \hline \vdots \\ \hline \eta_g Z_N \end{array}\right)$$

This means that $\tilde{\rho}$ *acting* on $\mathbb{R}^{bN}$ consists of η acting on each *block component* independently[4]. A similar calculation also works for matrices in $\mathcal{M}_{c\times(bN)}(\mathbb{R})$.

The idea of *not* mixing up the NN's *units* with the intermediate group action aims at having a single common space $\mathcal{Z} = \mathbb{R}^D$ (shared among all units, as with the traditional MF limit) where the behaviour of a *type* parameter might be studied. In particular, it is precisely what's needed for the expression $\Phi_\theta^N(x) := \frac{1}{N}\sum_{k=1}^N w_k\sigma(a_k^T X)$ to make sense as $N \to \infty$, allowing us to interpret it as a **LLN** for the *exchangeable units* (see theorem 3). The study

[4] Note that when $b = 1$, this network architecture corresponds to the usual *shallow NN* from section 2.2. In that case, η is an action on $\mathbb{R}$, which we'll assume reduces to the trivial action: $\eta = 1$ and $\tilde{\rho} := \mathrm{Id}_N$

of more interesting group actions over $\mathbb{R}^N$ is definitely an open problem to tackle as future work.

Despite this limitation, as we've *expanded* the usual setting of *shallow NNs*, (with all our *hidden units* living in a *more interesting* space), we are able to consider relatively *interesting* equivariant network architectures, such as *shallow and (very) wide* CNNs, DeepSets, GNNs, among others[5].

Equipped with these *actions* $G \circlearrowright_{\rho} \mathbb{R}^d, G \circlearrowright_{\tilde{\rho}} \mathbb{R}^{Nb}$, and $G \circlearrowright_{\hat{\rho}} \mathbb{R}^c$, we can define (as in [32]), for $(W, A) \in \mathcal{L}$:

$$\overline{\rho}_0(g).A = \tilde{\rho}_g A \rho_g^{-1} = (\mathrm{Id}_N \otimes \eta) A \rho_g^T = \left(\begin{array}{c} \eta_g A_1 \\ \hline \vdots \\ \hline \eta_g A_N \end{array}\right) \rho_g^T = \left(\begin{array}{c} \eta_g A_1 \rho_g^T \\ \hline \vdots \\ \hline \eta_g A_N \rho_g^T \end{array}\right)$$
$$\overline{\rho}_1(g).W = \hat{\rho}_g W \tilde{\rho}_g^{-1} = \hat{\rho}_g W (\mathrm{Id}_N \otimes \eta_g)^T = \hat{\rho}_g (W_1 \eta_g^T, \dots, W_n \eta_g^T) = (\hat{\rho}_g W_1 \eta_g^T, \dots, \hat{\rho}_g W_n \eta_g^T)$$

This leads to the *complete* action of G over the set of parameters of the network as:

$$\overline{\rho}_g.(W, A) = (\overline{\rho}_1(g).W, \overline{\rho}_0(g).A)$$

Now, given the this action behaves *independently* on *each unit block* of the A and W, we can extend $\overline{\rho}$ to each of these components $(W_k)_{k=1}^N$ and $(A_k)_{k=1}^N$ of W and A respectively[6]. For each $k \in \{1, \dots, N\}$, $\forall g \in G$:

$$\begin{aligned} \overline{\rho}(g).(W_k, A_k) &= (\hat{\rho}_g W_k \eta_g^T, \eta_g A_k \rho_g^T) \\ &= (\hat{\rho}_g w_k \eta_g^T, (\rho_g a_k \eta_g^T)^T) \end{aligned}$$

This definition is consistent with the one above, since the actions of $\overline{\rho}_0(g)$ and $\overline{\rho}_1(g)$ are naturally *by blocks*.

Remembering that $\forall k \in \{1, \dots, N\}$, $w_k \in \mathbb{R}^{c \times b}$ and $a_k \in \mathbb{R}^{d \times b}$, $\overline{\rho}_g$ actually defines an action over $\mathbb{R}^{(c+d) \times b}$ (here, we use block notation to describe $\begin{pmatrix} w_k \\ a_k \end{pmatrix} \in \mathbb{R}^{(c+d) \times b}$):

$$\overline{\rho}_g . \begin{pmatrix} w_k \\ a_k \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_k \eta_g^T \\ \rho_g a_k \eta_g^T \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g & 0 \\ 0 & \rho_g \end{pmatrix} \begin{pmatrix} w_k \\ a_k \end{pmatrix} \eta_g^T = \hat{M}_g \begin{pmatrix} w_k \\ a_k \end{pmatrix} \eta_g^T$$

where we've defined the block matrix: $\hat{M}_g := \begin{pmatrix} \hat{\rho}_g & 0 \\ 0 & \rho_g \end{pmatrix} \in \mathcal{M}_{(c+b) \times (c+b)}(\mathbb{R})$.

By *vectorizing* our parameter $\theta_k = \begin{pmatrix} w_k \\ a_k \end{pmatrix}$ into $\vec{\theta_k} \in \mathbb{R}^D$, we see that the action of $\overline{\rho}_g$, thanks to the properties of $\otimes$, corresponds to:

$$\overline{\rho}_g \vec{\theta_k} = \mathrm{vec}(\hat{M}_g \theta_k \eta_g^T) = (\eta_g \otimes \hat{M}_g) \vec{\theta_k} := M_g \vec{\theta_k}$$

[5] This setting makes sense as long as there's a *linear unit* that's *infinitely repeatable*, and thus interchangeable (such as a single convolution kernel that could potentially have *infinitely many channels*). This are referred to as networks with *tensor order* 1 in Maron et al. [55].

[6] This is why it's *natural* to consider $G \circlearrowright_{\tilde{\rho}} \mathbb{R}^{bN}$ with $\tilde{\rho}_. := \mathrm{Id}_N \otimes \eta_.$; so that each *fundamental unit* $\theta_k \in \mathbb{R}^D$ will be acted upon in the *same way* by G. Possibly, for a more *interesting* action $\tilde{\rho}$, a more intricate network structure will be needed.

i.e. the acion of G over $\mathbb{R}^D$ corresponds to multplying with a matrix of the form $M_g := \eta_g \otimes \hat{M}_g$.

We can see, in particular, that this action $\overline{\rho}$ over $\mathbb{R}^D$ is orthogonal. This is due to the fact that $\hat{M}_g \in O(c+d)$, as can be seen from the following calculation:

$$\hat{M}_g \hat{M}_g^T = \begin{pmatrix} \hat{\rho}_g & 0 \\ 0 & \rho_g \end{pmatrix} \cdot \begin{pmatrix} \hat{\rho}_g^T & 0 \\ 0 & \rho_g^T \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g \hat{\rho}_g^T & 0 \\ 0 & \rho_g \rho_g^T \end{pmatrix} = \mathrm{Id}_{c+d}$$

On the other hand, by properties of the Kronecker product, we have:

$$M_g M_g^T = (\eta_g \otimes \hat{M}_g)(\eta_g \otimes \hat{M}_g)^T = (\eta_g \otimes \hat{M}_g)(\eta_g^T \otimes \hat{M}_g^T) = (\eta_g \eta_g^T) \otimes (\hat{M}_g \hat{M}_g^T) = \mathrm{Id}_b \otimes \mathrm{Id}_{(c+d)} = \mathrm{Id}_D$$

so that indeed, $\forall g \in G,\ M_g \in O(D)$. i.e. $G \circlearrowright_M \mathbb{R}^D$ orthogonally.

After this calculations, we've managed to *succesfully construct an action* $G \circlearrowright_M \mathbb{R}^D$ (via an orthogonal representation). With this in place, we can now determine whether the *type* parameter, in the MF limit, will inherit some *interesting symmetries* from the G-invariance of the data distribution.

First, as noted at the end of Chapter 3, the action of G on the *parameter space* that we have defined (following [32, 31]) leads to the following characterization:

Definition 4.1 (Space of *Equivariant* Neural Network Parameters) *Let $G \circlearrowright_M \mathbb{R}^D$, we define:*

$$\mathcal{E}^G := \{\theta \in \mathbb{R}^D \ : \ \forall g \in G,\ M_g.\theta = \theta\}$$

This is the space of NN parameters that are left fixed under the action of G via M.

In particular, the following characterization will be key for our purposes in upcoming sections:

Proposition 19 *For any $\theta = (\theta_i)_{i=1}^N \in (\mathbb{R}^D)^N$, we have that:*

$$\Phi_\theta^N : \mathbb{R}^d \to \mathbb{R}^{bN} \to \mathbb{R}^c \text{ defines a } G\text{-equivariant NN} \iff \forall i \in \{1, \dots, N\},\ \theta_i \in \mathcal{E}^G$$

Proof. The proof's direct from the definition of the group actions employed. In any case, a simple proof is presented in Annex C.10. □

4.1.3 Generalization is key

Both previous sections provide a key motivation for what will follow: starting from the most *down-to-earth* example of a *shallow NN*, we arrived at a characterization (proposition 19) of *equivariant shallow NNs* that merely depends on some *equivariant subspace* $\mathcal{E}^G$ (and, in particular, doesn't heavily rely on the underlying description of the network)[7].

This allows us to actually generalize the ideas, hiding the (potential) complexity of the NN definition (in terms of layers, activation functions, pooling, truncation, etc.) behind the

[7] We consider such motivation as a key piece to understanding our work and its applicability; that's why we include both sections in the body of our work, rather than relegating them to an annex

all-purpose activation function $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ that's usually considered in the Mean Field literature.

We could then simply start by considering some (orthogonal) representation M of a compact group G acting on some space $\mathcal{Z}$ (a separable Hilbert Space we will refer to as the *parameter* space; in the previous example we had $\mathcal{Z} = \mathbb{R}^D$); i.e. $G \circlearrowright_M \mathcal{Z}$. We can then define, as in definition 4.1, the *equivariant parameter space* for our networks as:

$$\mathcal{E}^G := \{\theta \in \mathcal{Z} \ : \ \forall g \in G, \ M_g.\theta = \theta\}$$

Then, we can directly *define* the notion of an *equivariant shallow NN*:

Definition 4.2 (General Definition of an Equivariant Neural Network) *Let a compact group G act on a separable Hilbert space $\mathcal{Z}$ through an orthogonal representation $G \circlearrowright_M \mathcal{Z}$. Let $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be an activation function. We will say $f : \mathcal{X} \to \mathcal{Y}$ is a G-equivariant shallow NN if:*

$$f := \Phi_\theta^N = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(\cdot;\theta_i)$$

For some $N \in \mathbb{N}$ and $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$ such that $\forall i \in \{1, \ldots, N\}$, $\theta_i \in \mathcal{E}^G$. Coherently with what we introduced in section 2.2.2, we will denote the set of equivariant shallow NNs as $\mathcal{N}_{\sigma_*}(\mathcal{E}^G)$.

As seen in Chapter 2, any *shallow* neural network can be seen as an integral against the *empirical measure* of the parameters; i.e. for $\theta \in \mathcal{Z}^N$ consider $\nu_\theta^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$, and we have $\Phi_\theta^N(x) = \langle \sigma_*(x;\cdot), \nu_\theta^N \rangle$. Particularly, in this context we get the following characterization:

Proposition 20 *Φ_θ^N is G-equivariant $\iff$ ν_θ^N concentrates in $\mathcal{E}^G$ (i.e. $\nu_\theta^N(\mathcal{E}^G) = 1$).*

In the *more general context* of minimizing in the space of probability measures (as seen in Chapter 2), we will be particularly interested in considering measures that *concentrate* on $\mathcal{E}^G$ (a notion we will develop further in upcoming sections).

Just as we saw in section 3.4.3, $\mathcal{E}^G = \bigcap_{g \in G} \mathrm{Ker}(M_g - \mathrm{Id}_{\mathcal{Z}})$ is a *closed linear subspace* of $\mathcal{Z}$, and *the orthogonal projection $P_{\mathcal{E}^G}$* onto $\mathcal{E}^G$ can be explicitly defined as:

$$\begin{aligned} P_{\mathcal{E}^G} : \mathcal{Z} &\to \mathcal{E}^G \\ \theta &\mapsto \int_G M_g.\theta d\lambda_G(g) \end{aligned}$$

where λ_G is the unique *normalized* Haar measure on G.

Considering this definition, a first question one may ask is whether these *equivariant NNs* (as in definition 4.2) are themselves *equivariant models*. The affirmative answer comes under the assumption that σ_* *respects* the symmetry in some sense:

Proposition 21 *Let G be a compact group such that: $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_M \mathcal{Z}$ and $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$. Let $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be a jointly G-equivariant function. Then $\forall \theta \in (\mathcal{E}^G)^N$,*

$$\Phi_\theta^N \ \ \textit{is a G-equivariant function.}$$

More generally, $\forall \mu \in \mathcal{P}(\mathcal{E}^G)$, $f = \int_{\mathcal{Z}} \sigma_*(\cdot;\theta)d\mu(\theta)$ is a G-equivariant function.

Proof. We can provide a proof of the first part only requiring the definitions we've already introduced (and it shall be found on section C.10). The second part requires elements that will be introduced in upcoming sections. □

We might then wonder whether the condition of having σ_* being jointly G-equivariant is at all *reasonable*. Fortunately, for the example presented in section 4.1.2 this is indeed the case:

Proposition 22 *Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Z} = \mathbb{R}^D = \mathbb{R}^{(c+d)b}$, $\mathcal{Y} = \mathbb{R}^c$ and let G be a compact group such that $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_\eta \mathbb{R}^b$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$. Consider the activation function of shallow NNs $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ given (for $\theta = (w,a)$) by $\sigma_*(x,\theta) = w\sigma(a^T x)$, with $\sigma : \mathbb{R}^b \to \mathbb{R}^b$ being G-equivariant to the action of η (e.g. in many cases it is enough for σ to be applied pointwise). Then:*

$$\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y} \quad \textit{is jointly } G\textit{-equivariant}$$

Proof. See Annex C.10. □

In particular, this allows to say that the previous proposition 19 is actually a consequence of proposition 22 and definition 4.2. Now, as hinted by the last part of proposition 21, the properties of *equivariant* neural networks come from the underlying *activation function σ_** and the *measures* involved in their definition.

In particular, beyond measures in $\mathcal{P}(\mathcal{E}^G)$, *any G-invariant measure $\mu \in \mathcal{M}^G(\mathcal{Z})$ yields a G-equivariant model $f = \int_{\mathcal{Z}} \sigma_*(\cdot;\theta)d\mu(\theta)$* (when σ_* is jointly G-equivariant; see proposition 37). It is then crucial for our purposes to understand the *interplay* between G-invariant probability measures and measures *concentrated in $\mathcal{E}^G$*. In the following section we will dive a bit deeper into the study of these two quite interesting measure spaces.

4.2 Wasserstein Spaces and Invariance

Recall the setting of Wasserstein Spaces described in section 2.2.4 (based on Santambrogio [73] and Ambrosio et al. [1]). Let $\mathcal{Z}$ be a Hilbert space with norm $\|\cdot\|$ (often simply $\mathcal{Z} = \mathbb{R}^D$). Consider $\mathcal{P}(\mathcal{Z})$ the space of probability measures over $\mathcal{Z}$ and, for $p \in \mathbb{N}^*$, $\mathcal{P}_p(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) \ : \ \int_{\mathcal{Z}} \|\theta\|^p \mu(d\theta) < +\infty\}$ the space of probability measures with finite p-th moment. This space, endowed with the p-th *Wasserstein metric*[8] W_p, is what we call the p-th Wasserstein Space.

From here, also consider a compact group G such that $G \circlearrowright_M \mathcal{Z}$ (via the representation M). Recall that $\mathcal{M}^G(\mathcal{Z})$ is the set of G-invariant measures over $\mathcal{Z}$; and define:

$$\mathcal{P}_p^G(\mathcal{Z}) := \{\mu \in \mathcal{P}_p(\mathcal{Z}) \ : \ \forall g \in G, \ M_g\#\mu = \mu\} = \mathcal{M}^G(\mathcal{Z}) \cap \mathcal{P}_p(\mathcal{Z})$$

[8] Which is defined as: $W_p(\mu,\nu) := \left[\inf_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_\gamma[\|X-Y\|^p]\right]^{\frac{1}{p}}$, $\forall \mu,\nu \in \mathcal{P}_p(\mathcal{Z})$ and $\Pi(\mu,\nu)$ the space of *couplings between μ and ν*

which is the set of G-invariant probability measures with finite p-th moment. We will be interested in *projecting* any measure in $\mathcal{P}(\mathcal{Z})$ onto $\mathcal{P}^G(\mathcal{Z})$; for this purpose, we introduce the *symmetrization* of a measure:

Definition 4.3 (**Symmetrization of a Measure**) *Let G be a compact group, with Haar measure λ_G, and such that $G \circlearrowright_M \mathcal{Z}$. For $\mu \in \mathcal{P}(\mathcal{Z})$, we define its symmetrization with respect to G, $\mu^G \in \mathcal{P}(\mathcal{Z})$, as:*

$$\forall B \in \mathcal{B}_{\mathcal{Z}},\ \mu^G(B) := \int_G \mu(M_g^{-1}(B))d\lambda_G$$

Or, equivalently, $\forall f : \mathcal{Z} \to \mathbb{R}$ positive and measurable:

$$\langle f, \mu^G \rangle = \int_G \langle f \circ M_g, \mu \rangle d\lambda_G(g)$$

Proposition 23 *$\forall \mu \in \mathcal{P}(\mathcal{Z})$, μ^G is a well defined probability measure over $\mathcal{Z}$ and satisfies:*

1. $\forall g \in G,\ \forall \mu \in \mathcal{P}(\mathcal{Z}),\ (M_g \# \mu)^G = \mu^G$
2. $\forall a, b \in \mathbb{R},\ \forall \mu, \nu \in \mathcal{P}(\mathcal{Z}),\ (a\mu + b\nu)^G = a\mu^G + b\nu^G$

Proof. See Annex C.11, or just see Kallenberg [44] for more about *intensity measures.* □

Remark Notice that μ^G is actually the *intensity measure* associated to the random measure $\omega \mapsto M_{g(\omega)}\#\mu$, where g is a G-valued random variable with law λ_G. It is thus naturally well defined. More information about this fact, and other measure-theoretic insights can be found in Kallenberg [45].

We will also be interested in considering probability measures that *concentrate* their weight on a fixed subspace of parameters. We will define, for any measurable $\mathcal{E} \subseteq \mathcal{Z}$, the set of measures that *concentrate* on $\mathcal{E}$ (and those with finite p-th moment)[9]:

$$\mathcal{P}^{\mathcal{E}}(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z})\ :\ \mu(\mathcal{E}) = 1\} = \mathcal{P}(\mathcal{E}), \quad \textit{and} \quad \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z}) := \mathcal{P}_p(\mathcal{Z}) \cap \mathcal{P}^{\mathcal{E}}(\mathcal{Z}) = \mathcal{P}_p(\mathcal{E})$$

Whenever $\mathcal{E}$ is a *closed* linear subspace of $\mathcal{Z}$, we can consider the orthogonal projection[10] onto $\mathcal{E}$, $P_{\mathcal{E}} : \mathcal{Z} \to \mathcal{E}$. For any $\mu \in \mathcal{P}(\mathcal{Z})$, we denote its pushforward by $P_{\mathcal{E}}$ as $\mu^{\mathcal{E}} := P_{\mathcal{E}}\#\mu$. We can get an interesting characterization of $\mathcal{P}^{\mathcal{E}}(\mathcal{Z})$ in terms of $P_{\mathcal{E}}$:

Lemma 12 *Let $\mu \in \mathcal{P}(\mathcal{Z})$ and $\mathcal{E}$ be a closed linear subspace of $\mathcal{Z}$ with an associated orthogonal projection $P_{\mathcal{E}} : \mathcal{Z} \to \mathcal{E}$. Then:*

$$\mu(\mathcal{E}) = 1 \iff \mu^{\mathcal{E}} := P_{\mathcal{E}}\#\mu = \mu$$

Proof. See Annex C.11. □

[9] We will interchangeably denote these spaces both by $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ and $\mathcal{P}_p(\mathcal{E})$; the first notation highlights the fact that $\mathcal{E}$ is part of a larger ambient space $\mathcal{Z}$, while the second one emphasizes the fact that these probability measures only assign mass to $\mathcal{E}$.

[10] Recall that the orthogonal projection is the *bounded operator* (continuous linear map) that satisfies $P_{\mathcal{E}}^2 = P_{\mathcal{E}}$ and $\langle P_{\mathcal{E}}x, y \rangle = \langle x, P_{\mathcal{E}}y \rangle$, $\forall x, y \in \mathcal{Z}$

In our case, this will be interesting in the setting of *equivariant architectures* for neural networks. Recall the definition of the set of *equivariant shallow NNs*:

$$\mathcal{E}^G := \{\theta \in \mathcal{Z} \ : \ M_g.\theta = \theta, \ \forall g \in G\}$$

This is the space of parameters that will constitute a G-equivariant shallow neural network. In particular, the space $\mathcal{P}_p^{\mathcal{E}^G}(\mathcal{Z})$ has all probability measures that only assign weight to explicitly G-equivariant architectures. Analogously, $\mathcal{P}^G(\mathcal{Z})$ contains all probability measures corresponding to *symmetrized* models of NNs. It is therefore clear that both $\mathcal{P}_p^{\mathcal{E}^G}(\mathcal{Z})$ and $\mathcal{P}_p^G(\mathcal{Z})$ will be fundamental to grasp a complete understanding of our problem.

We can now check some properties satisfied by these spaces (considering $\mathcal{E}$ a generic closed linear subspace of $\mathcal{Z}$). First, by noticing that the Wasserstein metric is a *convex function on its individual arguments.* i.e.

Proposition 24 *Given a fixed $\mu \in \mathcal{P}_p(\mathcal{Z})$, the function $\nu \in \mathcal{P}_p(\mathcal{Z}) \mapsto W_p(\nu, \mu)$ is convex and continuous. Also, whenever $\mathcal{Z} = \mathbb{R}^D$, $\mu \lll \lambda$ and $p > 1$, it is strictly convex.*

Finally, if $G \circlearrowright_M \mathcal{Z}$ orthogonally, the function $W_p : \mathcal{P}_p(\mathcal{Z}) \times \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ is jointly G-invariant (in the sense of Definition 4.4).

Proof. See Annex C.11. □

Proposition 25 *Let G be a compact group with Haar measure λ_G. Let $\mathcal{E}$ be a closed linear subspace of $\mathcal{Z}$ with orthogonal projection $P_{\mathcal{E}}$. Let $\mu \in \mathcal{P}_p(\mathcal{Z})$. We have that:*

1. *$\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ and $\mathcal{P}_p^G(\mathcal{Z})$ are closed (under the topology induced by W_p) and convex subspaces of $\mathcal{P}_p(\mathcal{Z})$.*
2. *$\mu^{\mathcal{E}} \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ and $\mu^G \in \mathcal{P}_p^G(\mathcal{Z})$*
3. *$\mu^{\mathcal{E}}$ is a projection of μ onto $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$; in the sense that it minimizes $W_p(\mu, \cdot)$ over $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$. If $\mathcal{Z} = \mathbb{R}^D$, $\mu \lll \lambda$ and $p > 1$, then it is the unique such projection.*
4. *$\mu \in \mathcal{P}_p^G(\mathcal{Z}) \iff \mu = \mu^G$ and $\mu \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z}) \iff \mu = \mu^{\mathcal{E}}$*

Proof. See Annex C.11. □

Remark Notice that, by the last point, a measure $\mu \in \mathcal{P}(\mathcal{Z})$ is G-invariant (i.e. $\forall g \in G,\ M_g\#\mu = \mu$) **iff** $\mu^G = \mu$. Therefore, it is *also equivalent* to the fact that[11]:

$$\exists G' \subseteq G \text{ with } \ \lambda_G(G') = 1 \text{ s.t. } \forall g \in G', \ M_g\#\mu = \mu$$

Remark Notice that we don't show that μ^G is *a projection of $\mu \in \mathcal{P}(\mathcal{Z})$ onto $\mathcal{P}^G(\mathcal{Z})$*. Unfortunately, the same techniques employed in proposition 25 for $\mu^{\mathcal{E}}$ don't work in the case of μ^G. Intermediate results such as proposition 23, proposition 24 and proposition 25 have emerged in our search for such proof. Unfortunately, for the time being we leave this to future work.

[11] Indeed, G-invariance clearly implies this last property; and whenever this property holds, then for any measurable $f : \mathcal{Z} \to \mathbb{R}$, $\langle f, \mu^G \rangle = \int_G \langle f, M_g\#\mu \rangle d\lambda_G(g) = \int_{G'} \langle f, M_g\#\mu \rangle d\lambda_G(g) = \int_{G'} \langle f, \mu \rangle d\lambda_G(g) = \langle f, \mu \rangle$

In particular, the measure spaces we're interested in satisfy:

Lemma 13 *Let G be a compact group with Haar measure λ_G, such that $G \circlearrowright_M \mathcal{Z}$; let $\mathcal{E}^G \subseteq \mathcal{Z}$ be the linear subspace of elements invariant to the action of G over $\mathcal{Z}$, and $P_{\mathcal{E}^G}$ be the orthogonal projection onto it. We have the following properties:*

- $\mathcal{P}^{\mathcal{E}^G}(\mathcal{Z}) \subseteq \mathcal{P}^G(\mathcal{Z})$
- $\forall \mu \in \mathcal{P}(\mathcal{Z}),\ \mu^{\mathcal{E}^G} = (\mu^G)^{\mathcal{E}^G} = (\mu^{\mathcal{E}^G})^G$

Proof. See Annex C.11. □

Remark Notice that we have provided some sort of *canonical* way of transforming any measure $\mu \in \mathcal{P}(\mathcal{Z})$, to make it satisfy the properties we desire: we can either make it G-invariant or *concentrated on* $\mathcal{E}^G$. An interesting question that immediately arises is about *how we can quantify the distance between μ and its transformed version.*

This is still an open question in our work (which we hope to tackle in the future); but for the moment we know the following estimate holds: In $\mathcal{Z} = \mathbb{R}^D$, with $\mu \lll \lambda$ and $p > 1$, as $W_p(\cdot, \mu)$ is *strictly convex* and *admits a first variation* (see Santambrogio [73], proposition 7.17), by a variant of *Jensen's inequality* (which we'll introduce in proposition 32), we know that:

$$W_p(\mu^G, \mu) \leqslant \int_G W_p(M_g\#\mu, \mu) d\lambda_G(g)$$

A similar estimate for $\mu^{\mathcal{E}^G}$ is yet to be found.

Now, one might also notice that *absolute continuity* plays a heavily important role in some of the previous results. In particular, one might be interested in looking at some properties related to the *densities* of measures in our spaces of interest. Consider $\mathcal{Z} = \mathbb{R}^D$, and consider λ to be the Lebesgue measure over $\mathbb{R}^D$. Notice, as it is well known:

Lemma 14 *For any linear automorphism $A : \mathbb{R}^D \to \mathbb{R}^D$, we have that: $A\#\lambda = |\det(A)|\lambda$.*

In particular, if A is such that $|\det(A)| = 1$ (e.g., if $A \in O(D)$), we have: $A\#\lambda = \lambda$

Proposition 26 *Let G be a compact group with Haar measure λ_G, such that $G \circlearrowright_M \mathcal{Z}$ orthogonally. Let M be non-trivial, and let λ be the Lebesgue measure over $\mathcal{Z} = \mathbb{R}^D$. Let $\mu \in \mathcal{P}(\mathcal{Z})$ be such that $\mu \lll \lambda$; in particular it has a (measurable) density $u : \mathcal{Z} \to \mathbb{R}_+$, (with respect to λ).*

- ***[Case of** $\mathcal{P}(\mathcal{E}^G)$**]** We have:*

  $$\mu^{\mathcal{E}^G} \textit{ has a density wrt } \lambda_{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\lambda \textit{ (restricted to } \mathcal{E}^G)$$

  In particular, as $\mathcal{E}^G$ is a strict subspace of $\mathcal{Z}$, $\mu^{\mathcal{E}^G} \in \mathcal{P}^{\mathcal{E}^G}(\mathcal{Z})$ is degenerate and doesn't have a density with respect to λ.

- ***[Case of*** $\mathcal{P}^G(\mathcal{Z})$*] We have:*

$$\mu^G \in \mathcal{P}^G(\mathcal{Z}) \text{ has density } u^G := \int_G u \circ M_g d\lambda_G(g) \text{ wrt } \lambda$$

As a consequence:

$$\mu \in \mathcal{P}^G(\mathcal{Z}) \iff u \text{ is } G\text{-invariant } (\lambda - a.s.)$$

Proof. See Annex C.11. □

Now, from the theory of invariant measures developped in Kallenberg [45] (which is partially described in section 3.2.1), we can extract even more information about the space $\mathcal{P}^G(\mathcal{Z})$, as we will see in the following section.

4.2.1 Invariant Measures as Orbit Measures

Recalling some of the key *measure-theoretic* results from section 3.2.1; let G be a compact group acting on a (separable) Hilbert Space $\mathcal{Z}$ via a representation M. From the ergodic decomposition theorem for invariant measures (theorem 10) we know that any G-invariant measure on $\mathcal{Z}$ can be seen as a mixture of *orbit measures* (which *fill* each orbit with *uniform* weight). It is thus natural to ask whether any measure in $\mathcal{P}^G(\mathcal{Z})$ might be simply understood as a measure on the *orbit space* $G\backslash\mathcal{Z}$. With the following results we will show that this is indeed the case (under relatively natural assumptions).

As *orbit measures* are *invariant*, we can actually precisely reduce ourselves to the *orbit space* $G\backslash\mathcal{Z}$. For this to make sense in this context, assume the conditions for proposition 12 (point 4.) hold; in particular[12] let $\mathcal{B}_{G\backslash\mathcal{Z}} = \sigma(\mathcal{T}_{G\backslash\mathcal{Z}})$. Let $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ be the *natural projection* and $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ be a (Borel-measurable) *cross-section* for p (i.e. such that $p \circ s = \mathrm{Id}_{G\backslash\mathcal{Z}}$; it exists thanks to proposition 12)

Proposition 27 *As* $z \in \mathcal{Z} \mapsto \varphi_z \in \mathcal{M}(\mathcal{Z})$*, is a* G*-invariant function, there exists a unique* $\overline{\varphi} : G\backslash\mathcal{Z} \to \mathcal{M}(\mathcal{Z})$ *that is measurable and satisfies* $\varphi_x = \overline{\varphi}_{p(x)}$, $\forall x \in \mathcal{Z}$*. In particular,* $\overline{\varphi} : G\backslash\mathcal{Z} \to \mathcal{Z}$ *is a kernel (we call it the factorized orbit measure kernel).*

Proof. See Annex C.11. □

Corollary 5 *Let* G *be a compact group acting on* $\mathcal{Z}$ *and let* $\overline{\varphi} : G\backslash\mathcal{Z} \to \mathcal{Z}$ *be the factorized orbit measure kernel. For every* $\nu \in \mathcal{M}^G(\mathcal{Z})$*, there exists a unique* $\overline{\nu} \in \mathcal{M}(G\backslash\mathcal{Z})$ *such that:*

$$\nu = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(\cdot) d\overline{\nu}(\overline{x})$$

In particular, $\overline{\nu} = p_{\#}\nu$

[12] Recall that $\mathcal{B}_{G\backslash\mathcal{Z}} := \{B \subseteq G\backslash\mathcal{Z} \,:\, p^{-1}(B) \in \mathcal{B}_{\mathcal{Z}}\}$ and $\mathcal{T}_{G\backslash\mathcal{Z}} = \{B \subseteq G\backslash\mathcal{Z} \,:\, p^{-1}(B) \in \mathcal{T}_{\mathcal{Z}}\}$

Proof. See Annex C.11. □

This allows us to put the sets $\mathcal{P}^G(\mathcal{Z})$ and $\mathcal{P}(G\backslash\mathcal{Z})$ in direct correspondance.

Proposition 28 *Let G be a compact group acting on $\mathcal{Z}$ a separable Hilbert Space via a representation M. Let them satisfy the conditions of proposition 12 (point 4.). Consider the map:*

$$\begin{aligned} \Psi : \mathcal{P}^G(\mathcal{Z}) &\to \mathcal{P}(G\backslash\mathcal{Z}) \\ \mu &\mapsto p\#\mu \end{aligned}$$

Then, Ψ is a bijection between $\mathcal{P}^G(\mathcal{Z})$ and $\mathcal{P}(G\backslash\mathcal{Z})$ (whose inverse is defined as $\overline{\mu} \mapsto (s\#\overline{\mu})^G$). Moreover, Ψ is continuous and has measurable inverse. That is, Ψ is actually a bimeasurable bijection.

If we further assumed $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ to be a continuous map[13]*, then Ψ would be a homeomorphism between both spaces.*

Proof. See Annex C.11. □

Also notice, as a particular case, that:

Corollary 6 *Under the same hypothesis of proposition 28, $p|_{\mathcal{E}^G} : \mathcal{E}^G \to G\backslash\mathcal{Z}$ is injective. Let $\overline{\mathcal{E}^G} := p(\mathcal{E}^G)$; we have that: $\Psi|_{\mathcal{P}(\mathcal{E}^G)} : \mathcal{P}(\mathcal{E}^G) \to \mathcal{P}(\overline{\mathcal{E}^G})$ is a bimeasurable bijection (and also, if $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ is continuous, it is a homeomorphism).*

Proof. It follows directly from the fact that, $\forall z \in \mathcal{E}^G,\ G.z = \{z\}$, so that $s(p(z)) = z$; and the use of proposition 28. □

Remark These results will allow us to *reduce* our problem from searching a probability measure over the whole space $\mathcal{Z}$, to simply finding one over $G\backslash\mathcal{Z}$.

This might not seem like a great advantage, but in many cases the space $G\backslash\mathcal{Z}$ has a *smaller dimension* than $\mathcal{Z}$. In the remarkable example given by Mei et al. [57], they use this fact to reduce a problem on $\mathbb{R}^D$ under an $O(D)$-symmetry, into simply a problem over $\mathbb{R}_+$ (which is homeomorphic to $O(D)\backslash\mathbb{R}^D$). After having solved the *unconstrained* problem on $\mathbb{R}_+$ they obtained a measure over the *whole space* by making it *uniform over each orbit.*

In other words, one can *reduce* the problem just to $G\backslash\mathcal{Z}$ (potentially on smaller dimension), solve it over $\mathcal{P}(G\backslash\mathcal{Z})$ (unconstrained) and then *fill the rest of the space uniformly* using the *orbit measures* in order to retrieve a solution to the original problem. i.e. from an interesting measure $\overline{\mu} \in \mathcal{P}(G\backslash\mathcal{Z})$ we apply $\Psi^{-1} : \mathcal{P}(G\backslash\mathcal{Z}) \to \mathcal{P}^G(\mathcal{Z})$, defined as $\Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G$ (which corresponds exactly to this *uniform* way of filling the *rest* of the space) to recover the *natural symmetric extension* of such a measure on the whole space $\mathcal{P}^G(\mathcal{Z})$.

[13] Continuity of the cross-section is, however, a really strong assumption to make

This sort of approach has many potential advantages (as noted by the really interesting work by Chossat [18]), and it shall be explored more deeply in the upcoming sections, as well as in future work.

Remark From corollary 6, it seems like the *dimensionality reduction* discussed in the previous remark won't be effective for $\mathcal{E}^G$ (as it *injects* into its image in $G\backslash\mathcal{Z}$). Points in $\mathcal{E}^G$ are the only ones whose orbits are exactly *singletons*.

Before going on along these lines, we weill consider some key results that will be really useful for the theory that follows.

4.3 Differentials and Integrals of Equivariant Functions

In this work, the notion of *G-equivariant function* will have a crucial role in the understanding the *symmetric properties* of *shallow* NNs. Most remarkably, we have seen that quite *reasonable* variants of σ_* are actually *jointly* G-equivariant (see Proposition 22). It is then crucial to understand how *gradients* and *integrals* of such functions behave, as they will be the explicit drivers of our training dynamic. In some sense, the *gradient/integal* of an equivariant function will *still* be (in some sense) equivariant.

The following lemma characterizes the differential of *jointly equivariant* functions with respect to the action of some group G.

Proposition 29 *Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be (separable) Hilbert Spaces and G be a lcsH group. Let $G \circlearrowright_{\chi} \mathcal{X}$, $G \circlearrowright_{\tilde{\chi}} \mathcal{Z}$, $G \circlearrowright_{\check{\chi}} \mathcal{Y}$ via some representations $\chi, \tilde{\chi}$ and $\check{\chi}$ respectively.*

Let $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be jointly G-equivariant with respect to these actions (i.e. $\forall g \in G,\ \forall x \in \mathcal{X},\ \forall z \in \mathcal{Z},\ \check{\chi}_g.f(x,z) = f(\chi_g.x, \tilde{\chi}_g.z)$) and Fréchet-differentiable on its first argument. Then[14]*:*

$$\forall g \in G,\ \forall x \in \mathcal{X},\ \forall z \in \mathcal{Z},\ \ D_x f(\chi_g.x, \tilde{\chi}_g.z) = \check{\chi}_g.D_x f(x,z)\chi_g^{-1}$$

Proof. See Annex C.12. □

This, in particular, allows us to characterize the differential of G-equivariant functions.

Corollary 7 *If $G \circlearrowright_{\chi} \mathcal{X}$, $G \circlearrowright_{\tilde{\chi}} \mathcal{Y}$, and $f : \mathcal{X} \to \mathcal{Y}$ is a G-equivariant and Fréchet-differentiable function, then:*

$$\forall g \in G,\ \forall x \in \mathcal{X},\ \ D_x f(\chi_g.x) = \tilde{\chi}_g D_x f(x)\chi_g^T$$

Proof. See Annex C.12. □

[14] This condition could also be described as saying that: $D_x f : \mathcal{X} \times \mathcal{Z} \to \mathrm{BL}(\mathcal{X};\mathcal{Y})$; $(x,z) \mapsto D_x f(x,z)$ is a G-equivariant map (with respect to the right group actions).

We can also get some interesting *integral* properties of *jointly* G-equivariant functions.

Proposition 30 *Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be (separable) Hilbert Spaces and G be a lcsH group. Let $G \circlearrowright_{\chi} \mathcal{X}$, $G \circlearrowright_{\tilde{\chi}} \mathcal{Z}$, $G \circlearrowright_{\check{\chi}} \mathcal{Y}$ via some representations $\chi, \tilde{\chi}$ and $\check{\chi}$ respectively.*

Let $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be a jointly G-equivariant function (with respect to these actions). Consider a measure $\mu \in \mathcal{P}(\mathcal{Z})$ and let f be Bochner integrable on its second argument with respect to μ. Then[15]*:*

$$\forall x \in \mathcal{X},\ \forall g \in G,\ \check{\chi}_g \langle f(x; \cdot), \mu \rangle = \langle f(\chi_g x; \cdot), \tilde{\chi}_g \# \mu \rangle$$

Proof. See Annex C.12. □

4.4 Optimizing a Symmetric Functional

Recall that, after *convexifying* our original learning problem, we were faced with the task of optimizing a functional over the space of probability measures, $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$. In particular, with the setting described in section 2.2.4, and all the properties we have established along this chapter, we would expect to be able to extract interesting properties of our problem when we assume R to be *symmetric* in *some* sense. The *natural* way to do so is through the following definition:

Definition 4.4 *Let G be a compact group acting over $\mathcal{Z}$ (a separable Hilbert) via the representation M. We say a functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is G-invariant whenever $\forall g \in G$, $\forall \mu \in \mathcal{P}(\mathcal{Z})$: $R(M_g \# \mu) = R(\mu)$*

As an extension of the previous section, under the lens of G-invariant functionals, we have the following result about linear functional derivatives and intrinsic derivatives:

Proposition 31 *Let $R : \mathcal{P}(\mathcal{Z}) \longrightarrow \mathbb{R}$ be G-invariant and of class $\mathcal{C}^1$. Then: $\forall \theta \in \mathcal{Z}$, $\forall \mu \in \mathcal{P}(\mathcal{Z})$, $\forall g \in G$:*

$$\frac{\partial R}{\partial \mu}(M_g \# \mu, M_g.\theta) = \frac{\partial R}{\partial \mu}(\mu, \theta) \quad \textit{and} \quad D_\mu R(M_g \# \mu, M_g.\theta) = M_g.D_\mu R(\mu, \theta)$$

i.e. $\frac{\partial R}{\partial \mu}$ is jointly G-invariant and $D_\mu R$ jointly G-equivariant.

Proof. See Annex C.13. □

In particular, we would like for G-invariant funtionals to *achieve their minimum* at least within the set $\mathcal{P}^G(\mathcal{Z})$. For this purpose, the following version of *Jensen's Inequality* will be key:

[15] This could also be understood as saying that: the map $(x, \mu) \in \mathcal{X} \times \mathcal{P}(\mathcal{Z}) \mapsto \langle f(x; \cdot), \mu \rangle \in \mathcal{Y}$ is G-equivariant (with respect to the *good* G-actions).

Proposition 32 (**Jensen's Inequality**) *Let $R : \mathcal{P}(\mathcal{Z}) \longrightarrow \mathbb{R}$ be such that Lemma 4 holds (e.g. it is enough to take R convex and of class $\mathcal{C}^1$). Let $\lambda \in \mathcal{P}(S)$ and $\{\mu_s\}_{s\in S} \subseteq \mathcal{P}(\mathcal{Z})$; and define $\tilde{\mu} \in \mathcal{P}(\mathcal{Z})$ as the intensity measure*[16]*: $\tilde{\mu}(\cdot) = \int_S \mu_s(\cdot)d\lambda(s) \in \mathcal{P}(\mathcal{Z})$. Then, a Jensen-like inequality is satisfied:*

$$R(\tilde{\mu}) \leqslant \int_S R(\mu_s)d\lambda(s)$$

Proof. See Annex C.13. □

Remark To our knowledge, such a result is novel in this context (of functionals over the space of probability measures); however, the proof is somewhat *standard* and it wouldn't be surprising to find a previous statement of the same result deep in the literature. The question about *what happens when there's equality in the above expression* is open (to our knowledge) and left for future work[17].

In any case, thanks to this *Jensen inequality*, we get the following *general* result:

Corollary 8 *If $R : \mathcal{P}(\mathcal{Z}) \longrightarrow \mathbb{R}$ is convex, $\mathcal{C}^1$-regular and G-invariant in the sense of definition 4.4, then $\forall \mu \in \mathcal{P}(\mathcal{Z})$:*

$$R(\mu^G) \leqslant R(\mu)$$

Proof. Direct from the definition of $(\cdot)^G$. □

This, in turn, leads to the following general result for *convex* G-invariant functionals:

Proposition 33 (Optimality of Invariant Measures) *Whenever R is convex, $\mathcal{C}^1$ and $G-$ invariant, it holds that:*

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$$

In particular, if $\mu_* \in \arg\min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$, then $\mu_*^G \in \arg\min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$ as well.

Proof. See Annex C.13. □

Remark Notice that if there was a *unique minimizer* μ_* of R over $\mathcal{P}(\mathcal{Z})$, then it would have to satisfy: $\mu_* = \mu_*^G \in \mathcal{P}^G(\mathcal{Z})$, thus forcing the unique solution to be G-invariant. Also, understanding what happens when there's *equality* in proposition 32 might give us insight into the *structure* that μ^* would have in such a case.

In the following section we will see that, under the right conditions, the *population risk* on a learning problem, $R(\mu) = \mathbb{E}_\pi[\ell(\langle\sigma_*(x,\cdot),\mu\rangle), Y)]$, will be G-invariant (in the sense of definition 4.4.

[16] If $X \sim \lambda \in \mathcal{P}(S)$, it is the intensity measure of the random measure μ_X

[17] Possibly even the $\mathcal{C}^1$ restriction might be lifted by employing properties of the subdifferential of R instead of its linear functional derivative.

Example Some other remarkable examples of G-invariant functionals over $\mathcal{P}(\mathcal{Z})$ include:

1. Let $V(\mu) = \int_{\mathcal{Z}} r(\theta)d\mu(\theta)$ with $r : \mathcal{Z} \to \mathbb{R}$ a G-invariant regularization functional (e.g $r(\theta) = \|\theta\|^2$ for *orthogonal representations*). Indeed, notice that for $g \in G$:

$$V(M_g\#\mu) = \int_{\mathcal{Z}} r(\theta)d(M_g\#\mu)(\theta) = \int_{\mathcal{Z}} r(M_g\theta)d\mu(\theta) = \int r(\theta)d\mu(\theta) = V(\mu)$$

   thanks to the G-invariance of r. i.e. V is G-invariant.

2. Consider the *relative entropy* $H_\nu(\mu) = \int_{\mathcal{Z}} \log(\frac{d\mu}{d\nu})d\mu$ with $\mu \lll \nu$ and $\nu \in \mathcal{M}^G(\mathcal{Z})$ (e.g. $\nu = \lambda$ for *orthogonal representations*). Fix some $g \in G$; it is a known fact[18] that, as ν is G-invariant, $\frac{d(M_g\#\mu)}{d\nu}(x) = \frac{d\mu}{d\nu}(M_g^{-1}x)$. Therefore:

$$\begin{aligned} H_\nu(M_g\#\mu) &= \int \log\left(\frac{d(M_g\#\mu)}{d\nu}(\theta)\right) d(M_g\#\mu)(\theta) = \int \log\left(\frac{d\mu}{d\nu}(M_g^{-1}\theta)\right) d(M_g\#\mu)(\theta) \\ &= \int \log\left(\frac{d\mu}{d\nu}(M_g^{-1}M_g\theta)\right) d\mu(\theta) = H_\nu(\mu) \end{aligned}$$

   Which proves that H_ν is $G-$invariant (whenever the reference measure ν is as well).

3. In particular, recalling the definition of our *regularized risk* $R^{\tau,\beta} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, which is $\forall \mu \in \mathcal{P}(\mathcal{Z})$:

$$R^{\tau,\beta}(\mu) = R(\mu) + \tau \int r d\mu + \beta H_\nu(\mu)$$

   Whenever $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, $r : \mathcal{Z} \to \mathbb{R}$ and $\nu \in \mathcal{M}(\mathcal{Z})$ are G-invariant (in their respective sense); then $R^{\tau,\beta}$ is G-invariant.

This example shows that, when the *good conditions* are satisfied, a minimum for the *regularized problem* (or *unregularized* if we take $\tau, \beta = 0$) can be found within $\mathcal{P}^G(\mathcal{Z})$.

From the correspondence established in proposition 28, one might even *reduce the problem further:* as we know that $\mathcal{P}^G(\mathcal{Z})$ and $\mathcal{P}(G\backslash\mathcal{Z})$ are in *correspondence* (through the bijection $\Psi : \mathcal{P}^G(\mathcal{Z}) \to \mathcal{P}(G\backslash\mathcal{Z})$), we expect to be able to minimize by considering *only measures over the orbit space* $G\backslash\mathcal{Z}$. This actually holds, as shown by the following proposition:

Proposition 34 *Let G be a compact group acting on $\mathcal{Z}$ a separable Hilbert Space via a representation M. Let them satisfy the conditions of proposition 12 (point 4.). Then:*

$$\inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\overline{\mu} \in \mathcal{P}(G\backslash\mathcal{Z})} R((s\#\overline{\mu})^G)$$

Proof. See Annex C.13. □

[18] Indeed, for positive and measurable $f : \mathcal{Z} \to \mathbb{R}$, denoting $u = \frac{d(M_g\#\mu)}{d\nu}$, we have that (thanks to the G-invariance of ν): $\langle f, M_g\#\mu\rangle = \int f(M_g z)u(z)d\nu(z) = \int f(z)u(M_g^{-1}z)d\nu(z)$

This might not seem as much of an improvement, but at least *conceptually*, when R is convex, $\mathcal{C}^1$ and G-invariant, this allows us to solve a *smaller* problem, over $\mathcal{P}(G\backslash\mathcal{Z})$ rather than over the entire space $\mathcal{P}(\mathcal{Z})$. In particular, we only need to look at: $\overline{R} : \mathcal{P}(G\backslash\mathcal{Z}) \to \mathbb{R}$ defined by $\overline{R}(\overline{\mu}) = R((s\#\overline{\mu})^G)$ and solve:

$$\inf_{\overline{\mu}\in\mathcal{P}(G\backslash\mathcal{Z})} \overline{R}(\overline{\mu})$$

Remark A common example, that appears in [57], is one with $G \circlearrowright_{O(D)} \mathbb{R}^D$, for which this *reduction* amounts to *reducing the problem* onto the *ray* $\mathbb{R}_+ \cong G\backslash\mathcal{Z}$. In particular, this *reduction to the orbit space* might also allow for a *dimensionality reduction* that could be potentially helpful (e.g. the involved ODEs of the MF training dynamic might be easier to solve *numerically* in a smaller dimension. See Chossat [18] for other reference ideas about this approach).

Example Notice that, for the relevant elements that appear in the theory of WGF; as well as for the learning problem, we get the *following results*:

1. Consider $V(\mu) = \int r(\theta)d\mu$ with $r : \mathcal{Z} \to \mathbb{R}$ G-invariant. Then:

$$V((s\#\overline{\mu})^G) = \int r(\theta)d(s\#\overline{\mu})^G(\theta) = \int_G\int_{\mathcal{Z}} r(M_g\theta)d(s\#\overline{\mu})(\theta)d\lambda_G(\mu)$$

   As r is $G-$invariant, we have:

$$V((s\#\overline{\mu})^G) = \int_G\int_{\mathcal{Z}} r(M_g\theta)d(s\#\overline{\mu})(\theta)d\lambda_G(\mu) = \int_{\mathcal{Z}} r(\theta)d(s\#\overline{\mu})(\theta) = \int_{G\backslash\mathcal{Z}} r(s(\overline{\theta}))d\overline{\mu}(\overline{\theta})$$

   So, if we define $\overline{r} : G\backslash\mathcal{Z} \to \mathbb{R}$ to be: $\overline{\theta} \mapsto r(s(\overline{\theta}))$ we can further define:

$$\overline{V}(\overline{\mu}) := V((s\#\overline{\mu})^G) = \int_{G\backslash\mathcal{Z}} r(s(\overline{\theta}))d\overline{\mu}(\overline{\theta}) = \int_{G\backslash\mathcal{Z}} \overline{r}(\overline{\theta})d\overline{\mu}(\overline{\theta})$$

   In particular the *regularization term*, restricted to *extensions* of measures in $G\backslash\mathcal{Z}$, can be seen directly as an *analogous regularization term* defined directly for measures over $G\backslash\mathcal{Z}$[19].

2. Consider $H_\nu(\mu) = \int_{\mathcal{Z}} \log(\frac{d\mu}{d\nu})d\mu$ with $\mu \lll \nu$ and $\nu \in \mathcal{M}^G(\mathcal{Z})$. We have that:

$$H_\nu((s\#\overline{\mu})^G) = \int \log\left(\frac{d(s\#\overline{\mu})^G}{d\nu}\right) d(s\#\overline{\mu})^G$$

   As $(s\#\overline{\mu})^G$ and ν are G-invariant, we get from theorem 12 that $f(\theta) := \frac{d(s\#\overline{\mu})^G}{d\nu}(\theta)$ is G-invariant (it can be chosen that way WLOG a.s. uniquely). This implies (through the classic *factorization* theorem) that $\exists!\overline{f} : G\backslash\mathcal{Z} \longrightarrow \mathbb{R}_+$ such that

$$\forall z \in \mathcal{Z},\ f(z) = \overline{f}(p(z)) \quad \text{or, equivalently,} \quad \forall \overline{z} \in G\backslash\mathcal{Z},\ f(s(\overline{z})) = \overline{f}(\overline{z})$$

[19] Notice that this *new* regularization term $\overline{r}$ will be, in general, *less regular* than the original r; unless $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ is sufficiently regular

As intuition tells, we get that: $\overline{f} = \frac{d\overline{\mu}}{d\overline{\nu}}$, with $\overline{\nu} = p\#\nu$. Indeed, let $A \in \mathcal{B}_{G\backslash\mathcal{Z}}$ and recall that, for $\mu := (s\#\overline{\mu})^G$, we have $\psi(\mu) = p\#\mu = \overline{\mu}$. So, in particular (using the fact that f is the density of μ and the definition of $\overline{f}$), we get:

$$\overline{\mu}(A) = \mu(p^{-1}(A)) = \int_{p^{-1}(A)} f(z)d\nu(z) = \int_{p^{-1}(A)} \overline{f}(p(z))d\nu(z)$$

By the definition of the pushforward measure[20] we finally recover

$$\overline{\mu}(A) = \int_{p^{-1}(A)} \overline{f}(p(z))d\nu(z) = \int_A \overline{f}(\overline{z})d(p\#\nu)(\overline{z}) = \int_A \overline{f}d\overline{\nu}$$

which allows us to conclude that $\overline{\mu} \lll \overline{\nu}$ and $\frac{d\overline{\mu}}{d\overline{\nu}} = \overline{f}$.

Now, consider (for $f = \frac{d(s\#\overline{\mu})^G}{d\nu}$):

$$H_\nu((s\#\overline{\mu})^G) = \int_z \log(f(z))d(s\#\overline{\mu})^G(z) = \int_z \int_G \log(f(M_g z))d\lambda_G(g)d(s\#\overline{\mu})(z)$$

As f is G-invariant, we get:

$$H_\nu((s\#\overline{\mu})^G) = \int_{\mathcal{Z}} \log(f(z))d(s\#\overline{\mu})(z) = \int_{s(G\backslash\mathcal{Z})} \log(f(z))d(s\#\overline{\mu})(z)$$

Again, by definition of pushforward measures (and using the injectivity of s):

$$H_\nu((s\#\overline{\mu})^G) = \int_{s^{-1}(s(G\backslash\mathcal{Z}))} \log(f(s(\overline{z})))d\overline{\mu}(\overline{z}) = \int_{G\backslash\mathcal{Z}} \log(f(s(\overline{z})))d\overline{\mu}(\overline{z})$$

Finally, by the property $\overline{f}(\overline{z}) = f(s(\overline{z}))$:

$$H_\nu((s\#\overline{\mu})^G) = \int_{G\backslash\mathcal{Z}} \log(\overline{f}(\overline{z})d\overline{\mu}(\overline{z}) =: \overline{H}_{\overline{\nu}}(\overline{\mu})$$

So again, the *entropy regularization term* (with respect to $\nu \in \mathcal{M}^G(\mathcal{Z})$), restricted to *extensions* of measures in $G\backslash\mathcal{Z}$, can be seen directly as an *entropy term* defined directly for measures in $G\backslash\mathcal{Z}$ (with respect to the *projection* $p\#\nu$).

3. In learning setting, consider: $R(\mu) := \mathbb{E}_\pi[\ell(\langle\sigma_*(X,\cdot),\mu\rangle),Y)]$ with $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ a *jointly* G-equivariant function (with respect to $G \circlearrowleft_\rho \mathcal{X}$, $G \circlearrowleft_M \mathcal{Z}$, $G \circlearrowleft_{\hat{\rho}} \mathcal{Y}$); then, $\forall x \in \mathcal{X}$, $\overline{\mu} \in \mathcal{P}(G\backslash\mathcal{Z})$:

$$\begin{aligned}
\langle\sigma_*(x,\cdot),(s\#\overline{\mu})^G\rangle &= \int_G \int_{G\backslash\mathcal{Z}} \sigma_*(x, M_g s(\overline{z}))d\overline{\mu}(\overline{z})d\lambda_G(g) \\
\text{By proposition 30 } &= \int_G \int_{G\backslash\mathcal{Z}} \hat{\rho}_g \sigma_*(\rho_g^{-1}x, s(\overline{z}))d\overline{\mu}(z)d\lambda_G(g) \\
\text{By Fubini } &= \int_{G\backslash\mathcal{Z}} \int_G \hat{\rho}_g \sigma_*(\rho_g^{-1}x, s(\overline{z}))d\lambda_G(g)d\overline{\mu}(z) \\
&= \int_{G\backslash\mathcal{Z}} \overline{\sigma_*}(x,\overline{z})d\overline{\mu}(\overline{z})
\end{aligned}$$

[20] Recall that for $f : \mathcal{X}_1 \to \mathcal{X}_2$: $\int_{\mathcal{X}_2} gd(f\#\mu) = \int_{f^{-1}(\mathcal{X}_2)}(gof)d\mu$

Where we've defined $\overline{\sigma_*} : \mathcal{X} \times G\backslash\mathcal{Z} \to \mathcal{Y}$ by

$$\overline{\sigma_*}(x, \overline{z}) := \int_G \hat{\rho}_g \sigma_*(\rho_g^{-1} x, s(\overline{z})) d\lambda_G(g)$$

Notice the striking ressemblance between $\overline{\sigma_*}$ and the *symmetrization operation* typical to **FA** (we'll comment further about this connection in upcoming sections). We conclude that:

$$R((s\#\overline{\mu})^G) = \mathbb{E}_\pi[\ell(\langle\sigma_*(X, \cdot), (s\#\overline{\mu})^G\rangle), Y)] = \mathbb{E}_\pi[\ell(\langle\overline{\sigma_*}(x, \cdot), \overline{\mu}\rangle), Y] =: \overline{R}(\overline{\mu})$$

So, similar as before, the *risk* we will *effectively* be minimizing (over $\mathcal{P}(G\backslash\mathcal{Z})$) corresponds to *the same form*, only with *slight modifications*.

4. Thus, in general, as we've proven (under the right assumptions of G-invariance of r, ν and R) $\min_{\mu\in\mathcal{P}(\mathcal{Z})} R_\nu^{\tau,\beta}(\mu) = \min_{\overline{\mu}\in\mathcal{P}(G\backslash\mathcal{Z})} \overline{R_\nu^{\tau,\beta}}(\overline{\mu})$, where $\overline{R_\nu^{\tau,\beta}}$ can be seen *directly* as the *functional to minimize*, as it has (essentially) the same structure as an unrestricted problem over $\mathcal{P}(G\backslash\mathcal{Z})$. In particular, based on all previous examples, we can *understand* this problem *directly* over $G\backslash\mathcal{Z}$ (with the corresponding modified functionals $\overline{R}, \overline{V}$ and $\overline{H}_{\overline{\nu}}$) in a rather straightforward manner.

From proposition 33 and proposition 34, one might expect that a similar result shall hold for the space $\mathcal{P}(\mathcal{E}^G)$. Unfortunately, as the following counterexample shows, in general

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) < \inf_{\nu\in\mathcal{P}(\mathcal{E}^G)} R(\nu)$$

Proposition 35 *There are instances of the learning problem (as introduced in section 2.2.3; with $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle\sigma_*(X;\cdot), \mu\rangle, Y)\right]$), considering a finite group G acting on $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and $\mathcal{Z} = \mathbb{R}^{(d+1)}$ via orthogonal representations, having the data distribution π being G-invariant (and compactly supported), the activation function σ_* being jointly G-equivariant, $\mathcal{C}^\infty$ and bounded; and using the quadratic loss ℓ, such that:*

$$\min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) < \min_{\nu\in\mathcal{P}^{\mathcal{E}^G}(\mathcal{Z})} R(\nu)$$

Proof. See Annex C.13. □

A similar question was asked at the end of Elesedy and Zaidi [28] (related to their result on *shallow* NNs that we referenced in section 3.4.3), suggesting that it was an open problem. This counterexample suggests that, if *equivariant architectures* are chosen to be *too restrictive*, a potentially important cut in *generalization power* might be undergone: in general, we *won't attain an optimum of* R within the space $\mathcal{P}(\mathcal{E}^G)$.

A rather *simple* way of solving this issue is assuming that the class $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{E}^G))$ is *universal*, in which case the best possible value of the learning problem will always be optimal:

Proposition 36 *Consider the learning problem ($R(\mu) = \mathbb{E}_\pi\left[\ell(\langle\sigma_*(X;\cdot), \mu\rangle, Y)\right]$) under a quadratic loss, for a data distribution $\pi \in \mathcal{P}_2^G(\mathcal{X} \times \mathcal{Y})$. Assume that $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{E}^G))$ is universal* *on $L^2_G(\mathcal{X},\mathcal{Y};\pi|_{\mathcal{X}})$ (this in particular implies that $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$ is universal as well on that space). Then:*

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{\nu\in\mathcal{P}^{\mathcal{E}^G}(\mathcal{Z})} R(\nu) = R_*$$

Proof of Proposition 36. As done in Lemma 2, using Lemma 1 we can extract that

$$R(\mu) = \mathbb{E}_\pi[\|Y - \langle\sigma_*(X;\cdot),\mu\rangle\|^2_{\mathcal{Y}}] = R_* + \mathbb{E}_\pi[\|f^*(X) - \langle\sigma_*(X;\cdot),\mu\rangle\|^2_{\mathcal{Y}}]$$

where $f_*(x) := \mathbb{E}_\pi[Y|X=x]$ and R_* is the *Bayes risk* of the problem. From Proposition 13, we know that $f^* \in L^2_G(\mathcal{X},\mathcal{Y};\pi|_{\mathcal{X}})$, and so by universality of $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{E}^G))$ onto that space (as well as that of $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$), we conclude directly as in Lemma 2. □

Remark As stated in theorem 1, there are conditions on $\mathcal{E}^G$ and σ_* that would eventually allow us to satisfy assumption 1 and have the *desired* universality result (on $L^2_G(\mathcal{X},\mathcal{Y};\pi|_{\mathcal{X}})$). On the other hand, Maron et al. [55], Yarotsky [92], Zaheer et al. [93] among others have been able to provide conditions for *equivariant NNs* (of different kind) to be *universal* on the set of G-equivariant functions from $\mathcal{X}$ to $\mathcal{Y}$. Particularly, as noted in Maron et al. [55], *equivariant NNs* of *tensor order* 1 (as are the ones presented in our setting of section 4.1.2) are unable to achieve universality for certain types of group actions (see Theorem 2 from [55]). Bridging this gap (eventually allowing for *arbitrary order tensors* in our MF formulation) is part of our future challenges to make our approach more broadly applicable. In any case, *first order universality* (which is what we hope to have in our setting) has been established for relevant examples such as *Deep Sets* ([93, 92]) and *CNNs* ([92]) (for illustrative purposes, an example of a DeepSet architecture is presented in section B.1).

Remark A really interesting question at this point is whether under this setting (where *universality* holds) any sort of *explicit relation* can be established between the minimizers of both problems (if they explicitly exist).

For instance, one might wish that the optimum $\mu^* \in \arg\min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$ and $\nu^* \in \arg\min_{\nu\in\mathcal{P}^{\mathcal{E}^G}} R(\nu)$ could satisfy $(\mu^*)^{\mathcal{E}^G} = \nu^*$. With this in place, there could be an explicit way of finding *equivariant* optima (and with *uniqueness* of the minimizer, this would force the *global* optimum to be concentrated on $\mathcal{E}^G$). However, for such a relationship to hold, really constraining conditions must be assumed, and it's not clear under which context it would even be reasonable. Conversely, one could ask whether there is a *canonical way* of *extending* ν^* to make it *non-degenerate* on $\mathcal{Z}$ and also optimal for the *global* problem (as μ^*). Is there some measure $\tilde{\nu}$ over $(\mathcal{E}^G)^\perp$ such that $\nu\otimes\tilde{\nu} = \mu^*$? A deeper exploration of this topic could shed light onto whenever the training of explicitly equivariant NN models could have explicit advantages (and no *loss of information*) over *fully-connected* NN models. This is definitely something to tackle as part of our future work.

The following chapter will bring together multiple of the theoretical elements presented until this point, in order to explore the different forms of *exploiting symmetries* in the setting of NNs under our new lens. In particular, we will use them to study how the MF dynamics relates with the potential G-invariance of our functionals of interest.

Chapter 5

Symmetries of the Mean Field Dynamic

The main objective of this work is to try and leverage the properties described on the previous chapters, in order to explore the *symmetries* that the MF Limit (on a learning problem) may have when the input data is also *symmetric*.

Most notably, as mentioned in 2.2.7, Mei et al. [57] suggest that training with *G-invariant* data should lead to a mean field limit that is also *G-invariant* in some sense. In what follows, we will precisely detail what this statement refers to, and shall demonstrate its truth in an even stronger sense that what was initially proposed by Mei et al. [57]. Moreover, we will study the ramifications of these ideas to the study of **DA, FA** and **EA** of Neural Networks.

Recall the setting of the MF Limit of *shallow* NNs as described in Chapter 2. We will let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be separable Hilbert spaces and G a compact group that acts on all three spaces via *orthogonal representations* $G \circlearrowright_{\rho} \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$ and $G \circlearrowright_{M} \mathcal{Z}$. We'll often just consider $\mathcal{Z} = \mathbb{R}^D$. We will consider, as most of the literature, *shallow NNs* that *depend solely* on an *activation function* $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$; i.e.: for $\theta = (\theta_i)_{i=1}^N \in \mathcal{Z}^N$ the function $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$ given by:

$$\Phi_\theta^N(x) = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x; \theta_i), \quad \forall x \in \mathcal{X}$$

Recall that, if we define the empirical measure of θ by $\nu_\theta^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$, we have:

$$\Phi_\theta^N(x) = \langle \sigma_*(x; \cdot), \nu_\theta^N \rangle$$

Also recall that we defined (definition 4.2) a *shallow NN* to be *equivariant* if $\forall i \in \{1, \ldots, N\}$, $\theta_i \in \mathcal{E}^G$; or, equivalently, iff $\nu_\theta^N \in \mathcal{P}^{\mathcal{E}^G}(\mathcal{Z})$.

5.1 Symmetrizing NN Models

As we saw in proposition 21, *shallow equivariant NNs* yield G-equivariant models whenever $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is jointly G-invariant (and actually, this is true for all models of the form

$f = \int_{\mathcal{Z}} \sigma_*(\cdot;\theta)d\mu(\theta)$ with $\mu \in \mathcal{P}(\mathcal{E}^G)$). In reality, this isn't a property exclusive to measures in $\mathcal{P}(\mathcal{E}^G)$, as the following holds true (and its proof is exactly the same as for proposition 21):

Proposition 37 *Let $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be jointly G-invariant. Then $\forall \mu \in \mathcal{P}^G(\mathcal{Z})$ $f = \int_{\mathcal{Z}} \sigma_*(\cdot;\theta)d\mu(\theta)$ is a G-equivariant function.*

With such a result under our radar, we might ask ourselves whether the space $\mathcal{P}(\mathcal{E}^G)$ is *special* at all in the setting of *shallow NNs*. For instance, one may ask whether a network obtained from *proyecting* a given parameter onto $\mathcal{E}^G$ (i.e. for any $\theta \in \mathcal{Z}$, the NN given by[1] $\Phi^N_{P_{\mathcal{E}^G}\theta}$) has anything to do with the *symmetrization* of the original NN model, as defined in definition 3.8 (i.e. $\mathcal{Q}\Phi^N_\theta$). That is to say: if we consider the version of a model *that uses the closest possible equivariant parameter* and compare it to the *symmetrized version of the original model* (as from definition 3.8); are these two objects related in any way? One can actually check the following:

Proposition 38 *Let G be a compact group such that: $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_M \mathcal{Z}$ and $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$. Let $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and, for $\mu \in \mathcal{P}(\mathcal{Z})$ define $f_\mu : \mathcal{X} \longrightarrow \mathcal{Y}$ by $f_\mu(x) = \langle \sigma_*(\cdot, x), \mu\rangle$. Then, the symmetrization of f_μ (as in definition 3.8) is given by, $\forall x \in \mathcal{X}$:*

$$(\mathcal{Q}f_\mu)(x) = \langle \sigma_*^G(x,\cdot), \mu\rangle$$

where $\sigma_*^G : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is defined as: $\sigma_*^G(x,\theta) := \int_G \hat{\rho}_g.\sigma_*(\rho_g.x,\theta)d\lambda_G(g)$.

Furthermore, if we assume σ_* to be jointly G-equivariant, we get:

$$(\mathcal{Q}f_\mu)(x) = f_{\mu^G}(x) = \langle \sigma_*(x,\cdot), \mu^G\rangle$$

Where μ^G represents the symmetrization of measure μ (as in definition 4.3). In particular, simetrizing a model from $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$ in the usual sense corresponds to considering the model given by the symmetrized version of the measure (which lives in $\mathcal{F}_{\sigma_*}(\mathcal{P}^G(\mathcal{Z}))$).

Proof. See Annex C.14. □

Example (Symmetrizing a NN) Consider a NN model: Φ^N_θ with $\theta \in \mathcal{Z}^N$; which is $\forall x \in \mathcal{X}$ given by $\Phi^N_\theta(x) = \langle \sigma_*(x,\cdot), \nu^N_\theta\rangle$.

Notice that $\forall g \in G$ and measurable $f : \mathcal{Z} \to \mathbb{R}_+$:

$$\langle f, M_g\#\nu^N_\theta\rangle = \int f(M_g\theta)d\nu^N_\theta(\theta) = \frac{1}{N}\sum_{i=1}^N f(M_g\theta_i)$$

Thus,

$$\langle f, (\nu^N_\theta)^G\rangle = \int_G \frac{1}{N}\sum_{i=1}^N f(M_g\theta_i)d\lambda_G(g) = \frac{1}{N}\sum_{i=1}^N \int_G f(M_g\theta_i)d\lambda_G(g)$$

Recall that for $\theta \in \mathcal{Z}$, $T_\theta : G \to \mathcal{Z}$ given by $T_\theta(g) = M_g.\theta$ allows us to define the *orbit*

[1] Where we understand, for $\theta \in \mathcal{Z}^N$ $\mathcal{P}_{\mathcal{E}^G}\theta := (\mathcal{P}_{\mathcal{E}^G}\theta_i)_{i=1}^N$

measure as $\varphi_\theta = T_\theta \# \lambda_G$, so we have:

$$\begin{aligned}\langle f, (\nu_\theta^N)^G \rangle &= \frac{1}{N}\sum_{i=1}^N \int_G f(T_{\theta_i}(g)) d\lambda_G(g) = \frac{1}{N}\sum_{i=1}^N \int_{G\theta_i} f(z) d(T_{\theta_i}\#\lambda_G)(z) \\ &= \frac{1}{N}\sum_{i=1}^N \langle f, \varphi_{\theta_i}\rangle = \langle f, \frac{1}{N}\sum_{i=1}^N \varphi_{\theta_i}\rangle\end{aligned}$$

That is, $(\nu_\theta^N)^G = \frac{1}{N}\sum_{i=1}^N \varphi_{\theta_i}$. So, essentially, a *symmetrized NN* corresponds to integrating the activation's value over each individual parameter's orbit. In particular, for a finite group G, this is equivalent to a network with (at most) $N \cdot |G|$ parameters. In other words, the network *has one parameter for each orbit element* (in some sense, the network *memorizes* all data orientations). This is in sharp contrast with $(\nu_\theta^N)^{\mathcal{E}^G} = \nu^N_{P_{\mathcal{E}^G}.\theta} = \frac{1}{N}\sum_{i=1}^N \delta_{P_{\mathcal{E}^G}.\theta_i}$, which only admits *explicitly* G-invariant parameters (with $\leqslant N$ distinct parameters).

To answer our previous question, the object that *naturally* appears as the *symmetrized* model for a NN has (in principle) nothing to do with the space of equivariant parameters $\mathcal{E}^G$. Moreover, in general there will be values of $\theta \in \mathcal{Z}$ such that $\Phi^N_{P_{\mathcal{E}^G}\theta} \neq \mathcal{Q}\Phi^N_\theta$. This leads (to some extent) to the problem that inspires the counterexample of proposition 35: when $\mathcal{E}^G$ is *too restricted*, it will turn the corresponding NN model actually *away* from it's *optimal* version[2].

5.2 Networks that exploit Symmetry

Consider a probability distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (for this part, *not necessarily G-invariant*) from which we'll draw i.i.d. samples of data of the form $(X, Y) \in \mathcal{X} \times \mathcal{Y}$; and a *smooth* loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, *convex* on its first argument, that we'll use to evaluate model predictions. As seen in Chapter 2, we want to *find an optimal set of parameters*[3] $\theta \in \mathcal{Z}^N$ *that minimizes the population risk* (i.e. generalizes well) $R(\theta) = \mathbb{E}_\pi\left[\ell(\Phi_\theta^N(X), Y)\right] = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \nu_\theta^N\rangle, Y)\right]$. As seen in Chapter 2, such a function is highly non-convex, so we recur to the *convexification* of the problem in which we seek to solve:

$$\min_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) := \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right] \tag{5.1}$$

Also, recall the notation introduced in section 2.2.6, where we defined, for $x \in \mathcal{X},\ y \in \mathcal{Y}$, $L_{x,y} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ given by: $L_{x,y}(\mu) := \ell\left(\langle \sigma_*(x,\cdot), \mu\rangle, y\right)$. So that we can rewrite our functional as: $R(\mu) := \mathbb{E}_\pi\left[L_{X,Y}(\mu)\right]$, and also[4]: $D_\mu R(\theta, \mu) = \mathbb{E}_\pi\left[D_\mu L_{X,Y}(\theta,\mu)\right]$.

The study of the MF Limit ultimately tries to justify that a solution to this problem may be found through an SGD training dynamic. In our setting, we will further try to study how *symmetries* intervene in such MF limit.

[2] Recall that, in our counterexample of proposition 35 $\Phi^N_{P_{\mathcal{E}^G}\theta} \equiv 0$, in sharp contrast with the *symmetrized* version of such model

[3] Having fixed the network *architecture* with N and σ

[4] Since we have $\forall x \in \mathcal{X},\ y \in \mathcal{Y}$, that $\frac{\partial L_{x,y}}{\partial \mu}(\theta, \mu) = \langle \nabla_1 \ell(\langle \sigma_*(x,\cdot), \mu\rangle, y), \sigma_*(x,\theta)\rangle_{\mathcal{Y}}$ and $D_\mu L_{x,y}(\theta,\mu) = \nabla_\theta \sigma_*(x,\theta) \cdot \nabla_1 \ell(\langle \sigma_*(x,\cdot), \mu\rangle, y)$.

For this purpose, the following proposition tells us that, under the right conditions, we might place our problem under the lens of the *G-invariant functionals over $\mathcal{P}(\mathcal{Z})$* as described in section 4.4.

Proposition 39 *Let the data distribution π be jointly G-invariant, ℓ be jointly G-invariant; and σ_* be jointly G-equivariant. Then, the functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ defined as $R(\mu) = \mathbb{E}_\pi \left[\ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y)\right]$ is G-invariant (in the sense of definition 4.4).*

If we further let ℓ be convex and smooth; and σ_* be smooth and bounded, then R is also convex and $\mathcal{C}^1$. In particular, for any $\mu \in \mathcal{P}(\mathcal{Z})$ we have $R(\mu^G) \leqslant R(\mu)$. Therefore, $\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu)$ and, whenever $\mu_* \in \mathcal{P}(\mathcal{Z})$ achieves this minimum, then $\mu_*^G \in \mathcal{P}^G(\mathcal{Z})$ achieves it as well.

Proof. See Annex C.15. □

Remark As previously noted, the conditions for proposition 39 to hold aren't overly restrictive in the setting of NN models, since:

- Assuming $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$ is standard in the literature concerning *symmetries of NNs.*
- A loss function such as the *quadratic loss* ($\ell(y, \hat{y}) = \frac{1}{2}||y - \hat{y}||^2$; or any other L^p norm for that matter) is known to be G-invariant under orthogonal representations.
- From proposition 22 we know that under a *standard* setting of *shallow NNs*, $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is *jointly G-equivariant* (with respect to the actions $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$, $G \circlearrowright_M \mathcal{Z}$).

Remark Notice, once again, that if there was a *unique minimizer* μ_* for problem (5.1), then it would have to satisfy: $\mu_* = \mu_*^G \in \mathcal{P}^G(\mathcal{Z})$, forcing the unique solution to be G-invariant.

The question of *what happens* whenever there's the equality $R(\mu^G) = R(\mu)$ is unfortunately still open. We know that this happens when $\pi_{\mathcal{X}}$-a.s. for all $x \in \mathcal{X}$ we have $\forall g, h \in G : \langle \sigma_*(x; \cdot), M_g \# \mu_* \rangle = \langle \sigma_*(x; \cdot), M_h \# \mu_* \rangle$. This is, however, not enough to ensure that $\mu_* \in \mathcal{P}^G(\mathcal{Z})$ (which is what we would ultimately want to have).

Remark As mentioned previously, it is interesting to ask whether a global solution to problem (5.1) can be found within the space $\mathcal{P}(\mathcal{E}^G)$; meaning that it could potentially be approximated only by using *equivariant shallow NNs* (by a procedure as in proposition 3). Unfortunately, as shown in proposition 35 this is (in general) *not* the case: the problem that naturally appears, even in *simple* settings, is that when *too strong* equivariance is imposed on the network's architecture, a lot of the network's *expressive power* is lost (e.g. in the counterexample of Proposition 35, we have $\forall \theta \in (\mathcal{E}^G)^N$, $\Phi_\theta^N \equiv 0$). In particular, the problem may rely on the fact that in such context the class $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{E}^G))$ (of *equivariant NNs*) is decidedly *not* universal (in contrast to what might happen with *free* shallow NNs).

When universality is assumed, from proposition 36, one would expect to *have a G-equivariant NN* as a solution to our learning problem (or, at the least, arbitrarily close). If we assumed that some $\nu^* \in \mathcal{P}(\mathcal{E}^G)$ achieved the *minimum* of problem (5.1), the question would now be whether such an optimal measure will be achieved (or not) through the optimization dynamics. These interesting questions are partially tackled in the following sections; however, some still remain open and left for future work.

Before directly tackling the properties of the *mean field dynamic* of NN training, we make a digression to introduce some elements from the three main settings introduced in section 3.4 for *leveraging* symmetries during NN training (**DA**, **FA** and **EA**); now observed under the lens of the properties introduced in chapter 4.

5.2.1 DA, FA and EA revisited

We refer the interested reader to section 3.4 for an introduction of the concepts to be presented throughout the following section.

Recall that under the setting of **DA**, the idea is to *symmetrize* the loss function in order to *penalize* models that differ from being *equivariant.* However, nothing *forces* the resulting model to be explicitly *equivariant* in any sense. More specifically, we seek to minimize the *symmetrized* loss function: $R^G(\theta) = \int_G \mathbb{E}_\pi \left[\ell\left(\Phi_\theta^N(\rho_g.X), \hat{\rho}_g.Y\right)\right] d\lambda_G(g)$, which we can see as a functional over probability measures (i.e. *convexified*) as:

$$R^G(\mu) := \int_G \mathbb{E}_\pi \left[\ell\left(\langle\sigma_*(\rho_g.X, \cdot), \mu\rangle, \hat{\rho}_g.Y\right)\right] d\lambda_G(g) = \int_G \mathbb{E}_\pi \left[L_{\rho_g.X,\hat{\rho}_g.Y}(\mu)\right] d\lambda_G(g)$$

Analogously, under the **FA** setting, instead of dealing with the loss function, we are rather *symmetrizing the model* at hand. With this, we explicitly ensure that our models will always stay *equivariant*, though we might end up being *inefficient* (e.g. in the amount of network parameters) under such a choice. More explicitly, recall that instead of considering our *vanilla* NN model Φ_θ^N, we take the *symmetrized* version:

$$\Phi_\theta^{N,FA}(x) := \int_G \hat{\rho}_g^{-1}.\Phi_\theta^N(\rho_g.x) d\lambda_G(g) = (\mathcal{Q}(\Phi_\theta^N))(x)$$

So that, using proposition 38, $\Phi_\theta^{N,FA}(x) = \langle\sigma_*^G(x, \cdot), \nu_\theta^N\rangle$. In particular, the **FA** problem of minimizing: $R^{FA}(\theta) = \mathbb{E}_\pi\left[\ell\left(\Phi_\theta^{N,FA}(X), Y\right)\right]$ can be convexified into that of minimizing (over $\mathcal{P}(\mathcal{Z})$):

$$R^{FA}(\mu) := \mathbb{E}_\pi\left[\ell\left(\langle\sigma_*^G(X, \cdot), \mu\rangle, Y\right)\right]$$

Finally, inspired by the work of Flinth and Ohlsson [32], one could think of **EA** as the quest for optimizing[5]: $R^{EA}(\theta) = \mathbb{E}_\pi\left[\ell\left(\Phi_{P_{\mathcal{E}^G}\theta}^N(X), Y\right)\right] = \mathbb{E}_\pi\left[\ell\left(\langle\sigma_*(X, \cdot), P_{\mathcal{E}^G}\#\nu_\theta^N\rangle, Y\right)\right]$ so that the convexified population risk reads:

$$R^{EA}(\mu) := \mathbb{E}_\pi\left[\ell\left(\langle\sigma_*(X, \cdot), P_{\mathcal{E}^G}\#\mu\rangle, Y\right)\right]$$

Putting all of this together, under the suitable hypothesis for G, σ_* and ℓ, we get that:

[5] Notice that we use the fact that $P_{\mathcal{E}^G}\#\nu_\theta^N = \frac{1}{N}\sum_{i=1}^N P_{\mathcal{E}^G}\#\delta_{\theta_i} = \frac{1}{N}\sum_{i=1}^N \delta_{P_{\mathcal{E}^G}\theta_i}$

Proposition 40 *Let G be a compact group acting on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ separable Hilbert spaces. Let σ_* be jointly G-equivariant and ℓ jointly G-invariant. Then:*

$$R^G(\mu) = \int_G R(M_g \# \mu) d\lambda_G(g), \quad R^{FA}(\mu) = R(\mu^G) \ \ \textit{and} \ \ R^{EA}(\mu) = R(\mu^{\mathcal{E}^G})$$

In particular, R^G, R^{FA} and R^{EA} are G-invariant functionals over $\mathcal{P}(\mathcal{Z})$ (as in definition 4.4)

Proof. See Annex C.15. □

The expression for R^G is developed following the line established by Flinth and Ohlsson [32], where a similar calculation is carried out for $R^G(\theta)$. One shall notice that most of the upcoming theory doesn't rely on the *specific structure* underlying R; it works as long as R^G, R^{FA} and R^{EA} are defined as functions of R as in proposition 40[6]. From this point onward, we assume that σ_* *is jointly G-equivariant* and ℓ is *jointly G-invariant.* With the expressions of proposition 40 at hand, we can check that:

Proposition 41 *If $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is a convex and $\mathcal{C}^1$ functional, then R^G, R^{FA} and R^{EA} are convex and $\mathcal{C}^1$ as well, with linear functional derivatives given by:*

$$\frac{\partial R^G}{\partial \mu}(\theta, \mu) = \int_G \frac{\partial R}{\partial \mu}(M_g.\theta, M_g \# \mu) d\lambda_G(g), \quad \frac{\partial R^{FA}}{\partial \mu}(\theta, \mu) = \int_G \frac{\partial R}{\partial \mu}(M_g.\theta, \mu^G) d\lambda_G(g)$$

$$\textit{and} \quad \frac{\partial R^{EA}}{\partial \mu}(\theta, \mu) = \frac{\partial R}{\partial \mu}(P_{\mathcal{E}^G}.\theta, \mu^{\mathcal{E}^G})$$

And intrinsic derivatives given by (when well defined):

$$D_\mu R^G(\mu, \theta) = \int_G M_g^T . D_\mu R(M_g \# \mu, M_g.\theta) d\lambda_G(g)$$

$$D_\mu R^{FA}(\mu, \theta) = \int_G M_g^T . D_\mu R(\mu^G, M_g.\theta) d\lambda_G(g) \ \ \textit{and} \ \ D_\mu R^{EA}(\theta, \mu) = P^*_{\mathcal{E}^G} . D_\mu R(P_{\mathcal{E}^G}.\theta, \mu^{\mathcal{E}^G})$$

In particular, we have $\forall \mu \in \mathcal{P}(\mathcal{Z}),\ R^{FA}(\mu) \leqslant R^G(\mu)$

Proof. See Annex C.15. □

From the last part of proposition 41, one might be tempted to say that ***FA*** *generally allows for better models to be found.* However, as R^G and R^{FA} are *convex, $\mathcal{C}^1$ and G-invariant functionals* (from proposition 40 and proposition 41), thanks to Proposition 33 we can see that:

$$\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^G(\mu) = \inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R^G(\mu) \quad \text{and} \quad \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{FA}(\mu) = \inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R^{FA}(\mu)$$

i.e. these functions can be optimized merely within the realm of G-invariant probability measures. This becomes truly relevant when noticing that $\forall \mu \in \mathcal{P}^G(\mathcal{Z})$:

$$R(\mu) = R^G(\mu) = R^{FA}(\mu)$$

So that the following holds:

[6] i.e. Given $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ define $R^G(\cdot) = \int_G R(M_g \# \cdot) d\lambda_G(g)$, $R^{FA}(\cdot) = R((\cdot)^G)$ and $R^{EA}(\cdot) = R((\cdot)^{\mathcal{E}^G})$

Proposition 42 *Assume that $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is a convex and $\mathcal{C}^1$, then:*

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^G(\mu) = \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R^G(\mu) = \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R^{FA}(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{FA}(\mu)$$

In particular, whenever R is also G-invariant, from proposition 33 we get:

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^G(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{FA}(\mu)$$

Proof. Direct from the previous paragraph and proposition 33. □

Remark The first part of proposition 42 tells us that finding the *optimal measure* for the **DA** and **FA** problems corresponds, in practice, to minimizing the original functional R over the space of $\mathcal{P}^G(\mathcal{Z})$. In particular, the best *model* we can expect to find from using either **DA** or **FA** will lead to essentially the same risk value, which is the *best* that can be done by minimizing R whilst only considering *symmetric models.*

The latter part says that optimizing the original risk functional (without any sort of regularization nor *symmetrization* of any kind) will lead to a model that's at least as good as one coming from **DA** or **FA**.

We have left R^{EA} outside our analysis, as it unfortunately *doesn't necessarily satisfy* $\forall\mu \in \mathcal{P}^G(\mathcal{Z}),\ R(\mu) = R^{EA}(\mu)$ (therefore not allowing for an analog of proposition 42 to hold). The best we can do in this case is to notice that[7]:

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{EA}(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{E}^G)} R(\mu) \geqslant \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{FA}(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^G(\mu)$$

and, as noted in proposition 35, this inequality can be *strict* in the general case.

Despite the interesting insight that can be obtained from proposition 42, the second set of equalities is somewhat *to be expected.* Indeed, the following *natural* characterization of G-invariance for R tells us that whenever the original functional R is G-invariant, **DA** does *esentially nothing* (as the risk functional is *already symmetric*; this was already noticed in section 3.4).

Proposition 43 *A funcional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is G-invariant* ***if and only if*** $R = R^G$

Proof. See Annex C.15. □

Going back to the case of proposition 42 in which R is not required to be G-invariant[8] we will, in general, have the following inequalities (which might be strict in general):

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) \leqslant \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) \leqslant \inf_{\mu\in\mathcal{P}(\mathcal{E}^G)} R(\mu)$$

[7] From the fact that $\forall\nu \in \mathcal{P}(\mathcal{E}^G)$, $\nu^{\mathcal{E}^G} = \nu$, so that $R(\nu) = R^{EA}(\nu)$, and that $\forall\mu \in \mathcal{P}(\mathcal{Z})$, $\mu^{\mathcal{E}^G} \in \mathcal{P}(\mathcal{E}^G)$, so that $R^{EA}(\mu) = R(\mu^{\mathcal{E}^G}) \geqslant \min_{\nu\in\mathcal{P}(\mathcal{E}^G)} R(\nu)$

[8] e.g. in the learning setting with a data distribution $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})\backslash\mathcal{P}^G(\mathcal{X}\times\mathcal{Y}))$

An interesting question that immediately arises is whether the first two values are *close* to each other (or not)[9]. In the specific case of the *learning problem*, one could particularly be interested in quantifying *how far the risk R is from being G-invariant whenever the data distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is **approximately** G-invariant*[10]. In particular, following the work of Chen et al. [14], Lyle et al. [53] (as in lemma 11), one gets the following bound (with a simple proof that we'll restate for completeness):

Proposition 44 *Let $L^\mu : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be given by $L^\mu(\cdot,\cdot) = L_{\cdot,\cdot}(\mu)$ and set $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}$, $\forall g \in G$, $g.(x,y) = (\rho_g.x, \hat{\rho}_g.y)$. Then $\forall \mu \in \mathcal{P}(\mathcal{Z})$:*

$$|R(\mu) - R^G(\mu)| \leqslant \int_G W_1(L^\mu \# g \# \pi, L^\mu \# \pi) d\lambda_G(g)$$

Furthermore, whenever there exists a constant $C > 0$ such that $\forall \mu \in \mathcal{P}(\mathcal{Z})$, L^μ is C-Lipschitz, then:

$$\sup_{\mu \in \mathcal{P}(\mathcal{Z})} |R(\mu) - R^G(\mu)| \leqslant C \int_G W_1(g \# \pi, \pi) d\lambda_G(g)$$

In particular, it is enough to suppose ℓ and σ_* to be Lipschitz for this to hold.

Proof. See Annex C.15. □

Remark This proposition has two quite major drawbacks. On the one hand, in order to have the *interesting bound* relating R and R^G uniformly over $\mathcal{P}(\mathcal{Z})$; we need to require *too much* of both ℓ and σ_* (e.g. even the quadratic loss *isn't globally Lipschitz over $\mathcal{Z}$*). On the other hand, the obtained bound isn't exactly the one to be expected: as shown in proposition 25, $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is G-invariant **if and only if** $\pi = \pi^G$; and one might therefore expect a bound involving $W_1(\pi, \pi^G)$ (which represents the *distance to G-invariance more clearly*) instead of the one we get. Despite this criticism, the current bound still allows for significant analysis: thanks to the remark done after proposition 25, we know that the bound becomes 0 if and only if $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$.

Remark Proposition 44, in the strongest case, allows us to *control*, from the *approximate invariance of* π, *how far* **DA** (and **FA**) (in the best case) are able to get to the real *optimum* of R (which is in principle NOT exactly G-invariant). In particular, if $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is such that for some $\varepsilon > 0$: $\int_G W_1(g \# \pi, \pi) d\lambda_G(g) < \varepsilon$; then, by proposition 44, we directly have:

$$\sup_{\mu \in \mathcal{P}(\mathcal{Z})} |R(\mu) - R^G(\mu)| \leqslant C\varepsilon$$

In particular, any infimizing sequence for R, $(\mu_n)_{n \in \mathbb{N}} \subseteq \mathcal{P}(\mathcal{Z})$ will satisfy:

$$\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^G(\mu) \leqslant R^G(\mu_n) \leqslant R(\mu_n) + C\varepsilon$$

[9] We know that, in general, the third value can be strictly larger than the other two; though the question of *under which conditions* we can ensure it will be *close* to them is definitely of high interest for future work.

[10] See chapter 3 for a broad idea about *approximate equivariance*; also Chen et al. [14] provides a good reference for the result presented.

so that, by taking the limit, we get: $\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^G(\mu) - \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) \leqslant C\varepsilon$ or, in particular:

$$\left| \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) - \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) \right| \leqslant C\varepsilon$$

i.e. whenever π is *close to being G-invariant*, *not much will be lost* by optimizing directly over $\mathcal{P}^G(\mathcal{Z})$ (or, in other words, the inductive bias introduced by the use of **DA** or **FA** shouldn't really *harm* the optimization procedure).

On a different note, by proposition 43 (or section 3.4), we know that whenever R is G-invariant, $R = R^G$, and so **DA** *changes nothing* about the problem itself. One may then ask oneself: *why* has **DA** become such a popular approach for *exploiting* a problem's symmetries, despite it changing *nothing* of the population risk? The answer is that the *SGD training dynamic might change* under **DA**, potentially allowing for a *more effective training* of the NN. In particular, recall that the usual (simple)[11] SGD training dynamic for a sample $(X_k, Y_k)_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \pi$, learning rate s_k^N and parameters $\theta^k = (\theta_i^k)_{i=1}^N$ of empirical measure ν_k^N (for each iteration $k \in \mathbb{N}$); can be written as:

$$\theta_i^{k+1} = \theta_i^k - s_k^N D_\mu L_{X_k,Y_k}(\nu_k^N, \theta_i^k)$$

In the particular case of **DA**, **FA** and **EA**, these iterations would take the following form:

$$\theta_i^{k+1} = \theta_i^k - s_k^N D_\mu L^G_{X_k,Y_k}(\nu_k^N, \theta_i^k) = \theta_i^k - s_k^N \int_G D_\mu L_{\rho_g.X_k,\hat{\rho}_g.Y_k}(\nu_k^N, \theta_i^k) d\lambda_G(g) \tag{5.2}$$

$$\theta_i^{k+1} = \theta_i^k - s_k^N D_\mu L^{FA}_{X_k,Y_k}(\nu_k^N, \theta_i^k) = \theta_i^k - s_k^N \int_G D_\mu L_{\rho_g.X_k,\hat{\rho}_g.Y_k}((\nu_k^N)^G, \theta_i^k) d\lambda_G(g) \tag{5.3}$$

and

$$\theta_i^{k+1} = \theta_i^k - s_k^N D_\mu L^{EA}_{X_k,Y_k}(\nu_k^N, \theta_i^k) = \theta_i^k - s_k^N P_{\mathcal{E}^G} D_\mu L_{X_k,Y_k}((\nu_k^N)^{\mathcal{E}^G}, P_{\mathcal{E}^G}\theta_i^k) \tag{5.4}$$

respectively. At the *training* level, these iterations are (indeed) *different* from the usual SGD training dynamics, and we would expect them to produce *some kind of advantage* when the data is known to be G-invariant.

On the other hand, at the level of the *MF training dynamics only the population risk R intervenes*, meaning that the **DA** and the *vanilla* distributional dynamics (corresponding to the WGF of R^G and R respectively) will exactly coincide. Furthermore, we will see in the upcoming section that whenever the *initial condition* for the dynamics is G-invariant (i.e. $\mu_0 \in \mathcal{P}^G(\mathcal{Z})$), this will also coincide with the WGF of R^{FA}.

With all of this, it seems like, when R is known to be G-invariant, the MF dynamic will be mostly *unaffected* by the use of **DA** and **FA**. Therefore, one should expect the *advantages* of **DA** and **FA** to play a role mostly in the *transition* between the discrete SGD Dynamics and the MF limit process: maybe *propagation of chaos* (as in theorem 3) occurs *faster* when these techniques are employed during training[12]; or maybe other similar quantitative advantages can be derived. Unfortunately these last questions have proven hard to tackle, and so truly

[11] The argument easily adapts to equation (2.2)

[12] For instance, one might expect to *improve the POC bounds* by some factor depending solely on the group.

understanding what the advantages of **DA** and **FA** look like concretely is an open question that we will have to leave for future work.

In the following section we will dive deeper into the properties of the NN training dynamic (a WGF at the MF level), whenever the functional involved is assumed to be *symmetric*. We suggest reviewing chapter 2 for a deeper dive into the different MF Limit results from the literature.

5.3 Symmetries for a WGF

Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be some *risk* functional (e.g. that of the *learning problem*). We will study the formal setting of the *regularized risk minimization*, with:

$$R^{\tau,\beta}(\mu) = R(\mu) + \tau \int r d\mu + \beta H_\lambda(\mu)$$

Where $r : \mathcal{Z} \to \mathbb{R}$ is a *regularization* term and $H_\nu(\mu) := D(\mu||\nu) = \int \log(\frac{d\mu}{d\nu}(z))d\mu(z)$ is the *relative entropy* between μ and ν (with $\mu \lll \nu$). Notice that we consider the *Lebesgue measure* λ for our entropy regularization term. Recall from section 2.2.7 that the WGF associated to this functional, which is what *ultimately* the MF limiting process for SGD training will follow, is given by (in the case of a *general learning rate* of the form $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$):

$$\partial_t \mu_t = \varsigma(t) \left[\text{div} \left(\left(D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r \right) \mu_t \right) + \beta \Delta \mu_t \right] \tag{2.5}$$

Notice that by setting $\tau, \beta = 0$ we recover the WGF for R. Whenever $\beta = 0$, this equation is often known to have a (unique) solution *distributionally*; but for $\beta > 0$ the solutions to this equation are actually *strong*. This equation (2.5) is actually equivalent to the following *McKean-Vlasov nonlinear SDE*:

$$dZ_t = \varsigma(t) \left[- \left(D_\mu R(\mu_t, Z_t) + \tau \nabla_\theta r(Z_t) \right) dt + \sqrt{2\beta} dB_t \right] \quad \text{with} \quad \mu_t = \mathbf{Law}(Z_t) \tag{2.6}$$

Where $(B_t)_{t \geqslant 0}$ is a D-dimensional standard Brownian Motion.

The following (general) result holds:

Proposition 45 *Let G be a compact group acting orthogonally on the (separable) Hilbert space $\mathcal{Z}$ (via the representation M). Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a G-invariant functional, such that the following WGF (distributional dynamics) is well defined and has a unique (weak) solution:*

$$\partial_t \mu_t = \varsigma(t) \left[\text{div} \left(D_\mu R(\mu_t, \cdot) \, \mu_t \right) \right]$$

If the initial condition satisfies $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, then: λ-a.s. $\forall t \geqslant 0,\ \mu_t \in \mathcal{P}_2^G(\mathcal{Z})$

Proof. See Annex C.16. □

This, in particular, lets us conclude the following:

Theorem 14 *Let $\mathcal{Z} = \mathbb{R}^D$, and let G be a compact group acting orthogonally on via the representation M. Consider a functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ that's convex and of class $\mathcal{C}^1$ and such that assumption 7 holds (or any other such that the WGF is well defined and has a unique solution, such as assumption 10). Then, $\forall \mu_0 \in \mathcal{P}_2(\mathcal{Z})$:*

1. *The DD associated to $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, $\partial_t \mu_t = \varsigma(t) \left[\operatorname{div}\left(D_\mu R(\mu_t, \cdot)\, \mu_t\right)\right]$ has a unique (weak) solution.*
2. *The DD for functional $R^\tau(\mu) = R(\mu) + \tau \int r d\mu$ (with $\tau > 0$ and regularizer $r : \mathcal{Z} \to \mathbb{R}$): $\partial_t \mu_t = \varsigma(t) \left[\operatorname{div}\left((D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r)\, \mu_t\right)\right]$ has a unique (weak) solution.*
3. *The DD for functional $R_\nu^{\tau,\beta}(\mu) = R(\mu) + \tau \int r d\mu + \beta H_\nu(\mu)$ (with $\tau, \beta > 0$, regularizer $r : \mathcal{Z} \to \mathbb{R}$ and $\nu \in \mathcal{P}(\mathcal{Z})$ the Gibbs measure ($\nu \lll \lambda$) of potential $U : \mathcal{Z} \to \mathbb{R}$): $\partial_t \mu_t = \varsigma(t) \left[\operatorname{div}\left((D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r + \beta \nabla_\theta U)\, \mu_t\right) + \beta \Delta \mu_t\right]$ has a unique strong solution.*

Assume R, r and U to be G-invariant functions. If the initial condition satisfies $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, then: λ-a.s. $\forall t \geqslant 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$.

In the case of strong solutions (of density wrt λ given by $(u_t)_{t \geqslant 0}$), this means that $\forall t \geqslant 0$, $u_t = u_t^G = \int_G u_t(M_g.\cdot) d\lambda_G(g)$ (i.e. $\mathcal{S}.u_t = u_t$ with $\mathcal{S}$ the symmetrization operator defined in definition 3.8).

Proof. See Annex C.16. □

Remark Many interesting remarks are in place:

1. This result tells us exactly what we expected (and what was famously hinted by Mei et al. [57] acquires a concrete meaning): if the functional we work with is G-invariant, then the trajectory of the WGF will be G-invariant as well. In particular, when solutions are strong (and have a density), this density is a G-invariant function (λ-a.s. on $\mathcal{Z}$). This in turn allows us to solve the Fokker Planck PDE by only looking for G-invariant solutions (i.e. we restrict our search to functions that *respect* the symmetry of the problem). This can potentially help us *reduce the dimensionality of the problem* (as remarkably done by Mei et al. [57] for the action of $O(D)$ on $\mathbb{R}^D$, allowing them to look for solutions only as densities of measures over $\mathbb{R}_+$).
2. Notice that, transporting these ideas to the learning setting, for the usual functional $R(\mu) = \mathbb{E}_\pi \left[\ell(\langle \sigma_*(X; \cdot), \mu \rangle, Y)\right]$, the result is valid for an *arbitrary* Neural Network, without employing any kind of *symmetry-leveraging technique* (i.e. no **DA**, **FA** or **EA** is required). The only requirement is for the data distribution π to be G-invariant. In some sense, by taking the MF limit (in which both the amount of hidden parameters and the SGD iterations go to infinity), the resulting process *has already incorporated the symmetries from the data* (from the *infinite* SGD iterations considered when taking the MF limit), and the result is a limiting dynamic that *respects the G-invariance of the initial distribution all along the training process.*
3. The condition of having $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ isn't truly restrictive from the point of view of applications. In reality, the training process of NNs is usually initialized with an i.i.d.

> *Gaussian distribution* (i.e. $\overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{Id}_D)$) which, in particular, is G-invariant with respect to *any* orthogonal representation. This comes also with the disadvantage that: the fact that $\forall t \geqslant 0,\ \mu_t \in \mathcal{P}^G(\mathcal{Z})$ might not be *very informative* about the nature of the distribution at play (notably, it could *not truly improve*, and remain i.i.d. gaussian along the entire training process). Now, understanding whether this is the case (or not) during training is a clearly open question that we shall tackle in future work.

Now, recalling chapter 2, there are also conditions under which the MF dynamic for the **regularized** problem satisfies *global convergence.* For leveraging such a result in our context, we consider the following general proposition:

Proposition 46 *Let $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_p(\mathcal{Z})$ be a flow of measures such that $W_p(\mu_t, \mu_*) \xrightarrow[t\to\infty]{} 0$ for some $\mu_* \in \mathcal{P}_p(\mathcal{Z})$ (i.e. it converges in the Wasserstein metric). Then, we must have:*

- *If $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_p^G(\mathcal{Z})$, then $\mu_* \in \mathcal{P}_p^G(\mathcal{Z})$*
- *If $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$, then $\mu_* \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$*

Proof. Direct from Proposition 25. □

Remark Proposition 46 has a particularly useful consequence: it guarantees that, whenever the WGF converges to some measure $\mu_* \in \mathcal{P}_2(\mathcal{Z})$, as long as the WGF *stayed* within a given space (either $\mathcal{P}^G(\mathcal{Z})$ or $\mathcal{P}(\mathcal{E}^G)$) during the dynamic, then μ_* will have to still lie within in the same space.

This leads to the following particularly interesting result:

Corollary 9 *Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be G-invariant, let $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ and suppose that assumption 9 holds. Consider the WGF of R initialized at μ_0, $(\mu_t)_{t\geqslant 0}$. By theorem 5, we know that if $(\mu_t)_t$ converges to some $\mu_\infty \in \mathcal{P}_2(\mathcal{Z})$ in W_2, then μ_∞ is a global minimizer of R over $\mathcal{P}(\mathcal{Z})$. In particular, by theorem 14 together with proposition 46, we know that $\mu_\infty \in \mathcal{P}_2^G(\mathcal{Z})$.*

Proof. Direct from theorem 14 and proposition 46. □

Remark In particular, *when* the dynamic converges (starting from $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$), the *optimum* that's achieved through the dynamic corresponds also to a G-invariant measure (and, in particular, realizes $\min_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu)$).

Remark In the *regularized* case, though the same result holds, not much insight is obtained, since we *already knew* (from proposition 33 together with the *uniqueness* from proposition 9), that *the* optimum that can be attained has to be G-invariant.

Now, we know that a given Fokker-Planck equation (the DD) always has an associated McKean-Vlasov non-linear SDE. Under the good conditions, we know that both systems are in correspondence (and the existence/uniqueness of solutions translates from one to the other). In particular, we highlight the following result:

Corollary 10 *Under the assumptions of theorem 14, let $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$ and consider the McKean-Vlasov non-linear SDE for the functional $R_\nu^{\tau,\beta}$, given by:*

$$dZ_t = \varsigma(t)\left[-\left(D_\mu R(\mu, Z_t) + \tau\nabla_\theta r(Z_t) + \beta\nabla_\theta U(Z_t)\right)dt + \sqrt{2\beta}dB_t\right] \quad \textit{with } \mu_t = \textbf{\textit{Law}}(Z_t)$$

If $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, $r : \mathcal{Z} \to \mathbb{R}$ and $U : \mathcal{Z} \to \mathbb{R}$ are G-invariant functions and the initial condition satisfies $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$; then the unique solution (strong if $\beta > 0$) of the system satisfies: λ-a.s. $\forall t \geqslant 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$ (also, when $\beta > 0$, the density functions are G-invariant).

Proof. Direct from the correspondence between the Fokker-Planck equation and the McKean-Vlasov SDE (see Sznitman [83]) and theorem 14.

We do however provide an independent proof in section C.16. □

In particular, cases which haven't been explicitly covered in the above description (such as the $\upsilon = 1$ case of Bortoli et al. [9]), also benefit from such result:

Corollary 11 *Consider the setting of Bortoli et al. [9] (under assumption 8 4.), let $\zeta \in [0, 1)$, $\upsilon \in [0, 1]$ and $\alpha > 0$; define $\alpha^N_{\zeta,\upsilon} = (\alpha N^{(\upsilon-1)})^{1/(1-\zeta)}$ and consider: $\varepsilon_N = \alpha^N_{\zeta,\upsilon}$, $\varsigma(t) = (1 + t)^{-\zeta}$, such that $\forall k \in \mathbb{N}$, $s^N_k = \varepsilon_N\varsigma(k\varepsilon_N)$. Also, consider a fixed batchsize $B \in \mathbb{N}^*$, an initial measure $\mu^0 \in \mathcal{P}_2(\mathcal{Z})$, and fix $\upsilon = 1$. Under this setting, the McKean-Vlasov dynamic reads:*

$$dZ_t = \varsigma(t)\left[-\left(D_\mu R(\mu_t, \cdot) + \tau\nabla_\theta r(Z_t)\right)dt + \left(\sqrt{\frac{\alpha}{B}}\sqrt{\Sigma}(\mu_t, Z_t)\right)dB_t + \sqrt{2\beta}d\tilde{B}_t\right] \tag{5.5}$$

Where $\mu_t = Law(Z_t)$, $(B_t)_{t\geqslant 0}$ and $(\tilde{B}_t)_{t\geqslant 0}$ are (independent) D-dimensional Brownian Motions, and $\Sigma(\mu, \theta) := \mathbb{E}_\pi\left[\left(D_\mu L^\tau_{X,Y}(\mu, \theta) - D_\mu R^\tau(\mu, \theta)\right) \otimes \left(D_\mu L^\tau_{X,Y}(\mu, \theta) - D_\mu R^\tau(\mu, \theta)\right)\right]$, where $\otimes$ represents the outer product (or generally, the tensor product) between vectors in $\mathcal{Z} = \mathbb{R}^D$.

As shown in Section D.1.3 (from Bortoli et al. [9]), it is known that this system has a unique (in a trajectory-wise sense) ***strong*** *solution. Moreover, if $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, $\forall t \geqslant 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$*

With these results at hand, one might be tempted to ask whether a similar result holds but replacing $\mathcal{P}^G(\mathcal{Z})$ with $\mathcal{P}(\mathcal{E}^G)$. The answer is rather *mixed*, since, on the **regularized case**, we have that $R_\nu^{\tau,\beta}(\mu) < \infty \implies \mu$ has a density with respect to λ. In particular, if $G \circlearrowright_M \mathcal{Z}$ is non-trivial, then $\mathcal{E}^G$ is a strict subspace of $\mathcal{Z}$ and, by proposition 26, $\forall \mu \in \mathcal{P}(\mathcal{E}^G)$, μ doesn't have a density wrt λ. In other words, initializing the WGF of $R_\nu^{\tau,\beta}$ at $\mu_0 \in \mathcal{P}(\mathcal{E}^G)$ (which, by the way, would mean *starting with* $R_\nu^{\tau,\beta}(\mu_0) = \infty$), by the *regularizing effect* of the entropy, implies that $\forall t > 0$, μ_t would have a density wrt λ, and hence $\mu_t \notin \mathcal{P}(\mathcal{E}^G)$. Nevertheless, in the **noiseless** setting, we do get the following positive result:

Theorem 15 *Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ a* ***convex****, $\mathcal{C}^1$ and G-invariant functional, and $r : \mathcal{Z} \to \mathbb{R}$ a G-invariant function; such that assumption 10 (or assumption 8 4.) holds. Consider the*

McKean-Vlasov *dynamic of $R^{\tau} = R + \tau\langle r, \cdot\rangle$ (which has a* ***pathwise-unique*** *solution according to Bortoli et al. [9])*[13]*:*

$$dZ_t = \varsigma(t)\left[-\left(D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r(Z_t)\right) dt\right] \quad \textit{and} \quad \textit{Law}(Z_0) = \mu^0 \ (\textit{initial condition}) \tag{5.6}$$

where $\mu_t = \textit{Law}(Z_t)$ and $\mu^0 \in \mathcal{P}_2(\mathbb{R}^D)$ is a fixed initial condition. If $\mu^0 \in \mathcal{P}_2(\mathcal{E}^G)$, then

$$\forall t \geqslant 0,\ \mu_t \in \mathcal{P}_2(\mathcal{E}^G)$$

Proof. See Annex C.16. □

Remark Once again, multiple remarks are in place:

1. The result of theorem 15 can be actually made stronger, as we have the following equivalence:

   $$\left[\mathbb{P}\left(\forall t \in [0,T], Z_t \in \mathcal{E}^G\right) = 1\right] \iff \left[\forall t \in [0,T],\ \mu_t(\mathcal{E}^G) = 1\right]$$

   This follows from a *continuity* argument for the solutions of the SDE. The proof shall be found in Annex C.16.

2. Once again, Theorem 15, translated to the learning problem, proves that, *when the training data is symmetric*, even a *vanilla shallow NN* can manage to *respect* the symmetry that's imposed in the initial condition.

   More remarkably, the theorem actually states the following: in the mean field limit, a neural network that *began* its training *concentrated* in the space of G-invariant parameters, will actually *stay concentrated in $\mathcal{E}^G$* all throughout its training, despite the fact that there is *no constraint on the network whatsoever during training.* That is: during training *any given parameter could be updated*, but thanks to the *symmetries* of the problem, they are only updated in a way such that they remain (all throughout training) within $\mathcal{E}^G$. Once again, the *macroscopic vision* provided by the mean field limit (where SGD iterations and hidden units have been sent to ∞), makes the *symmetries* (which are seen at training time only through *samples* of the data) *explicit* for the asymptotic training dynamic.

3. Unfortunately, if $\mu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, as $\mathcal{E}^G$ is (in general) a lower dimensional *subspace*, the conditions of assumption 9 do not hold in this setting (since μ_0 cannot *separate* many sets). Therefore, an analogue of corollary 9 isn't straightforward to establish. In any case, even if it were possible to establish such a result, the fact that by proposition 35 we might have $\inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu) > \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu)$, implies that the WGF starting from $\mu_0 \in \mathcal{P}(\mathcal{E}^G)$ *could not converge* in $(\mathcal{P}_2(\mathcal{Z}), W_2)$ (in such a particular example). If the dynamic did converge, the limit would have to be some $\nu_* \in \mathcal{P}(\mathcal{E}^G)$ (by Proposition 46) that, at the same time, should be a *global minimizer* of R over $\mathcal{P}(\mathcal{Z})$; a contradiction.

4. This result bears some resemblance to the findings of Flinth and Ohlsson [32] (e.g., their *Corollary 1*), where they discuss the *stability* of the space $\mathcal{E}^G$ under the gradient flow of the *augmented risk* $[\theta \mapsto R^G(\theta)]$ (which arises from the use of Data Augmentation).

[13] Notice that this is actually an ODE with the only *randomness* involved in the initial condition

Essentially, it implies that if we *start with parameters in* $\mathcal{E}^G$, then the dynamics will never take us out of it. Our result shares a similar flavor for the dynamics in the MF limit.

5. In many ways, the consequence of theorem 15 is significantly *stronger* than that of theorem 14 (telling us that the dynamic will respect symmetry in a much stricter way than anticipated). However, finding an initial condition $\mu^0 \in \mathcal{P}(\mathcal{E}^G)$ isn't as simple as in the case of theorem 14 (where a simple *iid gaussian* was enough). The question of actually computing what the space $\mathcal{E}^G$ looks like is itself a really interesting and complex question (which has been recently tackled by Finzi et al. [31]). In particular, it may involve significant computational burden which might not be desired.

   Beyond all of this, the most *significant* limitation (in our context) from considering $\mu_0 \in \mathcal{P}(\mathcal{E}^G)$, is that $\mathcal{E}^G$ is a degenerate subspace which might not satisfy all the properties we desire (at least, seen as a subspace of $\mathcal{Z}$). Overcoming this limitation is one of the big challenges that will have to be studied in our future work.

One might desire to *extend* the result of theorem 15 to a *noisy setting*, where *guarantees of convergence* could be established. However, as noted in a previous remark, this can't be done through the addition of entropy over the entire space $\mathcal{Z}$. The solution is to consider an entropy regularizer against a measure *concentrated on* $\mathcal{E}^G$.

More specificaly, consider the following SGD training dynamic for $(\theta^k) \in \mathcal{Z}^N$ and $\tau, \beta > 0$:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\frac{1}{B_k} \sum_{j=1}^{B_k} D_\mu L^\tau_{X_j^k, Y_j^k}(\nu^N_{\theta^k}, \theta_i^k) \right) + \sqrt{2\beta s_k^N} P_{\mathcal{E}^G} \xi_i^k$$

Where $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$ is the learning rate and $(\xi_i^k) \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{Id}_D)$. Notice that we are *projecting the noise onto the subspace* $\mathcal{E}^G$ (using the orthogonal projection $P_{\mathcal{E}^G}$), in order to *restrict the exploration* within the realm of *equivariant parameters*. We will call this the *projected noisy SGD* training dynamics. One can infer that, under this training regime, the corresponding *mean-field* limit dynamic (i.e. the McKean-Vlasov equation) is written as:

$$dZ_t = \varsigma(t)[-\left(D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r(Z_t)\right) dt + \sqrt{2\beta} P_{\mathcal{E}^G} dB_t] \ \ \textit{with} \ \ \mathrm{Law}(Z_t) = \mu_t \qquad (5.7)$$

With $(B_t)_{t \geqslant 0}$ a D-dimensional Brownian motion. This corresponds to a WGF of an entropy-regularized functional with respect to a measure concentrated in $\mathcal{E}^G$. We can prove an analogue of Theorem 15 in this context:

Theorem 16 *Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ a convex, $\mathcal{C}^1$ and G-invariant functional, and $r : \mathcal{Z} \to \mathbb{R}$ a G-invariant function; such that assumption 10 (or assumption 8 (4.)) holds. Consider the Projected Mean Field Langevin Dynamic given by equation* (5.7) *(which has a pathwise-unique solution according to Bortoli et al. [9]). Let $\mu^0 \in \mathcal{P}_2(\mathbb{R}^D)$ be a fixed initial condition. If $\mu^0 \in \mathcal{P}_2(\mathcal{E}^G)$, then*

$$\forall t \geqslant 0, \ \mu_t \in \mathcal{P}_2(\mathcal{E}^G)$$

Proof. The proof is esentially the same as for theorem 15; only that we can now invoke the *pathwise uniqueness* even with the presence of noise, thanks to the appearence of the projection $P_{\mathcal{E}^G}$ in front of the stochastic integral.

We refer the reader to Annex C.16 for the fleshed out details. □

Remark In this case, theorem 16 has the advantage of allowing us to incorporating *noise* into our training dynamic, while *still* leaving the resulting process $(\mu_t)_{t\geqslant 0}$ *concentrated* on $\mathcal{E}^G$ (if the initial condition satisfied that as well).

On the other hand, this result was to be expected, since we are *explicitly forcing our noise* to live in $\mathcal{E}^G$ (and so, we rely on exactly the same principle employed in theorem 15, whilst explicitly introducing a *bias* by making the noise *not leave* such space).

One shall also notice that in this case *still all of the network parameters are trainable* (the only thing we are projecting is the *noise*); but analogously to theorem 15, the projected **MFLD** is well behaved enough to *maintain* the parameters within $\mathcal{E}^G$.

To close off this section, one may wonder whether *seeing $\mathcal{E}^G$ as a subspace of an ambient space $\mathcal{Z}$* makes any sense at all: why not consider some $\tilde{\mathcal{Z}} = \mathcal{E}^G$ directly? Indeed, the general setting under which our theory has been developed allows straightforwardly for such an extension:

Corollary 12 *Let $\tilde{\mathcal{Z}} = \mathcal{E}^G \leqslant \mathcal{Z} = \mathbb{R}^D$ be the space of G-equivariant parameters (in this case, regarded directly as a vector space $\mathcal{E}^G \cong \mathbb{R}^{\tilde{D}}$). Notice that G acts trivially on this space, so no considerations of G-invariance will be needed. Consider a functional $R : \mathcal{P}(\tilde{\mathcal{Z}}) \to \mathbb{R}$ that's convex and of class $\mathcal{C}^1$ and such that assumption 7 holds (with $\tilde{\mathcal{Z}}$ instead of $\mathcal{Z}$). Then, $\forall \mu_0 \in \mathcal{P}_2(\tilde{\mathcal{Z}})$, the DD for functional $R^{\tau,\beta}_{\nu}(\mu) = R(\mu) + \tau \int r d\mu + \beta H_\nu(\mu)$ (with $\tau, \beta > 0$, regularizer $r : \tilde{\mathcal{Z}} \to \mathbb{R}$ and $\nu \in \mathcal{P}(\tilde{\mathcal{Z}})$ the Gibbs measure ($\nu \lll \lambda$) of potential $U : \tilde{\mathcal{Z}} \to \mathbb{R}$): $\partial_t \mu_t = \varsigma(t) \left[\mathrm{div}\left(\left(D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r + \beta \nabla_\theta U\right) \mu_t\right) + \beta \Delta \mu_t\right]$ has a unique strong solution $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}^G_2(\tilde{\mathcal{Z}})$. Further assume that $\tilde{\mathcal{Z}}$ satisfies assumption 4. By theorem 8, there is global W_2 convergence to the optimum of the problem* $\inf_{\mu\in\mathcal{P}(\tilde{\mathcal{Z}})} R(\mu)$

Proof. This comes directly from the standard results presented in chapter 2. □

Remark Actually, the really interesting version of this is when we apply it in the following setting: Let $\mathcal{Z} = \mathbb{R}^D$ and $\tilde{\mathcal{Z}} = \mathcal{E}^G$, and let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be the convex and $\mathcal{C}^1$ functional we are trying to optimize. Then we can apply corollary 12 to the convex and $\mathcal{C}^1$ functional $R|_{\mathcal{P}(\mathcal{E}^G)} : \mathcal{P}(\mathcal{E}^G) \to \mathbb{R}$. In particular, from theorem 5, if the **WGF** $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{E}^G)$ of R converges in W_2 to some measure $\mu_* \in \mathcal{P}(\mathcal{E}^G)$, then such measure achieves the infimum of the problem, $\inf_{\nu\in\mathcal{P}(\mathcal{E}^G)} R(\nu)$ (which, we recall, could be quite far away from the actual *global minimum*). A truly interesting question, which we haven't been able to tackle yet, is whether this value will be too far off the actual *global infimum*.

In the case where proposition 36 holds, solving the problem directly over $\mathcal{E}^G$ could actually be much more convenient, since we are working over a space of reduced dimensionality. We haven't, however, managed to fully grasp what the precise theoretical and practical advantages of working on a lower-dimensional space are. This shall be tackled as part of our future work.

On the other hand, the global convergence results obtained will give us a guarantee of *convergence* of the **WGF** $(\mu_t)_{t\geqslant 0}$ of $(R|_{\mathcal{P}(\mathcal{E}^G)})^{\tau,\beta}_{\nu}$ to its global minimum, i.e. $\inf_{\mu\in\mathcal{P}(\mathcal{E}^G)} R(\mu)+$

$\tau\langle r, \mu\rangle + \beta H_\nu(\mu)$. Thanks to proposition 10, one should expect that with *arbitrarily small* regularization parameters, this quantity should (in some sense) *approximate* the value of $\inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu)$. Under the setting of proposition 36, this could also allow for an interesting comparison between the optimizers of $(R|_{\mathcal{P}(\mathcal{E}^G)})_\nu^{\tau,\beta}$ and $R_\nu^{\tau,\beta}$ as they approach the common value of R^*. Unfortunately, this will also have to be left for future work.

We will finalize this chapter with a small digression about how different *techniques for leveraging symmetry* behave under the MF training dynamic.

5.3.1 Neural Networks that Exploit Symmetry

From the insights obtained from previous sections, we have seen that *vanilla* NNs are able to *learn* the invariance directly from training with *symmetric* data. It is therefore interesting to ask whether **DA** of **FA** give any advantage over these vanilla NNs in terms of the MF Limit.

As noted in section 5.2.1, whenever the original functional $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is G-invariant (e.g. when $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$), there's actually no difference at all between R and R^G. Thus, at least at the level of the MF training dynamic, there is no explicit *gain* from doing **DA** in this case. On the other hand, **FA** is *not exactly equal* to R, even when R is G-invariant, so one might expect to find some *difference* at the level of the WGF. However, the following proposition (coming directly from our results on section 5.3) argues strongly against this:

Corollary 13 *Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a convex, $\mathcal{C}^1$ and G-invariant functional; and let R^{FA} be its **FA** version (i.e. $R^{FA}(\mu) = R(\mu^G)$, $\forall \mu \in \mathcal{P}(\mathcal{Z})$).*

*Assume the conditions for theorem 14 hold. Then, if $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, the **WGF** dynamics of R^{FA} and R starting at μ_0 coincide.*

Proof. See Annex C.16. □

Remark Some remarks are in place:

1. This result, despite seeming *simple* (as from section 5.2.1 we already knew that R, R^G and R^{FA} coincide on $\mathcal{P}^G(\mathcal{Z})$), provides a really interesting insight into the relationship between these functionals:

   We manage to prove that, *whenever the dynamic lies completely within* $\mathcal{P}^G(\mathcal{Z})$, the *natural* conclusion follows, which is that the entire dynamic for R^{FA} and R looks *exactly the same*. However, this isn't *entirely* obvious a priori. What's truly remarkable (as well as intuitive) is that the proof crucially requires R to be G-invariant. In particular, this speaks to the fact that, even knowing that $R = R^{FA}$ on $\mathcal{P}^G(\mathcal{Z})$ and even if both dynamics are *launched* from the same $\mu_0 \in \mathcal{P}^G(\mathcal{Z})$; if R isn't G-invariant, the processes won't necessarily coincide overtime. In particular, this is because R^{FA} is *always* a G-invariant functional (and therefore such a dynamic will always *stay G-invariant overtime*); whereas if R isn't G-invariant, nothing guarantees that the WGF process will *stay* within $\mathcal{P}^G(\mathcal{Z})$.

2. Corollary 13 could actually be rephrased as follows:

   Corollary 14 *Let $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a convex, $\mathcal{C}^1$ functional. Let R^G and R^{FA} be its **DA** and **FA** versions respectively (i.e. $R^G(\mu) = \int_G R(M_g \# \mu)$ and $R^{FA}(\mu) = R(\mu^G)$, $\forall \mu \in \mathcal{P}(\mathcal{Z})$ respectively).*

   *Assume the conditions for theorem 14 hold. Then, if $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, the **WGF** dynamics of R^{FA} and R^G starting at μ_0 coincide. In particular, whenever R is G-invariant, they are both equal to the WGF process for R starting at μ_0.*

   The true connection is actually at the level of R^G and R^{FA} (in general): both are G-invariant functionals that coincide within the class $\mathcal{P}^G(\mathcal{Z})$. The conclusion is what's expected: launching them both from the same G-invariant measure leads to the exact same dynamic overtime.

3. Naturally, as R^{EA} *projects* measures onto $\mathcal{P}(\mathcal{E}^G)$ (and more crucially, doesn't coincide with the others over $\mathcal{P}^G(\mathcal{Z})$), we can't expect the same result to hold. Unfortunately, to our knowledge, even if $\mu_0 \in \mathcal{P}(\mathcal{E}^G)$, the dynamics of R and R^{EA} won't necessarily coincide. Despite this, a connection between the dynamics of $R|_{\mathcal{P}(\mathcal{E}^G)}$ and R^{EA} seems more promising, and shall be studied in future work.

Another immediate result is the following:

Corollary 15 *Even if $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is convex, $\mathcal{C}^1$, but not necessarily G-invariant, R^G, R^{FA} and R^{EA} are all G-invariant (by proposition 40). Thus, if the right conditions are assumed, all of theorem 14, theorem 15, theorem 16, corollary 9, corollary 10 and corollary 12 will hold true.*

Proof. Direct. □

Finally, in this work we have studied multiple of the different objects related to both the *mean field limit* of shallow NNs and the *geometric deep learning* analysis of *symmetries* in NN models. We have succesfully managed to put both theories under a common *umbrella*, in such a way that important properties and behaviours can be thoroughly understood from either viewpoint. In general, the results we present are novel, and hopefully they shall constitute a significant contribution to the ever-growing literature of Machine Learning and Neural Networks (at least at a theoretical level). Many questions are still left open (as has been mentioned throughout this work), and they shall be an unavoidable part of our future challenges to tackle.

Chapter 6

Conclusion

This thesis has gone to great depths to explore the intricate relationship between symmetries in data and their effect on the training process of overparametrized NNs. We have explored both the dynamics of NN training under the lens of the MF limit, and the use of symmetry-leveraging methods to profit from a problem's symmetries. Having established these theoretical grounds, we have been able to provide interesting and relevant contributions that improve the current understanding of symmetries in NN training.

We have presented an as-thorough-as-possible description of most of the key elements concerning the MF Limit of *overparametrized* Neural Networks. More importantly, we provide a *unified description* of the general setting that should allow newcomers to grab a hold of the topic with relative ease. We have described many commonly well-known theoretical properties from overparametrized neural networks, but reinterpreting them in our context and providing additional insights for the purpose of our work. Take, for instance, our extension of the well-known universality result (see Theorem 1, Proposition 3, Corollary 1) or our perspective on the usual *discriminatory assumption* in Proposition 1. Similarly, results such as Lemma 4, Proposition 9 and Proposition 10 (among many others from the literature) have been restated in our setting, with minimal modifications to the original proofs, to fit into the *general framework* of our work.

We have also presented an intense dive into some elements of group theory (and group actions) that allow for easily describing some of the main *symmetries* that are commonly encountered in real world problems. We have also complemented some of the literature's results on invariant/equivariant functions and measures; most remarkably, Proposition 13 is given an original proof an it is used to extend a well-known literature result (about the *symmetrization gap* of a learning problem) in Lemma 10. Similarly, we have been able to describe some of the main *symmetry-leveraging techniques* from the literature (in this setting), later translating their definition into the more general *Mean Field* framework.

In the bulk of our contributions, we managed to establish many interesting insights for the problem at hand, slowly building a theoretical background throughout Chapters 4 and 5 to reach the main results by the end of our work.

For instance, through an example, we introduce a novel way to construct *equivariant NNs* in the MF setting of *shallow* NNs (see Proposition 19). This, in turn, allows us to generalize the definition of such models (see Definition 4.2) and to grasp its properties from an abstract high-level perspective (see Proposition 20 and Proposition 21). This renewed definition from a broader perspective, led us to explore specific properties of the objects involved: invariant measures and equivariant functions.

Through a series of mostly original results (Lemma 12, Proposition 24 and specially Proposition 25, Lemma 13 and Proposition 26) we manage to grasp specific OT properties of *invariant measures* and *measures concentrated on the subspace of equivariant parameters.* We proceed analogously for the *derivatives and integrals of equivariant functions*; providing (mostly original) key properties in Proposition 29 and Proposition 30. On a parallel note, we draw from classic Ergodic Theory to reinterpret *invariant measures* simply as probability measures over the orbit space (with classic results such as Proposition 27 and Corollary 5 serving to establish Proposition 28 and Corollary 6).

On a similar note, *invariant functionals over the space of probability measures* are studied and heavily developed. For instance, an interesting variant of Jensen's inequality is proved with relatively standard arguments (Proposition 32) and it is then used (together with Proposition 31 and Corollary 8) to establish one of the main results, stating that invariant functionals can be minimized by only considering *symmetric* measures (see Proposition 33). An equivalent result is drawn for the orbit space in Proposition 34 (using Proposition 28); but for the case of *equivariant models*, Proposition 35 provides a significative (and novel) counterexample that forbids an analogous result.

We properly reinterpret *symmetrized models* in our context with Proposition 38; as well as the usual symmetry-leveraging techniques (**DA**, **FA** and **EA**) with Proposition 40. Further interesting properties of this setting are presented regarding *linear functional derivatives* (Proposition 41), optimization under **DA**, **FA** and **EA** (Proposition 42), and bounding the *approximate invariance* of a functional under *approximately symmetric data* (Proposition 44).

We culminate our work by examining the Wasserstein Gradient Flows (WGF) of *invariant* functionals. One of our main original results, Proposition 45, states that whenever an invariant functional R has a well defined WGF; then such WGF, when initialized on a *symmetric* measure, will remain *symmetric* overtime. Many subsequent results follow to exploit this fact in the specific setting of the MF Limit of overparametrized NNs (such as Theorem 14, Corollary 10, Corollary 11 and Corollary 9). On the same line, another of our main results states that WGFs initialized on a measure concentrated on the *subspace of equivariant parameters*, will remain concentrated on such a space overtime (see Theorem 15, Theorem 16 and Corollary 12). Finally, analyzing the WGF of the **DA** and **FA** versions of a given functional provides interesting insights about *when* they coincide (see Corollary 13, Corollary 14 and Corollary 15).

More generally, throughout our work, we have addressed many key questions regarding the impact of data symmetries on the MF limit of NN training. Each chapter has contributed to this objective, either building upon existing known frameworks or rigorously proving enlightening new results. All in all, we have provided a comprehensive analysis of symmetries in NNs that enriches both their theoretical understanding and their potential practical applications (such as the development of new NN architectures or new NN optimization algorithms). Despite all of this, our work remains far from complete, as many very interesting open questions are yet to be attacked in our future work.

Some of the main open questions from our work specially revolve around understanding under which conditions the use of *equivariant NNs* won't lead to a *loss of generalization power*. For instance, despite having the counterexample stating that $\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) < \inf_{\mu\in\mathcal{P}(\mathcal{E}^G)} R(\mu)$ (i.e. in general the optimal value of the *global problem* is *strictly smaller* than that of the *restricted problem*), one may want to *bound* the *distance* between both quantities under reasonable conditions (similar to what's done in Proposition 44). This could shed light onto the assumptions that need to be made for the *global optimal value* to be *sufficiently close* to the *restricted optimal value*; i.e. when is it reasonable to restrict ourselves to such a space?

Furthermore, the truly interesting questions appear at the level of the *relation between optimizers*. It would be interesting to understand under which conditions an optimizer of the *global problem* (say, μ_*) could satisfy that $(\mu_*)^{\mathcal{E}^G}$ is an optimizer for the *restricted* problem. Conversely, one might ask *how* an optimum of the *restricted problem* (say, ν_*) could be *canonically extended* (in some way) in order for it to become an *optimizer* of the *global problem*. In particular, even if we assume that *universality* holds for $\mathcal{E}^G$ (as in Proposition 36), does it help in establishing these relationships? In the regularized setting, how can we even compare both problems if the entropy terms force the minimizers to be concentrated on different spaces? (Also, can we leverage Proposition 10 in this case to find a relationship between them?).

Another one of the big challenges to be tackled in future work revolves around understanding the interplay between the Mean Field theory and the well-known *universality* results from the literature. Notably: to what extend is the *compactness* of $\mathcal{Z}$ required for Theorem 1 to hold? Are there specific conditions (notably, requiring an unbounded σ_* or a specific growth-rate for K) under which having *universality* might also imply that the infimum $\inf_{\gamma\in\mathcal{M}^S(\mathcal{Z})} R(\gamma)$ will be attained? Can this result be achieved at least on the case of bounded target $f^* = \mathbb{E}[Y|X = \cdot]$? One might even be willing to ask about the converse statement: if the *conditional expectation is attained* for every problem with arbitrary $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$, does this imply that the class $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is universal as well?

On the same line, having $\mathcal{E}^G$ satisfy *universality* (as in Proposition 36) is not at all trivial in our *MF shallow NN* setting. Remarkably, Maron et al. [55] have shown that *first order tensor NNs* (which are the ones we consider in Section 4.1) aren't always universal (beyond a handful of examples). Thus, a great open challenge that emerges, is understanding whether our *shallow NN model* can be used to represent more complex structures (notably, NNs with arbitrarily *large tensor order*) by considering a well-chosen activation function (σ_*). In other words: which other *universal* architectures can be represented using our *shallow NN model*?

The third major axis of our future research deals with the subject of: what exactly do we earn from *reducing the dimensionality of the problem* at hand? For instance, if we follow Proposition 34, we know that we can *reduce the global minimization problem to the orbit space* $G\backslash\mathcal{Z}$; now, how exactly does this make the problem *easier to solve*? In other words, is the resolution of the limiting DD easier to tackle?[1] Similarly, is there any specific quantitative

[1] We know from Chossat [18] that we could hope to achieve *some kind of simplification* of the systems of equations at hand.

advantage from only considering parameters in $\mathcal{E}^G$? Can we profit from the *dimensionality reduction* to improve some of the *bounds* of known results (e.g. of *global convergence* and *propagation of chaos*)? How do we manage to do this while circumventing the issue of the *degeneracy* of $\mathcal{E}^G$ as a strict subspace of $\mathcal{Z}$?

Along the same lines, it would be of crucial interest to truly understand the concrete advantages of employing **DA** and **FA** during NN training. These symmetry-leveraging techniques don't have major effects at the level of the WGF; however, when data is *symmetric*, one would expect that the use of these techniques could *accelerate training* in some way. One may then ask: can we achieve a quantitative advantage regarding the *convergence to the global minimum* and/or the *propagation of chaos* results by using these symmetry-leveraging techniques during training? Finally, how does the use of **EA** fit into this landscape? In particular, more thought shall be put into understanding how the dynamics of $R|_{\mathcal{P}(\mathcal{E}^G)}$ and R^{EA} might be tightly connected (as is the case for R and R^{FA} whenever R is G-invariant).

Beyond these three major axes, many other interesting questions have appeared throughout our study. Regarding other aspects of the MF theory of *shallow NNs*, one might be willing to ask whether any other initializations of the training dynamic are worth considering, beyond the classic i.i.d. initialization. Also, an adaptation of the CLT results to our *general* framework would be of crucial interest; as well as studying how symmetries take part in the limiting fluctuation process. Along the same lines, extending some of the known *global convergence* results to the SGD dynamics (as in Suzuki et al. [82]), or going deeper into the MF study of multilayer NNs (as in Nguyen and Pham [61]) are both relevant questions for the MF literature by themselves. Studying how symmetries could intervene in both such settings would undoubtedly be of the highest of interests.

Analogously, we believe that many of our results can still be extended/generalized. Notably, a deeper analysis of *our* version of Jensen's inequality (Proposition 32) should be considered: we believe that a *simpler* proof can be achieved by passing directly through the *Wasserstein subdifferential* of the relevant *convex functional* (without requiring it to be $\mathcal{C}^1$). This could also shed light onto *whatever happens when there's equality in Jensen's expression.* For instance, in the example of the *learning problem*: would the fact that $R(\mu) = R(\mu^G)$ force μ to be G-invariant? Such an insight could allow us to better grasp the structure of the minimizers of G-invariant functionals. On the same line, we would like to understand to what extent some of our results critically lie on the usage of a *quadratic loss*[2]: which of these could be extended to consider an arbitrary *convex loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$? On a similar note, we believe that attacking the topics of approximate and partial symmetries (and seeing how they can be expressed in the MF regime) would definitely be an interesting extension of our setting. The same holds for the research of novel symmetry-exploiting techniques from the literature (such as the *canonicalization* concept from Kaba et al. [42]): can we translate these to the MF setting in order to compare them to other *symmetry-leveraging techniques* at the MF level?

Other minor open questions have been proposed throughout this work, notably suggesting to establish that μ^G should be the *canonical way* of making $\mu \in \mathcal{P}(\mathcal{Z})$ *symmetric*[3] (i.e. its

[2] For instance, consider Lemma 1, Lemma 2, Lemma 10, among many others.

[3] We haven't managed to prove this yet, despite many intermediate results such as proposition 23, proposition 24 and proposition 25 bringing us closer to that objective.

projection onto $\mathcal{P}^G(\mathcal{Z})$). Along the same lines, one would like to find some way of *quantifying* the distance between μ and both μ^G and $\mu^{\mathcal{E}^G}$: e.g. could we bound $W_p(\mu, \mu^{\mathcal{E}^G})$? Finally, the study of symmetries in the MF regime of NNs, but while considering *non-trivial actions* on the intermediate layer (particularly, going beyond the ones of the form $\mathrm{Id}_N \otimes \eta$ with $G \circlearrowright_\eta \mathbb{R}^b$), could be a really interesting next step for establishing a robust theory regarding this topic.

Last but not least, the computational aspect of this work is yet to be fully fleshed out. Computing $\mathcal{E}^G$ isn't truly an easy task (as mentioned in Finzi et al. [31]), and so many of the *practical* aspects of our novel contributions are yet to be thoroughly explored. With this in mind, numerical simulations will be put in place in order to test and verify the bulk of our theoretical results from a *practical perspective.*

With all being said, we could consider all of these open questions and problems to be another significant part of the novel contributions produced by this work. Many of these interesting ideas either hadn't even been touched on the literature, or where simply *too complicated to write down in a concrete fashion.* The theoretical background provided by this work has allowed us to not only prove many interesting results ourselves, but also to properly describe (and write down) many remarkable questions and problems to be tackled in the months following this thesis.

This work has provided with significant contributions to both the literature regarding the MF Theory of *shallow NNs*, and the literature concerned with the leveraging of *symmetries* in *learning* problems. These include reinterpretations of known facts from the literature, novel results with interesting original proofs, and a big list of *open problems and questions* that have emerged during the development of this thesis. We hope that our work will serve as a catalyst for further research on this topic, ultimately leading to significant advancements in the area.

Bibliography

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, 2. ed edition, 2008. ISBN 978-3-7643-8722-8 978-3-7643-8721-1. OCLC: 254181287.

[2] Dyego Araújo, Roberto I. Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks, 2019.

[3] Jimmy Aronsson. Homogeneous vector bundles and g-equivariant convolutional neural networks. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2), jul 2022. doi: 10.1007/s43670-022-00029-3. URL `https://doi.org/10.1007%2Fs43670-022-00029-3`.

[4] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net, 2019.

[5] Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization, 2021.

[6] Andrew Barron. Barron, a.e.: Universal approximation bounds for superpositions of a sigmoidal function. ieee trans. on information theory 39, 930-945. *Information Theory, IEEE Transactions on*, 39:930 – 945, 06 1993. doi: 10.1109/18.256500.

[7] Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks, 2020.

[8] James V. Bondar. Borel cross-sections and maximal invariants. *The Annals of Statistics*, 4(5):866–877, 1976. ISSN 00905364. URL `http://www.jstor.org/stable/2958624`.

[9] Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for sgd in wide neural networks, 2020.

[10] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL `https://arxiv.org/abs/2104.13478`.

[11] Haoyang Cao and Xin Guo. Sde approximations of gans training and its long-run behavior, 2020. URL `https://arxiv.org/abs/2006.02047`.

[12] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games.* Probability Theory and Stochastic

Modelling. Springer International Publishing, 2018. ISBN 9783319589206. URL https://books.google.cl/books?id=fZFODwAAQBAJ.

[13] Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics, 2022. URL https://arxiv.org/abs/2212.03050.

[14] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation, 2020.

[15] Zhengdao Chen, Grant M. Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks, 2022.

[16] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport, 2018.

[17] Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing, 2022.

[18] Pascal Chossat. The reduction of equivariant dynamics to the orbit space for compact group actions. *Acta Applicandae Mathematicae*, 70:71–94, 01 2002. doi: 10.1023/A:1013970014204.

[19] Taco Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces, 2020.

[20] George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989. URL https://api.semanticscholar.org/CorpusID:3958369.

[21] Tri Dao, Albert Gu, Alexander J. Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation, 2019.

[22] Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks, 2018.

[23] Arden Dertat. Applied Deep Learning - Part 1: Artificial Neural Networks — towardsdatascience.com. https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6, 2017. [Accessed 26-01-2024].

[24] Arnaud Descours, Arnaud Guillin, Manon Michel, and Boris Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case, 2023.

[25] C. Druţu and M. Kapovich. *Geometric Group Theory*. Colloquium Publications. American Mathematical Society, 2018. ISBN 9781470411046. URL https://books.google.cl/books?id=9WXZnQAACAAJ.

[26] Edward G. Effros. Transformation groups and c*-algebras. *Annals of Mathematics*, 81 (1):38–55, 1965. ISSN 0003486X. URL http://www.jstor.org/stable/1970381.

[27] M. Einsiedler and T. Ward. *Ergodic Theory: with a view towards Number Theory.* Graduate Texts in Mathematics. Springer London, 2010. ISBN 9780857290212. URL https://books.google.cl/books?id=PiDET2fS7H4C.

[28] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models, 2021.

[29] Stewart N. Ethier and Thomas G. Kurtz. Markov processes: Characterization and convergence. 2005. URL https://api.semanticscholar.org/CorpusID:122218527.

[30] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data, 2020.

[31] Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups, 2021.

[32] Axel Flinth and Fredrik Ohlsson. Optimization dynamics of equivariant and augmented neural networks, 2023.

[33] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart, 2021.

[34] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[35] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes, 2019.

[36] Thomas Hofmann, Bernhard Schölkopf, and Alexander Smola. A tutorial review of rkhs methods in machine learning. 01 2006.

[37] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

[38] Kaitong Hu, Zhenjie Ren, David Siska, and Lukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks, 2020.

[39] Kevin H. Huang, Peter Orbanz, and Morgane Austern. Quantifying the effects of data augmentation, 2022.

[40] Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks, 2023.

[41] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020.

[42] Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions, 2023.

[43] O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer New York, 2002. ISBN 9780387953137. URL https://books.google.cl/books?id=L6fhXh13OyMC.

[44] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, Dordrecht, 2005. URL https://cds.cern.ch/record/1250328.

[45] Olav Kallenberg. *Random Measures, Theory and Applications*, volume 77. Springer, 01 2017. ISBN 978-3-319-41596-3. doi: 10.1007/978-3-319-41598-7.

[46] Jinwoo Kim, Tien Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance, 2023.

[47] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups, 2018.

[48] Wataru Kumagai and Akiyoshi Sannai. Universal approximation theorem for equivariant maps by group cnns. *ArXiv*, abs/2012.13882, 2020. URL https://api.semanticscholar.org/CorpusID:229679932.

[49] Leon Lang and Maurice Weiler. A wigner-eckart theorem for group equivariant convolution kernels, 2021.

[50] Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T. Kiani. Implicit bias of linear equivariant networks, 2022.

[51] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes, 2018.

[52] Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S. Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels, 2019.

[53] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks, 2020.

[54] James Lynch, Paul Dupuis, and Richard S. Ellis. A weak convergence approach to the theory of large deviations. 1997. URL https://api.semanticscholar.org/CorpusID:119566970.

[55] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks, 2019.

[56] Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80:309–323, 1995. URL https://api.semanticscholar.org/CorpusID:11979467.

[57] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

[58] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit, 2019.

[59] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models, 2021.

[60] Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows, 2021.

[61] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks, 2023.

[62] Tomohiro Nishiyama. Convex optimization on functionals of probability densities, 2020.

[63] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics, 2022.

[64] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes, 2020.

[65] Peter Orbanz and Daniel M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures, 2015.

[66] F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. ISSN 0022-1236. doi: https://doi.org/10.1006/jfan.1999.3557. URL https://www.sciencedirect.com/science/article/pii/S0022123699935577.

[67] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, 1991. doi: 10.1162/neco.1991.3.2.246.

[68] Fréderic Paulin. Introduction topologique à la géométrie, 2023. URL https://www.imo.universite-paris-saclay.fr/~frederic.paulin/notescours/cours_GeometrieM1Orsay.pdf.

[69] Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance, 2023.

[70] Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design, 2022.

[71] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing, 2017.

[72] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, jul 2022. doi: 10.1002/cpa.22074. URL https://doi.org/10.1002%2Fcpa.22074.

[73] Filippo Santambrogio. Optimal transport for applied mathematicians. calculus of variations, pdes and modeling. 2015. URL https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf.

[74] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

[75] Samuel S. Schoenholz. Fast and Easy Infinitely Wide Networks with Neural Tangents — blog.research.google. https://blog.research.google/2020/03/fast-and-easy-infinitely-wide-networks.html. [Accessed 26-01-2024].

[76] John Shawe-Taylor. Threshold network learning in the presence of equivalences. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/087408522c31eeb1f982bc0eaf81d35f-Paper.pdf.

[77] John Shawe-Taylor. Sample sizes for threshold networks with equivalences. *Information and Computation*, 118(1):65–72, 1995. ISSN 0890-5401. doi: https://doi.org/10.1006/inco.1995.1052. URL https://www.sciencedirect.com/science/article/pii/S0890540185710528.

[78] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers, 2018. URL https://arxiv.org/abs/1805.01053.

[79] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks, 2019. URL https://arxiv.org/abs/1903.04440.

[80] Justin Sirignano and Konstantinos Spiliopoulos. Asymptotics of reinforcement learning with neural networks, 2019. URL https://arxiv.org/abs/1911.07304.

[81] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem, 2019.

[82] Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction, 2023.

[83] Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg. ISBN 978-3-540-46319-1.

[84] M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28 – 76, 1994. doi: 10.1214/aop/1176988847. URL https://doi.org/10.1214/aop/1176988847.

[85] Carson Teitler, Dec 2020. URL http://math.columbia.edu/~mmiller/TProjects/CTeitler20s.pdf.

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[87] C. Villani. Topics in optimal transportation theory. 58, 01 2003. doi: 10.1090/gsm/058.

[88] Cédric Villani. Optimal transport: Old and new. 2008. URL https://api.semanticscholar.org/CorpusID:118347220.

[89] Soledad Villar, David W. Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics, 2023.

[90] Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns, 2021.

[91] Jeffrey Wood and John Shawe-Taylor. Representation theory and invariant neural networks. *Discrete Applied Mathematics*, 69(1):33–60, 1996. ISSN 0166-218X. doi: https://doi.org/10.1016/0166-218X(95)00075-3. URL https://www.sciencedirect.com/science/article/pii/0166218X95000753.

[92] Dmitry Yarotsky. Universal approximations of invariant maps by neural networks, 2018.

[93] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets, 2018.

[94] Berfin Şimşek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances, 2021.

Annexes

Annex A

Reading Guide and Summary of Contributions

Concerning the *specifics* of this work, the following chapter gives a detailed overview of the structure of the document, delving into the specific results presented in each section.

Chapter 1

Chapter 1 contains a global overview of the problem at hand, together with an introduction this work's objectives and overall structure.

Chapter 2

Chapter 2 presents a thorough review of the literature on the topic of the **Mean Field (MF) Limit** in the context **Shallow Neural Networks (NNs)**, as well as most relevant theoretical elements for defining such an object. More specifically, its different subsections can be described as follows:

- In Section 2.1 the general learning problem is introduced, and in Section 2.1.1 the same is done for a standard model of Multilayer NNs.
- Section 2.2 introduces the idea of the (multiple) overparametrized regimes for the training of NNs.
- The chapter follows by going deeper into the results related to the MF limit:
  - First, considering known universality results from the literature (which we **slightly adapt** to make them fit into the context that's useful for our subsequent results).

    A proposition is proved showing that the *usual* notion of *discriminatory activation function* (from the ML literature) *implies* the one established in assumption 1 (proposition 1).

A classic universality result (Theorem 1 and corollary 1) is presented, with a proof by Rotskoff and Vanden-Eijnden [72] replicated under more *general* assumptions. Similarly, an heuristic observation from Rotskoff and Vanden-Eijnden [72] is now presented in our context with an added layer of formality (Proposition 3; with a proof that formalizes from the ideas of [72]).

- Section 2.2.3 presents the ideas behind the usual NN optimization problem, and how it might be **convexified** to simplify its resolution. Also, some known results about the decomposition of the population risk (in the quadratic case) and what this implies in the setting of universality (Lemma 1 and Lemma 2), are presented (with minimal proofs for *clarity*).
- The basic theory about Wasserstein spaces is presented (Section 2.2.4). Two small basic results are presented (and proved) to illustrate properties of the *population risk* in our setting (Lemma 3 and Proposition 5).
- The theory of Wasserstein Gradient Flows is introduced in Section 2.2.5, largely based in Ambrosio et al. [1], Santambrogio [73] and Villani [88] .

  Linear Functional Derivatives and *Intrinsic Derivatives* are generally defined (and calculated for some remarkable examples); and the general result of Proposition 6 is proved (with an original proof, though it's likely that similar results have been previously established in the literature).

  A known result about $\mathcal{C}^1$ functionals is presented (Lemma 4; with its proof being replicated from Hu et al. [38] for clarity).

  The usual definition of a Wasserstein (sub-)Gradient Flow (WGF) is provided and a well known result about the existence and uniqueness of solutions is presented (Proposition 7).
- Multiple variants of Stochastic Gradient Descent (SGD) for neural network training are presented in Section 2.2.6. Some key notation for our work (inspired from Section 2.2.5) is introduced.
- Section 2.2.7 regroups all the elements introduced up until that point to present the *MF view* of neural network training. In particular, a well known result of *Propagation of Chaos* is presented (under multiple variants detected in the existing literature) in Theorem 4. This shows exactly *in what sense* SGD can be regarded as a WGF for the population risk functional.
- Section 2.2.8 finalizes our exploration of the MF limit of shallow NNs by presenting classical *global convergence* results from the literature (for instance, Theorem 5, Proposition 9, Proposition 11, Proposition 10, Theorem 6, Theorem 7, Theorem 8, Corollary 3). In particular, part of our work included placing most of these results (coming from a multitude of papers) under a common *setting* (and notation). In particular some proofs (such as that of Proposition 10) were replicated from their original works (e.g. Hu et al. [38]), but under the new notation (and overall setting), to ensure that they stayed true even under the more *general* description.
- Finally, the obtained results are put under the lens of *shallow NN models*: In Assumption 5, we provide enough conditions for some of the results of the previous sections to hold.

- A quick overview of the existing work on the *mean field limit of deep neural networks* is provided, though not much depth is sought.

Chapter 3

In a similar vein to **Chapter 2**, **Chapter 3** presents the ideas behind the *group theoretical* understanding of *symmetries* in the context of Neural Networks; displaying relevant theoretical results from the literature as well as introducing *relevant ideas* from a practical viewpoint. Its subsections are structured as follows:

- The (well known) idea of *symmetry* (as *invariance/equivariance*) is introduced and motivated.
- Some background on group theory is presented, as it is the usual way of studying *symmetries* in the NN context. In particular, group actions are defined and some of their properties (as well as those of *orbit spaces*) are presented (Proposition 12). Invariant/Equivariant Maps and Actions via Representations are also introduced.
- Some well known results coming from the theory of G-invariant measures (see Kallenberg [45]); such as Theorem 10, Theorem 11 and Theorem 12 are presented. Some derivative results (closer to the setting of the learning problem) are also introduced.

  In particular, Proposition 13 (esentially stating that the conditional expectation function is G-equivariant when one assumes the underlying *law* to satisfy the same property) is presented and **proved**. The proof relies mostly on standard arguments, but we however believe it to be original (as far as our knowledge reaches).
- Some of the *standard* theory used to understand *symmetric models* (particularly, from Elesedy and Zaidi [28]) is introduced. The *symmetrization operator* is defined and a Hilbert orthogonal decomposition lemma is presented (Lemma 8).

  The notion of *symmetrization gap* (standard in the literature) is presented, and a known result from Elesedy and Zaidi [28] is **extended** (using Proposition 13 and following ideas from Huang et al. [40]). This *new version* is presented (and **proved**) in Lemma 10
- Section 3.4 presents the different well known and mostly used *techniques for leveraging symmetries* in the NN context. In particular, Data Augmentation (**DA**), Feature Averaging (**FA**) and Equivariant Architectures (**EA**) are each presented in a different section, together with some well known results concerning their *theoretical properties.* Special care is put into the understanding of EA, and Flinth and Ohlsson [32] is widely used as a reference in this aspect.

Chapter 4

Chapter 4 starts presenting the results from our own *study* of *symmetries* in the NN context, with an Optimal transport (OT) and Mean Field view. In particular, the chapter is built on the following structure:

- In Section 4.1.2, a simple, yet really useful, example is fully fleshed out, to introduce the concept of G-equivariant NNs (how they should be described in the *shallow NN setting*; and why a more *complex* network structure is required to make G-actions *interesting*). A useful characterization of G-equivariant *shallow NNs* is provided (Proposition 19), and this serves to motivate the idea of **generalizing** the definition of G-equivariant NNs.

  A new (broader) definition of *shallow* Equivariant NN models is introduced (consider Definition 4.2) and some of its basic properties are **proven** (e.g. in Proposition 20 and Proposition 21). Furthermore, the previous *motivational* example is shown to satisfy the new (more general) definition (in Proposition 22).

- Some more concepts about G-invariant measures are introduced; but now under the lens of OT theory. The well known *symmetrization* of a measure is presented and some of its basic properties are **proven** (see Proposition 23). A similar approach is used for the *pushforward through an orthogonal projection onto a linear subspace* (Lemma 12) .

  Some (relatively) well know properties of the Wasserstein distance are **proven** (notably, Proposition 24); though the arguments used are relatively standard and even directly inspired from Santambrogio [73].

  Similarly, in Proposition 25 and Lemma 13, interesting properties of the measure spaces $\mathcal{P}^G(\mathcal{Z})$ and $\mathcal{P}(\mathcal{E}^G)$ are presented and **proven** (and, to our knowledge, they are mostly original). Most remarkably, we prove that, in some sense we can *canonically project* any measure onto any one of these spaces.

  Finally, Proposition 26 provides a characterization of *how the densities of the different projections of a given measure* look like (and we **prove** this properties as well, through relatively standard techniques).

- Section 4.2.1 provides a reinterpretation of the results of Section 3.2.1 by passing to the *orbit space* $G\backslash\mathcal{Z}$. Notably, Proposition 27 and Corollary 5 are **proved** and used to establish a correspondence between $\mathcal{P}^G(\mathcal{Z})$ and $\mathcal{P}(G\backslash\mathcal{Z})$ (Proposition 28 and Corollary 6). These kinds of results seem to already exist in the setting of *ergodic theory*, however, the proofs we provide are original (as far as we were able to identify in the related literature).

- Section 4.3 Provides key properties satisfied by G-invariant functionals under derivation and integration. In particular, Proposition 29 and Proposition 30 are introduced and **proven** (through relatively standard arguments).

- Finally, Section 4.4 provides insight into the properties of functionals $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, particularly in the case when they are assumed to be G-invariant.

  More specifically, a notion of G-invariance is introduced for these kinds of functionals (Definition 4.4) and many properties are **proven** under such definition. For this purpose, an interesting variant of Jensen's inequality is proved (Proposition 32; the proof is rather standard but, to our knowledge, it is original). Most remarkably Proposition 31, Corollary 8 and Proposition 33 are proved, ultimately showing that G-invariant functionals can be minimized by only considering *symmetric* measures on $\mathcal{P}^G(\mathcal{Z})$. Furthermore, some interesting examples of G-invariant functionals are provided. Analogously, Proposition 28 is used to further *narrow down* the problem of *optimizing a G-invariant functional* to that of finding an *arbitrary* measure on $\mathcal{P}(G\backslash\mathcal{Z})$

(Proposition 34). Examples of such a *reduction* at the level of G-invariant functionals are also provided.

Finally, Proposition 35 provides an original **counterexample** to the idea that a G-invariant functional can always be minimized over $\mathcal{E}^G$. i.e. We prove that in the general setting models based on $\mathcal{P}(\mathcal{E}^G)$ *aren't enough* to optimize the original population risk functional.

Chapter 5

Chapter 5 culminates our work by employing the discovered facts from all previous chapters to prove properties of the WGF of G-invariant functionals.

- First, some insight is provided about *what a symmetrized NN model would look like in our context* (notably, Proposition 38 is **proven** and complemented with an example).
- Proposition 39, (stating that under the right G-invariance conditions the population risk from the *learning problem* will *also* be G-invariant) is derived from results of the previous chapter.
- Section 5.2.1 brings back the ideas of **DA**, **FA** and **EA** and places them in the context of our work. In particular, Proposition 40 (describing the corresponding functionals R^G, R^{FA} and R^{EA}), Proposition 41(describing the linear functional derivatives of such functionals) and Proposition 42 (showing that the optimization under **DA** and **FA** coincide and correspond to optimizing the original functional over the space $\mathcal{P}^G(\mathcal{Z})$) are presented and **proven**. Finally, a rather interesting bound (useful in the context of *approximate G-invariance*) is provided in the *learning setting* (Proposition 44). This fact is heavily inspired from an analogous result by Chen et al. [14], and its proof largely follows that of [14] (with small adaptations to fit it in our context).
- Finally, Section 5.3 provides some of the main results that pull together all the theoretical machinery introduced throughout this work. Most remarkably, we **prove** Proposition 45, which states that, whenever a G-invariant functional R has a well defined WGF; then such WGF will always **respect** the symmetric properties of the input (i.e. if the dynamic *starts* at a G-invariant measure, then overtime the entire flow will remain G-invariant as well). This result is put into perspective with Theorem 14, Corollary 10, Corollary 11 and Corollary 9 (the latter stating that: *if* the WGF finds an optimum through a G-invariant dynamic, then such an optimum has to be G-invariant as well).

  These results can actually be made **stronger**, as we can state that a G-invariant functional will be such that its WGF, if initialized in $\mathcal{P}(\mathcal{E}^G)$, will remain in $\mathcal{P}(\mathcal{E}^G)$ overtime (see Theorem 15, Theorem 16) and Corollary 12.

  Last of all, we analyze how the WGF of the **DA** and **FA** versions of a functional R (R^G and R^{FA} resp.) behave in terms of their associated WGF. We prove Corollary 14 (or, in particular, Corollary 13), which states that: when initialized in $\mathcal{P}^G(\mathcal{Z})$, the WGF of R^{FA} and R^G coincide exactly. Similarly, Corollary 15 states that the properties of WGFs for G-invariant functionals are all satisfied by the **DA**, **FA** and **EA** versions of R (even if R isn't G-invariant).

Chapter 6

Chapter 6 provides a natural conclusion to our work, gathering and summarizing the bulk of our original contributions. It also contains a compilation of open questions to be attacked in our future work.

Annexes

Annex A corresponds to the current review of the work's structure.

Annex B provides an illustrative example of a Deep Set architecture. Also, some illustrative *calculations* are provided for a simple example of NN architecture in our setting.

Annex C contains all the Proofs for the different results presented in the work (they are original for the most part, and whenever elements were borrowed from anywhere in the literature, it is clearly stated as so).

Annex D contains some of the classic *technical assumptions* that are required for the results to hold (specially those of Section 2.2). It also contains some information about the *Central Limit Theorem* of the SGD training process of shallow NNs; and some results of Existence/Uniqueness of solutions for the McKean-Vlasov equation in the setting of Bortoli et al. [9].

Annex B

Application of our framework to some EA

B.1 Example of *Deep Sets*

An emblematic example of **neural networks with equivariant architecture** are the **Deep Sets**, introduced by Zaheer et al. [93]. Essentially, it is a neural network architecture designed to be **invariant** under the action of $G = \mathcal{S}_n$.

The input space is essentially $\mathbb{R}^{n\times d} \simeq (\mathbb{R}^n)^{\otimes d}$, which represents having n copies of length-d real-valued vectors. We aim to construct a network with 1 hidden layer of the form $\hat{y} : \mathbb{R}^{n\times d} \xrightarrow{\varphi,\sigma} \mathbb{R}^{n\times N} \xrightarrow{\mathcal{A}} \mathbb{R}^N \xrightarrow{W} \mathbb{R}$, where $\hat{y}(x) = \frac{1}{N} W\mathcal{A}(\sigma(\varphi(x)))$, and $\varphi \in \tilde{\mathbf{Hom}}_G(\mathbb{R}^{n\times d}, \mathbb{R}^{n\times N})$, $W \in \mathbb{R}^N$. The layer $\mathcal{A}$ corresponds to a **global average pooling**, which *reduces the dimension* of the layer's output by simply *averaging* over the domain on which the group acts (with no trainable parameters). Eventually, we would like to let N go to infinity.

As shown in [93], the **only** way to achieve a $\mathcal{S}_n$**-equivariant** layer is if the *matrices* $A \in \mathbb{R}^{(n\times N)\times(n\times d)}$ and $b \in \mathbb{R}^{n\times N}$ (from the definition of equivariant affine layer: $\varphi : x \to Ax+b$) are of the form:

$$A = \alpha \otimes I + \beta \otimes J, \qquad b = \gamma \otimes (1, \dots, 1)$$

Where $\alpha, \beta \in \mathbb{R}^{N\times d}, \gamma \in \mathbb{R}^N$ are the **trainable parameters** of the layer; $I = \mathbf{Id}_{n\times n}$ and $J = \vec{1}_n\vec{1}_n^T$ are two $n \times n$ matrices; and $\otimes$ is the usual tensor product.

Explicitly expressing these matrices in a *block-wise* representation, we have:

$$\alpha \otimes I = \begin{pmatrix} \alpha & 0 & \dots & 0 \\ 0 & \alpha & \dots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \alpha \end{pmatrix} \in \mathbb{R}^{(n\times N)\times(n\times d)}, \qquad \beta \otimes J = \begin{pmatrix} \beta & \beta & \dots & \beta \\ \beta & \beta & \dots & \beta \\ \vdots & \ddots & \ddots & \beta \\ \beta & \beta & \dots & \beta \end{pmatrix} \in \mathbb{R}^{(n\times N)\times(n\times d)}$$

$$\gamma \otimes (1, \dots, 1) = (\gamma, \dots, \gamma) \in \mathbb{R}^{(n\times N)}$$

So that:

$$\forall i \in \{1, \ldots, n\},\ \varphi(x)(i) = \alpha \cdot x(i) + \beta \cdot \left(\sum_{j=1}^{n} x(j)\right) + \gamma \in \mathbb{R}^N$$

In other words, for this equivariant layer, we have only $2(N \times d) + N = N(2d + 1)$ free parameters. We can further notice that:

$$\alpha = \left(\begin{array}{c} \alpha_1 \\ \hline \vdots \\ \hline \alpha_c \end{array}\right), \qquad \beta = \left(\begin{array}{c} \beta_1 \\ \hline \vdots \\ \hline \beta_c \end{array}\right), \qquad \gamma = (\gamma_k)_{k=1}^{N}$$

With $\alpha_k, \beta_k \in \mathbb{R}^d$ for all $k \in \{1, \ldots, N\}$. Thus, we can write the action of φ as:

$$\forall k \in \{1, \ldots, N\},\ \forall i \in \{1, \ldots, n\},\ (\varphi(x))_k(i) = \alpha_k \cdot x(i) + \beta_k \cdot \left(\sum_{j=1}^{n} x(j)\right) + \gamma_k$$

Therefore, our **1-layer $\mathcal{S}_n$-equivariant neural network** takes the form[1]:

$$\hat{y}(x) = \frac{1}{N} \sum_{k=1}^{N} W_k \cdot \frac{1}{n} \sum_{i=1}^{n} \sigma \left(\alpha_k \cdot x(i) + \beta_k \cdot \left(\sum_{j=1}^{n} x(j)\right) + \gamma_k \right)$$

In other words, we can group our parameter vector simply as: $\theta = (\theta_1, \ldots, \theta_N)$, where:

$$\forall i \in \{1, \ldots, N\},\ \theta_i = (W_i, \alpha_i, \beta_i, \gamma_i) \in \mathcal{E}^G := \mathbb{R}^{2d+2}$$

(notice that these are the parameters *invariant* to the corresponding action of $\mathcal{S}_n$ over the *ambient space* $\mathcal{Z} := \mathbb{R}^{n \times (n \times d)} \times \mathbb{R}^n$). This brings us closer to what we would like to study in the Mean Field limit. Moreover, if we define:

$$\sigma_*(x, \theta_k) := \sum_{i=1}^{n} W_k \sigma \left(\alpha_k \cdot x(i) + \beta_k \cdot \left(\sum_{j=1}^{n} x(j)\right) + \gamma_k \right)$$

we directly obtain:

$$\hat{y}_N(x, \theta) = \frac{1}{N} \sum_{k=1}^{N} \sigma_*(x, \theta_k)$$

which is essentially the same formulation used in our work.

It's not complex to notice that, in this network architecture, what we can consider as the *fundamental interchangeable unit* is somewhat more intricate than the *usual shallow NN model* described at the beginning of Section 2.2. Particularly, in this case, the intermediate layer takes the form $\mathbb{R}^{n \times N}$, with a **non-trivial action $\mathcal{S}_n \circlearrowleft_\eta \mathbb{R}^{n \times N}$** (so, it rather falls under our setting from section 4.1.2. A fundamental question for our future work will be to attempt to extend these concepts to more general actions of G (and therefore other kinds of architectures).

[1] Where, in this case, the **GAP** function $\mathcal{A}$ is defined as: $\mathcal{A}(y) = \frac{1}{n} \sum_{i=1}^{n} y(i)$

B.2 Some Explicit Calculations for the typical example

Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Z} = \mathbb{R}^D = \mathbb{R}^{(c+d)b}$, $\mathcal{Y} = \mathbb{R}^c$ and let G be a compact group such that $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_\eta \mathbb{R}^b$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$. Consider the *activation function* of *shallow NNs* $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ given (for $\theta = (w, a)$) by $\sigma_*(x, \theta) = w\sigma(a^T x)$, with $\sigma : \mathbb{R}^b \to \mathbb{R}^b$ being G-equivariant to the action of η (e.g. in many cases it is enough for σ to be applied *pointwise*).

Consider the quantity we defined as $\forall \mu \in \mathcal{P}(\mathcal{Z})$: $L_{x,y}(\mu) := \ell(\langle \sigma_*(x; \cdot), \mu \rangle, y)$. In this particular example, we can *explicitly calculate* relevant quantities such as $D_\mu L_{x,y}(\mu, \theta)$ for all $\theta \in \mathbb{R}^D$, $\forall \mu \in \mathcal{P}(\mathcal{Z})$, $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$. We find that the expression for this is given by:

$$J(x, y; \theta, \mu) := D_\mu L_{x,y}(\mu, \theta) = \begin{pmatrix} \nabla_1 \ell(\langle \sigma_*(x, \cdot), \mu \rangle, y) \sigma(a^T x)^T \\ x \nabla_1 \ell(\langle \sigma_*(x, \cdot), \mu \rangle, y)^T w \sigma'(a^T x) \end{pmatrix} \in \mathbb{R}^{(c+d) \times b}$$

Which can be seen, vectorized as $\vec{J}(x, y; \theta, \mu) = \mathrm{vec}(J(x, y; \theta, \mu)) \in \mathbb{R}^D$. We can derive this expression as follows:

Derivation of the expression of $J(x, y; \theta, \mu) := D_\mu L_{x,y}(\mu, \theta)$. The expression for $\vec{J}$ comes from the following derivation. Recall that:

$$\vec{J}(x, y; \theta, \mu) := (D_{\vec{\theta}} \sigma_*(x, \theta))^T U(x, y; \mu) \in \mathbb{R}^D$$

where we (temporarily) define $U(x, y; \mu) := \nabla_1 \ell(\langle \sigma_*(x, \cdot), \mu \rangle, y)$

We notice that

$$D_{\vec{\theta}} \sigma_*(x, \theta) = D_{\vec{\theta^T}} \sigma_*(x, \theta) \cdot D_{\vec{\theta}}(\mathrm{vec}(\theta^T))$$

And, as $\mathrm{vec}(\theta^T) = K^{(c+d,b)} \mathrm{vec}(\theta)$, we have: $D_{\vec{\theta}}(\mathrm{vec}(\theta^T)) = K^{(c+d,b)}$, so that:

$$D_{\vec{\theta}} \sigma_*(x, \theta) = D_{\vec{\theta^T}} \sigma_*(x, \theta) \cdot K^{(c+d,b)}$$

The idea of derivating with respect to $\vec{\theta^T}$ is that $\theta^T = (w^T, a^T)$ and so, $\mathrm{vec}(\theta^T) = \begin{pmatrix} \mathrm{vec}(w^T) \\ \mathrm{vec}(a^T) \end{pmatrix}$ the derivative has the components associated with each set of parameters separate; i.e.

$$D_{\vec{\theta^T}} \sigma_*(x, \theta) = \left(D_{\mathrm{vec}(w^T)} \sigma_*(x, \theta), D_{\mathrm{vec}(a^T)} \sigma_*(x, \theta) \right)$$

Now, notice that, as $\sigma_*(x, \theta) \in \mathbb{R}^c$:

$$\begin{aligned} \sigma_*(x, \theta) &= w\sigma(a^T x) = \mathrm{vec}(w\sigma(a^T x)) = \mathrm{vec}(\mathrm{Id}_c \cdot w \cdot \sigma(a^T x)) \\ &= (\sigma(a^T x)^T \otimes \mathrm{Id}_c) \mathrm{vec}(w) = (\sigma(a^T x)^T \otimes \mathrm{Id}_c) K^{(b,c)} \mathrm{vec}(w^T) \end{aligned}$$

Where we've used properties of the Kronecker product, and those of vectorization. With this, we get that:

$$D_{\mathrm{vec}(w^T)} \sigma_*(x, \theta) = (\sigma(a^T x)^T \otimes \mathrm{Id}_c) K^{(b,c)}$$

Similarly, by the chain rule we get:

$$D_{\mathrm{vec}(a^T)} \sigma_*(x, \theta) = w\sigma'(a^T x) D_{\mathrm{vec}(a^T)}(a^T x)$$

where we're denoting, for $y \in \mathbb{R}^b$:

$$\sigma'(y) := \begin{pmatrix} \sigma'(y_1) & 0 & \dots & 0 \\ 0 & \sigma'(y_2) & \dots & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \sigma'(y_b) \end{pmatrix}$$

This comes from the fact that the activation function σ is applied coordinatewise.

On the other hand, as $a^Tx \in \mathbb{R}^b$, $a^Tx = \mathrm{vec}(a^Tx) = \mathrm{vec}(\mathrm{Id}_b a^T x) = (x^T \otimes \mathrm{Id}_b)\mathrm{vec}(a^T)$. With this, we get:

$$D_{\mathrm{vec}(a^T)}\sigma_*(x,\theta) = w\sigma'(a^Tx)(x^T \otimes \mathrm{Id}_b)$$

Putting both together, we have:

$$D_{\vec{\theta}^T}\sigma_*(x,\theta) = \big((\sigma(a^Tx)^T \otimes \mathrm{Id}_c)K^{(b,c)}, w\sigma'(a^Tx)(x^T \otimes \mathrm{Id}_b)\big) \in \mathbb{R}^{c\times D}$$

This translates to:

$$D_{\vec{\theta}}\sigma_*(x,\theta) = \big((\sigma(a^Tx)^T \otimes \mathrm{Id}_c)K^{(b,c)}, w\sigma'(a^Tx)(x^T \otimes \mathrm{Id}_b)\big) \cdot K^{(c+d,b)}$$

So that, for $\vec{J}$

$$\begin{aligned} \vec{J}(x,y;\theta,\mu) &= (\big((\sigma(a^Tx)^T \otimes \mathrm{Id}_c)K^{(b,c)}, w\sigma'(a^Tx)(x^T \otimes \mathrm{Id}_b)\big) K^{(c+d,b)})^T U(x,y;\mu) \\ &= K^{(b,c+d)} \cdot \big((\sigma(a^Tx)^T \otimes \mathrm{Id}_c)K^{(b,c)}, w\sigma'(a^Tx)(x^T \otimes \mathrm{Id}_b)\big)^T U(x,y;\mu) \\ &= K^{(b,c+d)} \cdot \begin{pmatrix} K^{(c,b)}(\sigma(a^Tx) \otimes \mathrm{Id}_c) \\ (x \otimes \mathrm{Id}_b)\sigma'(a^Tx)w^T \end{pmatrix} U(x,y;\mu) \\ &= K^{(b,c+d)} \cdot \begin{pmatrix} K^{(c,b)}(\sigma(a^Tx) \otimes \mathrm{Id}_c)U(x,y;\mu) \\ (x \otimes \mathrm{Id}_b)\sigma'(a^Tx)w^T U(x,y;\mu) \end{pmatrix} \in \mathbb{R}^D \end{aligned}$$

On the other hand, we can see that the matrix $\bar{J}(x,y;\theta,\mu) = \begin{pmatrix} U(x,y;\mu)\sigma(a^Tx)^T \\ xU(x,y;\mu)^T w\sigma'(a^Tx) \end{pmatrix}$ satisfies:

$$\begin{aligned} \mathrm{vec}\left(\bar{J}(x,y;\theta,\mu)\right) &= \mathrm{vec}\left((\sigma(a^Tx)U(x,y;\mu)^T, \sigma'(a^Tx)w^T U(x,y;\mu)x^T)^T\right) \\ &= K^{(b,c+d)} \cdot \mathrm{vec}\left((\sigma(a^Tx)U(x,y;\mu)^T, \sigma'(a^Tx)w^T U(x,y;\mu)x^T)\right) \\ &= K^{(b,c+d)} \cdot \begin{pmatrix} \mathrm{vec}(\sigma(a^Tx)U(x,y;\mu)^T) \\ \mathrm{vec}(\sigma'(a^Tx)w^T U(x,y;\mu)x^T) \end{pmatrix} \\ &= K^{(b,c+d)} \cdot \begin{pmatrix} K^{(c,b)}\mathrm{vec}(U(x,y;\mu)\sigma(a^Tx)^T) \\ \mathrm{vec}(\sigma'(a^Tx)w^T U(x,y;\mu)x^T) \end{pmatrix} \end{aligned}$$

And, as:

$$\begin{aligned} \mathrm{vec}(U(x,y;\mu)\sigma(a^Tx)^T) &= \mathrm{vec}(\mathrm{Id}_c U(x,y;\mu)\sigma(a^Tx)^T) \\ &= (\sigma(a^Tx) \otimes \mathrm{Id}_c)\mathrm{vec}(U(x,y;\mu)) \\ &= (\sigma(a^Tx) \otimes \mathrm{Id}_c)U(x,y;\mu) \end{aligned}$$

and:

$$\begin{aligned}\mathrm{vec}(\sigma'(a^Tx)w^TU(x,y;\mu)x^T) &= \mathrm{vec}(\mathrm{Id}_b\sigma'(a^Tx)w^TU(x,y;\mu)x^T)\\ &= (x\otimes \mathrm{Id}_b)\mathrm{vec}\left(\sigma'(a^Tx)w^TU(x,y;\mu)\right)\\ &= (x\otimes \mathrm{Id}_b)\sigma'(a^Tx)w^TU(x,y;\mu)\end{aligned}$$

We get that:

$$\mathrm{vec}\left(\bar{J}(x,y;\theta,\mu)\right) = K^{(b,c+d)}\cdot\begin{pmatrix}K^{(c,b)}\mathrm{vec}(U(x,y;\mu)\sigma(a^Tx)^T)\\ \mathrm{vec}(\sigma'(a^Tx)w^TU(x,y;\mu)x^T)\end{pmatrix} = \vec{J}(x,y;\theta,\mu)$$

In short, the following expression holds:

$$\vec{J}(x,y;\theta,\mu) = \mathrm{vec}\left(\bar{J}(x,y;\theta,\mu)\right)$$

Which in particular, due to $\mathrm{vec}:\mathbb{R}^{(c+d)\times b}\to\mathbb{R}^D$ being an isomorphism, implies that:

$$J(x,y;\theta,\mu) = (D_\theta\sigma_*(x,\theta))^TU(x,y;\mu)\in\mathbb{R}^{(c+d)\times b}$$

can be perfectly identified with $\bar{J}(x,y;\theta,\mu)$. Indeed, if V is the matrix representing vec, we get:

$$D_\theta\sigma_*(x,\theta) = D_{\vec{\theta}}\sigma_*(x,\theta)D_\theta\mathrm{vec}(\theta) = D_{\vec{\theta}}\sigma_*(x,\theta)\cdot V$$

So that:

$$J(x,y;\theta,\mu) = V^T\vec{J}(x,y;\theta,\mu) = V^TV\bar{J}(x,y;\theta,\mu)$$

And as V is orthogonal[2], we get the expression we wanted:

$$J(x,y;\theta,\mu) = \bar{J}(x,y;\theta,\mu) = \begin{pmatrix}U(x,y;\mu)\sigma(a^Tx)^T\\ xU(x,y;\mu)^Tw\sigma'(a^Tx)\end{pmatrix}\in\mathbb{R}^{(c+d)\times b}$$

□

Hopefully, these calculations will not only serve to understand the different objects at play in our *slightly more robust* version of *shallow NNs*; but also to highlight the immense *burden of calculations* relieved by considering our theory from a more *general viewpoint.*

[2] This comes from the fact that the action over the basis of $\mathbb{R}^{(c+d)\times b}$ gives $\mathrm{vec}(e_{i,j}) = e_{b(j-1)+i}$ in the basis of $\mathbb{R}^D$. As for all $i_1,i_2\in\{1,\dots,(c+d)\},\ j_1,j_2\in\{1,\dots,b\},\ (i_1,j_1)\neq(i_2,j_2)\implies b(j_1-1)+i_1\neq b(j_2-1)+i_2$, we get the orthogonality condition we desire.

Annex C

Proofs

C.1 Proofs for Section 2.2.2

For the original *universality* result in Rotskoff and Vanden-Eijnden [72], the authors consider:

Assumption 6 *[Assumptions for the Universality result in Rotskoff and Vanden-Eijnden [72]] Let $\mu \in \mathcal{P}(\mathcal{X})$, and consider:*

- $\mathcal{X} \subseteq \mathbb{R}^d$ *and* $\mathcal{Z} \subseteq \mathbb{R}^D$ *are* ***closed smooth Riemannian manifolds*** *(in particular, compact).*
- $\mu - a.s.\ \forall x \in \mathcal{X},\ \sigma_*(x, \cdot) \in \mathcal{C}^1(\mathcal{Z})$*.*
- *Let σ_* be discriminating, in the sense that:*

$$\left[\forall z \in \mathcal{Z}\ \ a.e. \quad \int_{\mathcal{X}} g(x)\sigma_*(x, z)\mu(dx) = 0\right] \Longrightarrow [g = 0\ \ \mu - a.e.\ \ in\ \ \mathcal{X}]$$

And with this, they prove:

Theorem 17 (**Universality** as in Rotskoff and Vanden-Eijnden [72]) *Under assumption 6, $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is a* ***dense subspace*** *of $L^2(\mathcal{X}, \mathcal{Y}; \mu)$ (in the L^2-norm topology).*

In the body of this work, we consider a *slightly stronger* universality result, in the sense that it might be used in a setting with $\mathcal{Y}$ being a subset of vector space more interesting than $\mathbb{R}$.

Proof of Proposition 1. Let $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}^c$ and $\mathcal{Z} = \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} \times \mathbb{R}^b$, and $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be defined as:

$$\forall x \in \mathcal{X},\ \forall \theta = (w, a, b) \in \mathcal{Z},\ \sigma_*(x; \theta) := w\sigma(a^T x + b)$$

for $\sigma : \mathbb{R} \to \mathbb{R}$ an *activation function* that's applied **pointwise** and that's discriminatory in the sense that $\forall \mu \in \mathcal{M}^S(\mathcal{X})$:

$$\int_{\mathbb{R}^d} \sigma(a^T x + b) d\mu(x) = 0 \ \forall a \in \mathbb{R}^d, \ \forall b \in \mathbb{R} \Longrightarrow \mu \equiv 0$$

Let $g \in L^2(\mathcal{X}, \mathcal{Y}; \pi_x)$ be a function satisfying:

$$\forall z \in \mathcal{Z}, \ \int_{\mathcal{X}} \langle g(x), \sigma_*(x; z) \rangle_{\mathcal{Y}} d\pi_{\mathcal{X}}(x) = 0$$

We want to show that $g \equiv 0$ $\pi_{\mathcal{X}}$-a.e.

For this, notice that $g = \begin{pmatrix} g_1 \\ \vdots \\ g_c \end{pmatrix}$ with $\forall i \in \{1, \ldots, c\}, \ g_i : \mathbb{R}^d \to \mathbb{R}$. It will therefore be enough to prove that $\forall i \in \{1, \ldots, c\}, g_i \equiv 0$ $\pi_{\mathcal{X}}$-a.e. For this purpose, let $i \in \{1, \ldots, c\}$ and consider the measure given by $\mu_i(dx) = g_i(x)\pi_{\mathcal{X}}(dx)$, which lives in $\mathcal{M}^S(\mathcal{X})$. Now, consider arbitrary $A \in \mathbb{R}^d$ and $B \in \mathbb{R}$ and define:

$$\tilde{A} = (A|0|\ldots|0) \in \mathbb{R}^{d \times b}, \ \ \tilde{B} = \begin{pmatrix} B \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^b, \ \ \tilde{W} = (e_i|0|\ldots|0) \in \mathbb{R}^{c \times b}$$

where $e_i \in \mathbb{R}^c$ is the i-th canonical basis vector. With these, we have $\forall x \in \mathbb{R}^d$:

$$\tilde{A}^T x + \tilde{B} = \begin{pmatrix} A^T x + B \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^b, \text{ and thus: } \ \tilde{W}\sigma(\tilde{A}^T x + \tilde{B}) = \tilde{W} \begin{pmatrix} \sigma(A^T x + B) \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sigma(A^T x + B) e_i$$

i.e. for $\theta = (\tilde{W}, \tilde{A}, \tilde{B}) \in \mathcal{Z}$, we have:

$$\forall x \in \mathcal{X}, \ \sigma_*(x, \theta) = \tilde{W}\sigma(\tilde{A}^T x + \tilde{B}) = \sigma(A^T x + B) e_i$$

And evidently, for all $x \in \mathcal{X}$ we also have that:

$$\langle g(x), \sigma_*(x; \theta) \rangle_{\mathcal{Y}} = \sigma(A^T x + B) \langle g(x), e_i \rangle_{\mathcal{Y}} = \sigma(A^T x + B) g_i(x)$$

By our initial hypothesis on g, we know that, in particular, for $\theta \in \mathcal{Z}$:

$$\int_{\mathcal{X}} \langle g(x), \sigma_*(x; \theta) \rangle_{\mathcal{Y}} d\pi_{\mathcal{X}}(x) = 0$$

But rewriting the integrand, this tells us that:

$$\int_{\mathcal{X}} \langle g(x), \sigma_*(x; \theta) \rangle_{\mathcal{Y}} d\pi_{\mathcal{X}}(x) = \int_{\mathcal{X}} \sigma(A^T x + B) g_i(x) d\pi_{\mathcal{X}}(x) = \int_{\mathcal{X}} \sigma(A^T x + B) d\mu_i(x) = 0$$

But then, as $A \in \mathbb{R}^d$ and $B \in \mathbb{R}$ were arbitrary, we have that $\mu_i \in \mathcal{M}^S(\mathcal{X})$ is such that:

$$\forall A \in \mathbb{R}^d,\ \forall B \in \mathbb{R},\ \int_{\mathcal{X}} \sigma(A^T x + B) d\mu_i(x) = 0$$

So, by our discriminatory assumption, we get: $\mu_i \equiv 0$. In particular, as g_i is the density of μ_i with respect to $\pi_{\mathcal{X}}$, we must have that: $g_i \equiv 0$ $\pi_{\mathcal{X}}$-a.e. Now, as $i \in \{1, \ldots, c\}$ was arbitrary, we must conclude that $\forall i \in \{1, \ldots, c\},\ g_i \equiv 0$ $\pi_{\mathcal{X}}$-a.e. and therefore $g \equiv 0$ $\pi_{\mathcal{X}}$-a.e.

This concludes the proof, as we have shown that σ_* is discriminatory in the sense of assumption 1. □

Proof of Proposition 2 (as in Rotskoff and Vanden-Eijnden [72]). First, let's see that the class $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is indeed a subspace of $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$.

Let $f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$, by definition, $\exists \gamma \in \mathcal{M}^S(\mathcal{Z})$ such that $f(x) = \langle \sigma_*(x; \cdot), \gamma \rangle$ $\forall x \in \mathcal{X}$ $\pi_{\mathcal{X}} - a.s.$ (remember these are Bochner integrals). We then check:

$$\begin{aligned}
\|f\|^2_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} &= \int_{\mathcal{X}} \left\| \int_{\mathcal{Z}} \sigma_*(x,z)\gamma(dz) \right\|^2_{\mathcal{Y}} \pi_{\mathcal{X}}(dx) \\
&= \int_{\mathcal{X}} \langle \int_{\mathcal{Z}} \sigma_*(x,z)\gamma(dz), \int_{\mathcal{Z}} \sigma_*(x,z')\gamma(dz') \rangle_{\mathcal{Y}} \pi_{\mathcal{X}}(dx) \\
\textit{by linearity} &= \int_{\mathcal{X}} \int_{\mathcal{Z}} \int_{\mathcal{Z}} \langle \sigma_*(x,z), \sigma_*(x,z') \rangle_{\mathcal{Y}} \gamma(dz)\gamma(dz')\pi_{\mathcal{X}}(dx) \\
\textit{by Fubini} &= \int_{\mathcal{Z}\times\mathcal{Z}} K(z,z')\gamma(dz)\gamma(dz') \leqslant \|K\|_\infty |\gamma|^2_{TV} < \infty,
\end{aligned}$$

where we defined $K(z, z') = \int_{\mathcal{X}} \langle \sigma_*(x,z), \sigma_*(x,z') \rangle_{\mathcal{Y}} \pi_{\mathcal{X}}(dx)$. Thanks to assumption 1, $(z, z') \mapsto K(z, z')$ is well defined and **continuous**; thus by the compactness of $\mathcal{Z}$, we get that $\|K\|_\infty = \sup_{(z,z') \in \mathcal{Z}\times\mathcal{Z}} |K(z, z')| < \infty$. The fact that for $\gamma \in \mathcal{M}^S(\mathcal{Z})$, $|\gamma|_{TV} < \infty$ allows us to conclude.

Besides that, $\mathcal{F}(\mathcal{M}_{\mathcal{Z}})$ is clearly a **linear** subspace. □

Proof of theorem 1, as in Rotskoff and Vanden-Eijnden [72]. To show that $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is dense in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$, we proceed by contradiction. We know that the Bochner space $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ is a Hilbert Space.

Assuming that $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ is not dense, by the Hahn-Banach theorem, there exists a nonzero continuous linear functional $L : L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}}) \to \mathbb{R}$ such that $Lf = 0$ for all $f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$. By the Riesz representation theorem (on a Hilbert Space), the action of L on f can be represented as the inner product in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ between f and some $g \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$, $g \neq 0$. In particular, this $g \neq 0$ is such that for all $f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$: $\langle g, f \rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} = 0$. As for any $z \in \mathcal{Z}$, $\sigma_*(\cdot; z) = \langle \sigma_*, \delta_z \rangle \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$, we have then that $\forall z \in \mathcal{Z}$:

$$0 = \langle g, \sigma_*(\cdot; z) \rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} = \int_{\mathcal{X}} \langle g(x), \sigma_*(x,z) \rangle_{\mathcal{Y}}\ \mu(dx)$$

By our *discriminating* assumption, this would mean that $g = 0$ $\pi_{\mathcal{X}}$*-a.e. in* $\mathcal{X}$, which is a contradiction with the fact that $g \neq 0$ in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$. □

Proof of Proposition 3. Let $f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ and consider $\gamma \in \mathcal{M}^S(\mathcal{Z})$ such that $f = \int_{\mathcal{Z}} \sigma_*(\cdot, z)\gamma(dz)$. Consider the Jordan decomposition for γ, such that $\gamma = \gamma^+ - \gamma^-$, with $\gamma^+, \gamma^- \in \mathcal{M}(\mathcal{Z})$ (positive measures over $\mathcal{Z}$), satisfying $\operatorname{supp}(\gamma^+) \cup \operatorname{supp}(\gamma^-) = \operatorname{supp}(\gamma)$ and $\operatorname{supp}(\gamma^+) \cap \operatorname{supp}(\gamma^-) = \varnothing$.

Now, as $|\gamma|_{TV} = \int_{\mathcal{Z}}(\gamma^+(dz) + \gamma^-(dz)) < \infty$, consider the **probability measure** $\nu := (\gamma^+ + \gamma^-)/|\gamma|_{TV} \in \mathcal{P}(\mathcal{Z})$, and draw an independent sample $(Z_i)_{i\in\mathbb{N}}$ from it. Consider the function $\varphi : \mathcal{Z} \to \mathbb{R}$ defined as: $\varphi(z) := \begin{cases} +|\gamma|_{TV} & \text{if } \ z \in \operatorname{supp}(\gamma^+) \\ -|\gamma|_{TV} & \text{if } \ z \in \operatorname{supp}(\gamma^-) \end{cases}$ and set, for every $i \in \mathbb{N}$, $C_i = \varphi(Z_i)$. Therefore, the sample $(C_i, Z_i)_{i\in\mathbb{N}}$ is i.i.d. of law $\eta \in \mathcal{P}(\mathbb{R} \times \mathcal{Z})$ given, for any integrable $\hat{f} : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$, by: $\langle \hat{f}, \eta \rangle = \int_{\operatorname{supp}(\gamma^+)} \hat{f}(|\gamma|_{TV}, z)d\nu(z) + \int_{\operatorname{supp}(\gamma^-)} \hat{f}(-|\gamma|_{TV}, z)d\nu(z)$. Then, by the Law of Large Numbers, we get that for all integrable $\hat{f} : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}(C_i, Z_i) = \left\langle \hat{f}, \frac{1}{n}\sum_{i=1}^{n} \delta_{C_i, Z_i} \right\rangle \xrightarrow[n\to\infty]{\eta\text{-a.s.}} \left\langle \hat{f}, \eta \right\rangle$$

In particular, any function of the form $\hat{f}(c, z) = cf(z)$ with $f : \mathcal{Z} \to \mathbb{R}$ a ν-integrable function, will satisfy:

$$\left\langle \hat{f}, \eta \right\rangle = \int_{\operatorname{supp}(\gamma^+)} |\gamma|_{TV} f(z)d\nu(z) + \int_{\operatorname{supp}(\gamma^-)} -|\gamma|_{TV} f(z)d\nu(z)$$

Then, noticing that $\nu|_{\operatorname{supp}(\gamma^+)} = \frac{\gamma^+}{|\gamma|_{TV}}$ and $\nu|_{\operatorname{supp}(\gamma^-)} = \frac{\gamma^-}{|\gamma|_{TV}}$, we get that:

$$\begin{aligned} \left\langle \hat{f}, \eta \right\rangle &= \frac{1}{|\gamma|_{TV}} \int_{\operatorname{supp}(\gamma^+)} |\gamma|_{TV} f(z)d\gamma^+(z) - \frac{1}{|\gamma|_{TV}} \int_{\operatorname{supp}(\gamma^-)} |\gamma|_{TV} f(z)d\gamma^-(z) \\ &= \int_{\mathcal{Z}} f(z)d\gamma^+(z) - \int_{\mathcal{Z}} f(z)d\gamma^-(z) \\ &= \int_{\mathcal{Z}} f(z)d(\gamma^+ - \gamma^-)(z) = \langle f, \gamma \rangle \end{aligned}$$

This implies that for any γ-integrable $f : \mathcal{Z} \to \mathbb{R}$ we have:

$$\frac{1}{n}\sum_{i=1}^{n} C_i f(Z_i) = \left\langle f, \frac{1}{n}\sum_{i=1}^{n} C_i \delta_{Z_i} \right\rangle \xrightarrow[n\to\infty]{\eta\text{-a.s.}} \langle f, \gamma \rangle$$

Now, by a standard argument passing through a countable dense subset, we get that:

$$\eta\text{-a.s. } \forall f \in C_b(\mathcal{Z}),\ \frac{1}{n}\sum_{i=1}^{n} C_i f(Z_i) = \left\langle f, \frac{1}{n}\sum_{i=1}^{n} C_i \delta_{Z_i} \right\rangle \xrightarrow[n\to\infty]{} \langle f, \gamma \rangle$$

Choose **one particular realization of the random variables**, $(c_i, \theta_i)_{i\in\mathbb{N}}$ such that the above holds and recall that σ_* to be continuous and such that $\forall x \in \mathcal{X}$, $\pi_{\mathcal{X}}$-a.s., $\sigma_*(x, \cdot)$ is bounded (by assumption 1). Define $\gamma_n := \frac{1}{n}\sum_{i=1}^{n} c_i \delta_{\theta_i}$ and conclude that:

$$\pi_{\mathcal{X}}\text{-a.s., } \forall x \in \mathcal{X},\ f_n(x) := \langle \sigma_*(x, \cdot), \gamma_n \rangle \xrightarrow[n\to\infty]{} \langle \sigma_*(x, \cdot), \gamma \rangle = f(x)$$

which corresponds to the desired $\pi_{\mathcal{X}}$-a.s. approximation. If $\pi_{\mathcal{X}}$ is **compactly supported**, then by the continuity of σ_* over $\mathcal{X} \times \mathcal{Z}$ we get that σ_* is bounded on $\operatorname{supp}(\pi_{\mathcal{X}})$. Therefore $\pi_{\mathcal{X}}$-a.s. $\forall x \in \mathcal{X}$, $|f_n(x)| \leqslant \frac{1}{n}\sum_{i=1}^{n} |c_i||\sigma_*(x, \theta_i)| \leqslant |\gamma|_{TV} \|\sigma_*\|_{\infty, \operatorname{supp}(\pi_{\mathcal{X}}) \times \mathcal{Z}}$, allowing us to use the **dominated convergence theorem** to conclude that $\|f_n - f\|_{L^p(\pi_{\mathcal{X}})} \xrightarrow[n\to\infty]{} 0$. □

Proof of Corollary 1. Let $f \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ and $\varepsilon > 0$, by theorem 1 we know there exists $\gamma_\varepsilon \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ such that $\|f_{\gamma_\varepsilon} - f\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} < \frac{\varepsilon}{2}$. Also, by proposition 3 (thanks to the compact support of $\pi_{\mathcal{X}}$), there exists a set of parameters $(c_i^\varepsilon, \theta_i^\varepsilon)_{i \in \mathbb{N}} \subseteq \mathbb{R} \times \mathcal{Z}$ such that for some $N_\varepsilon \in \mathbb{N}$: $\|f_{\gamma_\varepsilon} - \Phi^{N_\varepsilon}_{(c^\varepsilon, \theta^\varepsilon)}\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} < \frac{\varepsilon}{2}$. Notice that $\Phi^{N_\varepsilon}_{(c^\varepsilon, \theta^\varepsilon)} \in \mathcal{N}_{\sigma_*}(\mathcal{Z})$. The triangle inequality allows us to conclude by noticing that:

$$\|f - \Phi^{N_\varepsilon}_{(c^\varepsilon, \theta^\varepsilon)}\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} \leqslant \|f_{\gamma_\varepsilon} - \Phi^{N_\varepsilon}_{(c^\varepsilon, \theta^\varepsilon)}\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} + \|f_{\gamma_\varepsilon} - f\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} \leqslant \varepsilon$$

i.e. $\forall f \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$, $\exists \Phi^{N_\varepsilon}_{(c^\varepsilon, \theta^\varepsilon)} \in \mathcal{N}_{\sigma_*}(\mathcal{Z})$ such that: $\|f - \Phi^{N_\varepsilon}_{(c^\varepsilon, \theta^\varepsilon)}\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} \leqslant \varepsilon$ allowing us to conclude. □

C.2 Proofs for Section 2.2.3

Proof of Lemma 1. By the characterization of the norm in a Hilbert space:

$$\begin{aligned}
R(f) &= \mathbb{E}_\pi[\|Y - f(X)\|_{\mathcal{Y}}^2] \\
&= \mathbb{E}_\pi[\|Y - f^*(X) + f^*(X) - f(X)\|_{\mathcal{Y}}^2] \\
&= \mathbb{E}_\pi[\|Y - f^*(X)\|_{\mathcal{Y}}^2 + 2\langle Y - f^*(X), f^*(X) - f(X)\rangle_{\mathcal{Y}} + \|f^*(X) - f(X)\|_{\mathcal{Y}}^2] \\
&= \mathbb{E}_\pi[\|Y - f^*(X)\|_{\mathcal{Y}}^2] + 2\mathbb{E}_\pi[\langle Y - f^*(X), f^*(X) - f(X)\rangle_{\mathcal{Y}}] + \mathbb{E}_\pi[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2]
\end{aligned}$$

The first term clearly doesn't depend on f, and the second term satisfies (by using the bilinearity of $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ and the measurability of f and f^*):

$$\begin{aligned}
\mathbb{E}_\pi[\langle Y - f^*(X), f^*(X) - f(X)\rangle_{\mathcal{Y}}] &= \mathbb{E}_\pi[\mathbb{E}_\pi[\langle Y - f^*(X), f^*(X) - f(X)\rangle_{\mathcal{Y}}|X]] \\
&= \mathbb{E}_\pi[\mathbb{E}_\pi[\langle Y|X] - f^*(X), f^*(X) - f(X)\rangle_{\mathcal{Y}}] \\
&= \mathbb{E}_\pi[\langle f^*(X) - f^*(X), f^*(X) - f(X)\rangle_{\mathcal{Y}}] = 0
\end{aligned}$$

Therefore, we conclude that: $R(f) = R_* + \mathbb{E}_\pi[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2]$, where the term $R_* = \mathbb{E}_\pi[\|Y - f^*(X)\|_{\mathcal{Y}}^2]$ is the Bayes risk, as it doesn't depend in f and the other term can be made zero by $f^* \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$.

Now, if any other f satisfies $R(f) = R_*$, then it means that $\mathbb{E}_\pi[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2 = 0$, which allows us to conclude that $f = f^*$ in $L^2(\mathcal{X}, \mathcal{Y}; \pi\mathcal{X})$. i.e. f^* is the *unique* minimizer. □

Proof of Lemma 2. As $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is such that $\pi|_{\mathcal{Y}}$ has a finite second moment, then $f^* = \mathbb{E}_\pi[Y|X = \cdot] \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$. Also, by lemma 1, we know that for any $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$:

$$R(f) = \mathbb{E}_\pi[\|Y - f(X)\|_{\mathcal{Y}}^2] = R_* + \mathbb{E}_\pi[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2]$$

By *universality*, given any $\varepsilon > 0$, we have that $\exists f_\varepsilon \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ such that $\|f_\varepsilon - f^*\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} < \varepsilon$. i.e: $\forall \varepsilon > 0$:

$$R_* \leqslant \inf_{f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))} R(f) \leqslant R(f_\varepsilon) < R_* + \varepsilon$$

By taking $\varepsilon \to 0$ we see that: $\inf_{f \in \mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))} R(f) = \inf_{\gamma \in \mathcal{M}^S(\mathcal{Z})} R(\gamma) = R^*$ (where we denote $R(\gamma)$ for $R(f_\gamma)$ with: $f_\gamma(x) = \langle \sigma_*(x, \cdot), \gamma\rangle \; \forall x \in \mathcal{X}$).

For the last part, it's clear that if there exists some measure $\mu^* \in \mathcal{M}^S(\mathcal{Z})$ such that $\inf_{\gamma \in \mathcal{M}^S(\mathcal{Z}))} R(\gamma)$ is attained, then in particular: $R(\mu^*) = \inf_{\gamma \in \mathcal{M}^S(\mathcal{Z}))} R(\gamma) = R_*$. By uniqueness of f^* as the minimizer of R in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$, this can only happen if

$$\forall x \in \mathcal{X} \; \pi_{\mathcal{X}}\text{-a.e.}, \; \langle \sigma_*(x, \cdot), \mu^*\rangle = f^*(x) = \mathbb{E}_\pi[Y|X = x]$$

□

C.3 Proofs for Section 2.2.4

Proof of Lemma 3. The implication to the left is direct from a previous characterization. Now, given $W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$, we know that $\forall f : \mathcal{Z} \to \mathbb{R}$ with *p-growth*, $\int_{\mathcal{Z}} f d\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{Z}} f d\mu$. Let $\mathcal{Y}$ be a separable real hilbert space, with $(e_k)_{k\in\mathbb{N}}$ an orthornormal Hilbert basis. For $y \in \mathcal{Y}$ we define: $\forall k \in \mathbb{N},\ P_k.y = \langle y, e_k \rangle$; and we know that $\|y\|_{\mathcal{Y}}^2 = \sum_{k\in\mathbb{N}} |P_k.y|^2$. In particular, let $f : \mathcal{Z} \to \mathcal{Y}$ be an arbitrary continuous function with p-growth, and define $y_n := \int_{\mathcal{Z}} f d\mu_n$ and $y = \int_{\mathcal{Z}} f d\mu$. We want to prove that $\|y_n - y\| \xrightarrow[n\to\infty]{} 0$. Indeed, notice that: $\forall k \in \mathbb{N},\ P_k.y_n = P_k. \int_{\mathcal{Z}} f d\mu_n = \int_{\mathcal{Z}} P_k.f d\mu_n$ (by linearity of the Bochner integral). Also, clearly: $\forall k \in \mathbb{N},\ \forall z \in \mathcal{Z},\ |P_k f(z)| \leqslant \|P_k\|_{BL(\mathcal{Y},\mathcal{Y})} \|f(z)\|_{\mathcal{Y}} \leqslant \|f(z)\|_{\mathcal{Y}} \leqslant C + C'\|z\|_{\mathcal{Z}}^p$. So, $\forall k \in \mathbb{N},\ P_k.f : \mathcal{Z} \to \mathbb{R}$ is a function with p-growth, and therefore: $P_k.y_n = \int_{\mathcal{Z}} P_k.f d\mu_n \to \int_{\mathcal{Z}} P_k.f d\mu = P_k.y$. In particular, $\forall N \in \mathbb{N},\ \sum_{k=1}^N |P_k.y_n - P_k.y|^2 \xrightarrow[n\to\infty]{} 0$. Notice that $\forall n \in \mathbb{N},\ \lim_{N\to\infty} \sum_{k=1}^N |P_k.y_n - P_k.y|^2 = \|y_n - y\|_{\mathcal{Y}}$; and as the series is *convergent*, it means that $\forall n \in \mathbb{N},\ \forall \varepsilon > 0,\ \exists M_n \in \mathbb{N} : \sum_{k=M_n}^\infty |P_k.y_n - P_k.y|^2 \leqslant \varepsilon$. From a standard diagonal argument one concludes that $\|y_n - y\|_{\mathcal{Y}} \xrightarrow[n\to\infty]{} 0$

□

Proof of Proposition 5. Indeed, let $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_p(\mathcal{Z})$ be a sequence converging to $\mu \in \mathcal{P}_p(\mathcal{Z})$. As σ_* is assumed to be continuous with p-growth on its second argument (i.e. $\forall x \in \mathcal{X},\ \forall z \in \mathcal{Z},\ \|\sigma_*(x,z)\|_{\mathcal{Y}} \leqslant C + C'\|z\|_{\mathcal{Z}}^p$ for some constants $C, C' > 0$), by lemma 3, we know that,

$$\pi_{\mathcal{X}}\text{-a.e. } \forall x \in \mathcal{X},\ \langle \sigma_*(x,\cdot), \mu_n \rangle \xrightarrow[n\to\infty]{} \langle \sigma_*(x,\cdot), \mu \rangle$$

As ℓ is assumed to be continuous, we get:

$$\pi_{\mathcal{X}}\text{-a.e. } \forall x \in \mathcal{X},\ \forall y \in \mathcal{Y},\ \ell(\langle \sigma_*(x,\cdot), \mu_n \rangle, y) \xrightarrow[n\to\infty]{} \ell(\langle \sigma_*(x,\cdot), \mu \rangle, y) \tag{C.1}$$

Now, as $\ell \geqslant 0$, we can employ Fatou's Lemma[1] to obtain that:

$$\mathbb{E}_\pi \left[\liminf_{n\to\infty} \left(\ell(\langle \sigma_*(X,\cdot), \mu_n \rangle, Y) \right) \right] \leqslant \liminf_{n\to\infty} \mathbb{E}_\pi \left[\ell(\langle \sigma_*(X,\cdot), \mu_n \rangle, Y) \right]$$

We can then notice that from equation (C.1), we have:

$$\mathbb{E}_\pi \left[\liminf_{n\to\infty} \left(\ell(\langle \sigma_*(X,\cdot), \mu_n \rangle, Y) \right) \right] = \mathbb{E}_\pi \left[\ell(\langle \sigma_*(X,\cdot), \mu \rangle, Y) \right] = R(\mu)$$

Putting both things together, we conclude the desired l.s.c. for W_p: $R(\mu) \leqslant \liminf_{n\to\infty} R(\mu_n)$
□

[1] Because all the involved functions are measurable

C.4 Proofs for Section 2.2.5

Proof of Proposition 6. We know that, $\forall \mu, \nu \in \mathcal{P}(\mathcal{Z}),\ h \in [0,1]$:

$$\begin{aligned}\frac{R((1-h)\mu + h\nu) - R(\mu)}{h} &= \frac{L(\langle \Phi, (1-h)\mu + h\nu\rangle) - L(\langle \Phi, \mu\rangle)}{h} \\ &= \frac{L(\langle \Phi, \mu\rangle + h\langle \Phi, \nu - \mu\rangle) - L(\langle \Phi, \mu\rangle)}{h}\end{aligned}$$

Let's denote by $q_\mu := \langle \Phi, \mu\rangle$ (analogously $q_{\nu-\mu} := \langle \Phi, \nu - \mu\rangle$) and $s_{\mu,\nu} := hq_{\nu-\mu}$, so we can write:

$$\frac{R((1-h)\mu + h\nu) - R(\mu)}{h} = \frac{L(q_\mu + s_{\mu,\nu}) - L(q_\mu)}{h}$$

As L is Gateaux differentiable, we have the following first order Taylor expansion $\forall x, s \in \mathcal{H}$, $\forall t \in \mathbb{R}$:

$$L(x + t\, s) = L(x) + t\, D_h L(x).s + o(|t| \|s\|)$$

In this particular case, we get:

$$\begin{aligned}\frac{R((1-h)\mu + h\nu) - R(\mu)}{h} &= \frac{L(q_\mu + h\, q_{\nu-\mu}) - L(q_\mu)}{h} \\ &= \frac{h.D_h L(q_\mu).q_{\nu-\mu} + o(|h|\|q_{\nu-\mu}\|)}{h} \\ &= D_h L(q_\mu).q_{\nu-\mu} + \frac{o(|h|\|q_{\nu-\mu}\|)}{h}\end{aligned}$$

As $\|q_{\nu-\mu}\| < \infty$ by hypothesis, we can say that: $o(|h|\|q_{\nu-\mu}\|) = o(h)$. Therefore, taking the limit with $h \to 0$, we get that:

$$\lim_{h\to 0} \frac{R((1-h)\mu + h\nu) - R(\mu)}{h} = D_h L(q_\mu).q_{\nu-\mu} + \lim_{h\to 0} \frac{o(h)}{h} = D_h L(q_\mu).q_{\nu-\mu}$$

Now, developping this last term (using, for instance, the linearity of the Bochner integral under bounded linear operators, as we know $D_h L(x, \cdot).(\cdot)$ to be linear and bounded as we're working on Hilbert Spaces), we get that:

$$D_h L(q_\mu).q_{\nu-\mu} = D_h L(q_\mu).\langle \Phi, \nu - \mu\rangle = \langle D_h L(q_\mu)(\Phi), \nu - \mu\rangle$$

Now, by the Riesz representation theorem, note that $\forall \theta \in \mathcal{Z}$: $D_h L(q_\mu)(\Phi(\theta)) = \langle \nabla_h L(q_\mu), \Phi(\theta)\rangle_{\mathcal{H}}$ and so:

$$\lim_{h\to 0} \frac{R((1-h)\mu + h\nu) - R(\mu)}{h} = \int_{\mathcal{Z}} \langle \nabla_h L(\langle \Phi, \mu\rangle), \Phi(\theta)\rangle_{\mathcal{H}}\, d(\nu - \mu)(\theta)$$

From where we deduce that:

$$\frac{\partial R}{\partial \mu}(\mu, \theta) = \langle \nabla_h L(\langle \Phi, \mu\rangle), \Phi(\theta)\rangle_{\mathcal{H}} - C_{R,\mu}$$

Where $C_{R,\mu}$ is a fixed constant, given by:

$$C_{R,\mu} = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, \theta) d\mu(\theta) = \int_{\mathcal{Z}} \langle \nabla_h L(\langle \Phi, \mu\rangle), \Phi(\theta)\rangle_{\mathcal{H}}\, d(\mu)(\theta) = \langle \nabla_h L(\langle \Phi, \mu\rangle), \langle \Phi, \mu\rangle\rangle_{\mathcal{H}}$$

On the other hand, for the intrinsic derivative, notice that $D_\theta(\frac{\partial R}{\partial \mu}(\mu,\theta)) : \mathcal{Z} \to \mathbb{R}$ is a bounded linear functional over $\mathcal{Z}$ a Hilbert space, so by Riesz representation, $\exists D_\mu R(\mu,\theta) := \nabla_\theta(\frac{\partial R}{\partial \mu}(\mu,\theta)) \in \mathcal{Z}$ such that:

$$\forall z \in \mathcal{Z},\ \left\langle \nabla_\theta \left(\frac{\partial R}{\partial \mu}(\mu,\theta) \right), z \right\rangle_{\mathcal{Z}} = D_\theta \left(\frac{\partial R}{\partial \mu}(\mu,\theta) \right)(z)$$

However, we can develop the RHS, and as the constant $C_{R,\mu}$ doesn't depend on θ, we get that:

$$D_\theta \left(\frac{\partial R}{\partial \mu}(\mu,\theta) \right)(z) = D_\theta \left(\langle \nabla_h L(\langle \Phi, \mu \rangle), \Phi(\theta) \rangle_{\mathcal{H}} \right)(z)$$

Now, by the chain rule and the definition of the *adjoint operator* of $D_\theta \Phi(\theta)$:

$$D_\theta \left(\frac{\partial R}{\partial \mu}(\mu,\theta) \right)(z) = \langle \nabla_h L(\langle \Phi, \mu \rangle), D_\theta \Phi(\theta)(z) \rangle_{\mathcal{H}} = \left\langle \left(D_\theta \Phi(\theta) \right)^* \left(\nabla_h L(\langle \Phi, \mu \rangle) \right), z \right\rangle_{\mathcal{Z}}$$

So, as they coincide for every $z \in \mathcal{Z}$, we conclude that:

$$D_\mu R(\mu,\theta) = \left(D_\theta \Phi(\theta) \right)^* \left(\nabla_h L(\langle \Phi, \mu \rangle) \right)$$

□

Proof of Lemma 4 (from Hu et al. [38]). Define $\mu^\varepsilon := (1-\varepsilon)\mu + \varepsilon\mu'$. Since R is convex, we have

$$\varepsilon \left(R(\mu') - R(\mu) \right) \geqslant R(\mu^\varepsilon) - R(\mu) = \int_0^\varepsilon \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^s, z) d(\mu' - \mu)(dz)\, ds.$$

Since the map $s \in [0,1] \mapsto \mu^s$ is continuous, it is of **compact image** (denoted $[\mu,\mu']$). In particular, as $\frac{\partial R}{\partial \mu}$ is continuous and bounded on its second argument, we get that, it is bounded on $[\mu,\mu'] \times \mathcal{Z}$. In particular the dominated convergence theorem and *Lebesgue differentiation theorem* (as $\varepsilon \to 0$) allows us to conclude. □

C.5 Proofs for Section 2.2.7

Proof of Theorem 2. In Theorem 2.4 from Hu et al. [38], using the hypothesis they prove that $\forall \mu \in \mathcal{P}_2(\mathcal{Z})$, $\forall Z = (Z_i)_{i=1}^N \overset{i.i.d.}{\sim} \mu$ over $\mathcal{Z}^N$, it holds that: $|\mathbb{E}[R(\nu_Z^N)] - R(\mu)| \leqslant \frac{2L}{N}$. From here, they assume that a minimizer exists (which might seem restrictive). It however holds without such assumption, as we can consider an arbitrary $\varepsilon > 0$ and some measure $\mu^* \in \mathcal{P}_2(\mathcal{Z})$ such that $R(\mu^*) \leqslant \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu) + \varepsilon$. Then, clearly $|\mathbb{E}[R(\nu_Z^N)] - \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu)| \leqslant \frac{2L}{N} + \varepsilon$. As $\mathbb{E}[R(\nu_Z^N)] \geqslant \inf_{\theta \in \mathcal{Z}^N} R\left(\nu_\theta^N\right) \geqslant \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu)$, we conclude that:

$$\left| \inf_{\theta \in \mathcal{Z}^N} R\left(\nu_\theta^N\right) - \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu) \right| \leqslant \frac{2L}{N} + \varepsilon$$

allowing us to conclude by taking $\varepsilon \to 0$. □

C.6 Proofs for Section 2.2.8

Proof of Proposition 9 (from Hu et al. [38]). To prove existence, consider any $\overline{\mu} \in \mathcal{P}(\mathcal{Z})$ such that $R_\nu^{\tau,\beta}(\overline{\mu}) < +\infty$. Consider the set:

$$S := \left\{ \mu \in \mathcal{P}(\mathcal{Z}) : \beta H_\nu(\mu) \leqslant R_\nu^{\tau,\beta}(\overline{\mu}) - \inf_{\mu' \in P(\mathcal{Z})} R^\tau(\mu') \right\}.$$

As a sublevel set of the relative entropy H, S is weakly compact. Together with the **weak lower semi-continuity** of $R_\nu^{\tau,\beta}$, the minimum of $R_\nu^{\tau,\beta}$ on S is attained[2]. As $\forall \mu \notin S$, $R_\nu^{\tau,\beta}(\mu) \geqslant R_\nu^{\tau,\beta}(\overline{\mu})$, the minimum of $R_\nu^{\tau,\beta}$ on S coincides with the global minimum. Since $R_\nu^{\tau,\beta}$ is strictly convex (thanks to the entropy), the minimizer is unique.

Now, let $\mu^* = \arg\min_{\mu \in P(\mathcal{Z})} R_\nu^{\tau,\beta}(\mu)$, we know $\mu^* \in S$, and thus $H(\mu^*) < +\infty$ as well as $E_{\mu^*}[U(X)] < \infty$. Therefore, μ^* is absolutely continuous with respect to the Gibbs measure, so also absolutely continuous with respect to the Lebesgue measure.

If further U satisfies assumption 3, the last point implies that $\mu^* \in \mathcal{P}_2(\mathcal{Z})$ □

Proof of Proposition 10. Let $(\tau_n)_{n \in \mathbb{N}}$ and $(\beta_n)_{n \in \mathbb{N}}$ be two positive sequences decreasing to 0. On the one hand, since R is continuous (weakly if $p = 0$ or in W_p for other $p \geqslant 1$) and $H_\nu(\mu) = D(\mu||\nu) \geqslant 0$, for all $\mu_n \to \mu$ (in the appropiate sense), we have

$$\liminf_{n \to +\infty} R_\nu^{\tau_n,\beta_n}(\mu_n) \geqslant \lim_{n \to +\infty} R(\mu_n) = R(\mu).$$

On the other hand, given $\mu \in \mathcal{P}_p(\mathcal{Z})$, consider ρ to be the heat kernel in $\mathcal{Z} = \mathbb{R}^D$ and $\rho_n(x) := \beta_n^{-D} \nu(x/\beta_n)$. In particular, from Ambrosio et al. [1] (as the heat kernel has finite p-th moments) we know that $\mu_n := \mu * \rho_n \xrightarrow[n \to \infty]{} \mu$ in W_p (or weakly if it is the case).

Now, since the function $h(x) := x \log(x)$ is convex, from Jensen's inequality we get that

$$\int_\mathcal{Z} h(\mu*\rho_n)dx \leqslant \int_\mathcal{Z}\int_\mathcal{Z} h\left(\rho_n(x-y)\right)\mu(dy)dx = \int_\mathcal{Z} h(\rho_n(x))dx = \int_\mathcal{Z} h(\rho(x))dx - D\,\log(\sqrt{2\beta_n}),$$

Besides, we have

$$\int_\mathcal{Z} (\mu * \rho_n)\log(g)dx = -\int_\mathcal{Z} \mu(dy) \int_\mathcal{Z} \rho_n(x) U(x-y)dx \geqslant -C\left(1 + \int_\mathcal{Z} |y|^2 \mu(dy)\right).$$

The last inequality is due to the quadratic growth of U; and by the same argument on r:

$$\int_\mathcal{Z} (\mu * \rho_n) r dx = \int_\mathcal{Z} \mu(dy) \int_\mathcal{Z} \rho_n(x) r(x-y)dx \leqslant C\left(1 + \int_\mathcal{Z} |y|^2 \mu(dy)\right).$$

Therefore, as R is W_p-continuous, $R(\mu_n) \xrightarrow[n \to \infty]{} R(\mu)$, and:

$$\begin{aligned}
&\limsup_{n \to +\infty} R_\nu^{\tau_n,\beta_n}(\mu * \nu_n) \\
&\qquad \leqslant R(\mu) + \limsup_{n \to +\infty} \tau_n \left(\int_\mathcal{Z} (\mu * \rho_n) r dx \right) \\
&\qquad\quad + \limsup_{n \to +\infty} \beta_n \left(\int_\mathcal{Z} h(\mu * \rho_n)dx - \int_\mathcal{Z} (\mu * \rho_n)\log(g)dx \right)
\end{aligned}$$

[2] Indeed, it is a lsc function over a compact set (under the weak topology).

And, as $\lim_{n\to\infty} \beta_n \log(\sqrt{2\beta_n}) = 0$ and the rest of the terms are bounded, we conclude that:

$$\limsup_{n\to+\infty} R_\nu^{\tau_n,\,\beta_n}(\mu * \rho_n) \leqslant R(\mu)$$

In particular, denoting by $\mu_*^{\tau,\beta,\nu}$ the unique minimizer of $R_\nu^{\tau,\beta}$, then from the previous expressions we get $\forall n \in \mathbb{N}$ and $\forall \mu \in \mathcal{P}_p(\mathcal{Z})$:

$$R(\mu_*^{\tau_n,\beta_n,\nu}) \leqslant R_\nu^{\tau_n,\beta_n}(\mu_*^{\tau_n,\beta_n,\nu}) \leqslant R_\nu^{\tau_n,\beta_n}(\mu * \rho_n)$$

So that,

$$\limsup_{n\to\infty} R(\mu_*^{\tau_n,\beta_n,\nu}) \leqslant \limsup_{n\to+\infty} R_\nu^{\tau_n,\beta_n}(\mu * \rho_n) \leqslant R(\mu), \quad \text{for all } \mu \in P_2(\mathcal{Z}).$$

□

C.7 Proofs for Section 3.2.2

Proof of Proposition 13. Indeed, , as $\mathbb{E}[\|Y\|^2] < \infty$, we know the conditional expectation $\mathbb{E}[Y|X]$ is well defined. From there, we know there exists a measurable $f^* : \mathcal{X} \to \mathcal{Y}$ s.t. $f^*(X) = \mathbb{E}[Y|X]$. In particular, we consider $\forall x \in \mathcal{X}$, $f^*(x) := \mathbb{E}[Y|X = x]$ and, consequently, $\xi = Y - f^*(X)$. We immediately notice that, as π is G-invariant, $f^* : \mathcal{X} \to \mathcal{Y}$ is G-equivariant, as (by properties of the conditional expectation): Given any $h : \mathcal{X} \to \mathbb{R}$ square integrable, we will show that: $\mathbb{E}_\pi[Yh(X)] = \mathbb{E}_\pi[\int_G \hat{\rho}_g^{-1}.f^*(\rho_g.X)d\lambda_G(g)h(X)]$. Indeed, notice that (in the first step using *Fubini's theorem*, as $f^* \in L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ and h square integrable):

$$\begin{aligned}
\mathbb{E}_\pi\left[\int_G \hat{\rho}_g^{-1}.f^*(\rho_g.X)d\lambda_G(g)h(X)\right] &= \int_G \mathbb{E}_\pi[\hat{\rho}_g^{-1}.f^*(\rho_g.X)h(X)]d\lambda_G(g) \\
\textit{linearity} &= \int_G \hat{\rho}_g^{-1}.\mathbb{E}_\pi[f^*(\rho_g.X)h(\rho_g.^{-1}.\rho_g.X)]d\lambda_G(g) \\
\pi_{\mathcal{X}} \textit{ is } G\textit{-invariant} &= \int_G \hat{\rho}_g^{-1}.\mathbb{E}_\pi[f^*(X)h(\rho_g.^{-1}.X)]d\lambda_G(g) \\
h \circ \rho_g^{-1} \geqslant 0, \textit{ measurable and def. of } f^* &= \int_G \hat{\rho}_g^{-1}.\mathbb{E}_\pi[Yh(\rho_g.^{-1}.X)]d\lambda_G(g) \\
\textit{linearity} &= \int_G \mathbb{E}_\pi[\hat{\rho}_g^{-1}.Yh(\rho_g.^{-1}.X)]d\lambda_G(g) \\
\pi \textit{ is } G\textit{-invariant} &= \int_G \mathbb{E}_\pi[Yh(X)]d\lambda_G(g) \\
&= \mathbb{E}_\pi[Yh(X)]
\end{aligned}$$

By uniqueness of the conditional expectation, we know therefore that $f^*(X) \overset{a.s.}{=} \int_G \hat{\rho}_g^{-1}.f^*(\rho_g.X)\lambda_G(g)$. In particular, $\pi_{\mathcal{X}}$-a.e. $f^* = \mathcal{Q}(f^*)$, implying that f^* is G-equivariant.

It is clear that ξ is centered, as by the tower property:

$$\mathbb{E}[\xi] = \mathbb{E}[Y - \mathbb{E}[Y|X]] = \mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y] - \mathbb{E}[Y] = 0$$

Analogously, notice that, from the property (with $\mathcal{H}$ some σ.field):

$$\begin{aligned}
\mathbb{E}[\|Y - \mathbb{E}[Y|\mathcal{H}]\|^2|\mathcal{H}] &= \mathbb{E}[\|Y\|^2|\mathcal{H}] + \mathbb{E}[\|\mathbb{E}[Y|\mathcal{H}]\|^2|\mathcal{H}] - 2\mathbb{E}[\langle Y, \mathbb{E}[Y|\mathcal{H}]\rangle|\mathcal{H}] \\
&= \mathbb{E}[\|Y\|^2|\mathcal{H}] + \|\mathbb{E}[Y|\mathcal{H}]\|^2 - 2\langle \mathbb{E}[Y|\mathcal{H}], \mathbb{E}[Y|\mathcal{H}]\rangle \\
&= \mathbb{E}[\|Y\|^2|\mathcal{H}] - \|\mathbb{E}[Y|\mathcal{H}]\|^2
\end{aligned}$$

We can derive a sort of *law of total variance*, in the sense that:

$$\mathbb{E}[\|Y - \mathbb{E}[Y]\|^2|] = \mathbb{E}[\mathbb{E}[\|Y - \mathbb{E}[Y|X]\|^2|X]] + \mathbb{E}[\|\mathbb{E}[Y|X] - \mathbb{E}[\mathbb{E}[Y|X]]\|^2|]$$

In particular, $\mathbb{E}[\mathbb{E}[\|Y - \mathbb{E}[Y|X]\|^2|X]] \leqslant \mathbb{E}[\|Y - \mathbb{E}[Y]\|^2|]$, leading us to:

$$\mathbb{E}[\|\xi\|^2] = \mathbb{E}[\|Y - \mathbb{E}[Y|X]\|^2] = \mathbb{E}[\mathbb{E}[\|Y - \mathbb{E}[Y|X]\|^2|X]] \leqslant \mathbb{E}[\|Y - \mathbb{E}[Y]\|^2|] < \infty$$

This implies, that ξ is a square integrable random vector (from properties of the Bochner integral). Let $h : \mathcal{X} \to \mathcal{Y}$ be any measurable function. Then, consider:

$$\begin{aligned}\mathbb{E}[\langle \xi, h(X)\rangle] &= \mathbb{E}[\langle Y - \mathbb{E}[Y|X], h(X)\rangle_{\mathcal{Y}}] \\ &= \mathbb{E}[\langle Y, h(X)\rangle_{\mathcal{Y}} - \langle \mathbb{E}[Y|X], h(X)\rangle_{\mathcal{Y}}] \\ &= \mathbb{E}[\langle Y, h(X)\rangle_{\mathcal{Y}}] - \mathbb{E}[\langle \mathbb{E}[Y|X], h(X)\rangle_{\mathcal{Y}}] \\ &= \mathbb{E}[\langle Y, h(X)\rangle_{\mathcal{Y}}] - \mathbb{E}[\mathbb{E}[\langle Y, h(X)\rangle_{\mathcal{Y}}|X]] \\ &= \mathbb{E}[\langle Y, h(X)\rangle_{\mathcal{Y}}] - \mathbb{E}[\langle Y, h(X)\rangle_{\mathcal{Y}}] \\ &= 0\end{aligned}$$

Where we've used standard properties of the conditional expectation (tower property, linearity, etc). The only property that may cause doubts is that for $\mathcal{H}$-measurable r.v. Z on $\mathcal{Y}$:

$$\mathbb{E}[\langle Y, Z\rangle_{\mathcal{Y}}|\mathcal{H}] \overset{a.s.}{=} \langle \mathbb{E}[Y|\mathcal{H}], Z\rangle_{\mathcal{Y}}$$

This can be shown by decomposing $\mathcal{Y}$ into its orthonormal basis (as its a separable Hilbert), $(e_n)_{n\in\mathbb{N}}$,and considering the projections $\forall n \in \mathbb{N}$, $\forall y \in \mathcal{Y}$ $P_n y = \langle y, e_n\rangle$, which are bounded linear operators. In particular, this implies (by using the *linearity* of the conditional expectation, as well as its known properties in $\mathbb{R}$, as $P_n(Z)$ is $\mathcal{H}$-measurable):

$$\forall n \in \mathbb{N},\ P_n(\mathbb{E}[Y|\mathcal{H}])P_n(Z) \overset{a.s.}{=} \mathbb{E}[P_n(Y)|\mathcal{H}]P_n(Z) \overset{a.s.}{=} \mathbb{E}[P_n(Y)P_n(Z)|\mathcal{H}]$$

So, for all $N \in \mathbb{N}$ (notice that, since there are countable cases, we can take a complete measure set where this holds for every $N \in \mathbb{N}$):

$$\sum_{n=1}^{N} P_n(\mathbb{E}[Y|\mathcal{H}])P_n(Z) \overset{a.s.}{=} \mathbb{E}[\sum_{n=1}^{N} P_n(Y)P_n(Z)|\mathcal{H}]$$

Finally, taking the limit with $N \to \infty$, we know that:

$$\begin{aligned}\langle \mathbb{E}[Y|\mathcal{H}], Z\rangle_{\mathcal{Y}} &= \sum_{n\in\mathbb{N}} P_n(\mathbb{E}[Y|\mathcal{H}])P_n(Z) \\ &= \mathbb{E}[\sum_{n\in\mathbb{N}} P_n(Y)P_n(Z)|\mathcal{H}] \\ &= \mathbb{E}[\langle Y, Z\rangle_{\mathcal{Y}}|\mathcal{H}]\end{aligned}$$

□

C.8 Proofs for Section 3.3

Proof of Lemma 10 (based on Elesedy and Zaidi [28] and Huang et al. [40]). As $H \leqslant G$ is a compact group and π is H-invariant, from proposition 13 we know that $(X, Y) \overset{a.s.}{=} (X, f^*(X) + \xi)$, with $f^* \in L^2(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})$ such that $\mathcal{Q}_H f^* = f^*$ (as functions in $L^2(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})$), and ξ being centered, with finite variance and such that for all measurable $h : \mathcal{X} \to \mathcal{Y}$, $\mathbb{E}[\langle \xi, h(X) \rangle_{\mathcal{Y}}] = 0$.

Consider any $f \in L^2(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})$ and decompose it, using Lemma 8 (thanks to the fact that $\pi|_{\mathcal{X}}$ is G-invariant), as $f = \overline{f}_G + f_G^{\perp}$, where $\overline{f}_G = \mathcal{Q}_G f$ is its symmetric part and $f_G^{\perp} = f - \mathcal{Q}_G f$ its antisymmetric part.

Since ξ is centered and has finite variance, we can write[3]:

$$\begin{aligned}
\Delta(f, \mathcal{Q}_G f) &= \mathbb{E}_\pi\left[\|Y - f(X)\|_{\mathcal{Y}}^2\right] - \mathbb{E}_\pi\left[\|Y - \overline{f}_G(X)\|_{\mathcal{Y}}^2\right] \\
&= \mathbb{E}_\pi\left[\|f^*(X) + \xi - f(X)\|_{\mathcal{Y}}^2\right] - \mathbb{E}_\pi\left[\|f^*(X) + \xi - \overline{f}_G(X)\|_{\mathcal{Y}}^2\right] \\
&= \mathbb{E}_\pi\left[\|f^*(X) - f(X) + \xi\|_{\mathcal{Y}}^2\right] - \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X) + \xi\|_{\mathcal{Y}}^2\right] \\
&= \mathbb{E}_\pi\left[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2 + 2\langle f^*(X) - f(X), \xi\rangle_{\mathcal{Y}} + \|\xi\|_{\mathcal{Y}}^2\right] \\
&\quad - \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2 + 2\langle f^*(X) - \overline{f}_G(X), \xi\rangle_{\mathcal{Y}} + \|\xi\|_{\mathcal{Y}}^2\right] \\
&= \mathbb{E}_\pi\left[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2\right] + 2\mathbb{E}_\pi[\langle f^*(X) - f(X), \xi\rangle_{\mathcal{Y}}] + \mathbb{E}_\pi[\|\xi\|_{\mathcal{Y}}^2] \\
&\quad - \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2 - 2\mathbb{E}_\pi[\langle f^*(X) - \overline{f}_G(X), \xi\rangle_{\mathcal{Y}}] - \mathbb{E}_\pi[\|\xi\|_{\mathcal{Y}}^2\right] \\
&= \mathbb{E}_\pi\left[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2\right] - \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2\right] \\
&\quad + 2\mathbb{E}_\pi[\langle f^*(X) - f(X), \xi\rangle_{\mathcal{Y}}] - 2\mathbb{E}_\pi[\langle f^*(X) - \overline{f}_G(X), \xi\rangle_{\mathcal{Y}}] \\
&= \mathbb{E}_\pi\left[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2\right] - \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2\right] + 0 - 0
\end{aligned}$$

Where we've used the defining properties of ξ and the fact that f, f^* and $\overline{f}_G$ are all measurable, as well. From here, the proof follows exactly as that of Huang et al. [40] (or Elesedy and Zaidi [28]); using the decomposition of f, we get that

$$\Delta(f, \mathcal{Q}_G f) = \mathbb{E}\left[\|f^*(X) - f(X)\|_{\mathcal{Y}}^2\right] - \mathbb{E}\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2\right]$$

can be written as:

$$\begin{aligned}
\Delta(f, \mathcal{Q}_G f) &= \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2 - 2\langle f^*(X) - \overline{f}_G(X), f_G^{\perp}(X)\rangle_{\mathcal{Y}} + \|f_G^{\perp}(X)\|_{\mathcal{Y}}^2\right] \\
&\quad - \mathbb{E}_\pi\left[\|f^*(X) - \overline{f}_G(X)\|_{\mathcal{Y}}^2\right] \\
&= -2\mathbb{E}_\pi\left[\langle f^*(X) - \overline{f}_G(X), f_G^{\perp}(X)\rangle_{\mathcal{Y}}\right] + \mathbb{E}_\pi\left[\|f_G^{\perp}(X)\|_{\mathcal{Y}}^2\right] \\
&= -2\langle f^* - \overline{f}_G, f_G^{\perp}\rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} + \|f_G^{\perp}\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}^2 \\
&= -2\langle f^*, f_G^{\perp}\rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} + \|f_G^{\perp}\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}^2
\end{aligned}$$

Where we used the definition of the inner product in $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ and also the fact that $\langle \overline{f}_G, f_G^{\perp}\rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} = 0$. The first term on the right hand side, $-2\langle f^*, f_G^{\perp}\rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}$, is what Huang et al. [40] call the *mismatch* between the real *underlying model* (which is only H-equivariant) and the *symmetrized* version of our model (which we *overdid*, as we made it entirely G-equivariant).

[3] We could actually skip all of these calculations, as they follow directly from Lemma 1. We do however include them to keep the proof similar to that of [28, 40]

Now, in the particular case of π being G-invariant, we would have that $\mathcal{Q}_G f^* = f^*$ (as functions in $L^2(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})$), and therefore, as $\overline{f}_G$ lives in the *orthogonal space*, we would get $-2\langle f^*, f_G^{\perp} \rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} = 0$, giving us the desired result:

$$\Delta(f, \mathcal{Q}_G f) = \|f_G^{\perp}\|^2_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}$$

□

C.9 Proofs for Section 3.4.3

Proof of the form of the orthogonal projection onto $\mathcal{E}^G$. To prove that the orthogonal projection from $\mathcal{L}$ onto $\mathcal{E}^G$ has the form:

$$\forall A \in \mathcal{L},\ P_{\mathcal{E}^G}(A) = \int_G \overline{\rho}(g).Ad\lambda_G(g)$$

we start by defining an operator $P : \mathcal{L} \to \mathcal{E}^G$ by:

$$\forall A \in \mathcal{L},\ P(A) = \int_G \overline{\rho}(g).Ad\lambda_G(g)$$

- **It is linear** as $\forall A, \tilde{A} \in \mathcal{L},\ \forall \lambda \in \mathbb{R}$
$$\int_G \overline{\rho}(g).(\lambda A - \tilde{A})d\lambda_G(g) = \int_G \lambda\overline{\rho}(g).A - \overline{\rho}(g).\tilde{A}d\lambda_G(g) = \lambda \int_G \overline{\rho}(g).Ad\lambda_G(g) - \int_G \overline{\rho}(g).\tilde{A}d\lambda_G(g)$$
so that:
$$P(\lambda A - \tilde{A}) = \lambda P(A) - P(\tilde{A})$$
- **We can see that** $P(A) \in \mathcal{E}^G,\ \forall A \in \mathcal{L}$; this follows from the fact that, for a fixed $g \in G$, and for arbitrary $A \in \mathcal{L}$:
$$\begin{aligned}\overline{\rho}(g)P(A) = \overline{\rho}(g)\int_G \overline{\rho}(h)Ad\lambda_G(h) &= \int_G \overline{\rho}(g)\overline{\rho}(h)Ad\lambda_G(h)\\ = \int_G \overline{\rho}(gh)Ad\lambda_G(h) &= \int_G \overline{\rho}(\tilde{h})Ad\lambda_G(\tilde{h}) = PA\end{aligned}$$
where we've used the invariance of the Haar measure λ_G. i.e. we have proven that:
$$\forall g \in G,\ \overline{\rho}(g)P = P$$
- **We can also see that** $P(A) = A,\ \forall A \in \mathcal{E}^G$. This is clear from the fact that $\forall A \in \mathcal{E}^G,\ \forall g \in G,\ \overline{\rho}(g)A = A$, so we get:
$$P(A) = \int_G \overline{\rho}(g).Ad\lambda_G(g) = \int_G Ad\lambda_G(g)$$
This means, in particular, that $P^2(A) = P(A),\ \forall A \in \mathcal{L}$, so $P : \mathcal{L} \to \mathcal{E}^G$ is a projection.
- **Finally, we see that** P **is orthogonal**, as for $\langle \cdot, \cdot \rangle$ the inner product on $\mathcal{L}$, we have $\forall A \in \mathcal{L},\ \forall \tilde{A} \in \mathcal{E}^G$:
$$\begin{aligned}\langle P(A), \tilde{A}\rangle &= \langle \int_G \overline{\rho}(g)Ad\lambda_G(g), \tilde{A}\rangle\\ \textit{by linearity } &= \int_G \langle \overline{\rho}(g)A, \tilde{A}\rangle d\lambda_G(g)\\ \textit{by orthogonality } &= \int_G \langle A, \overline{\rho}(g)^T\tilde{A}\rangle d\lambda_G(g)\\ As\ \tilde{A} \in \mathcal{E}, &= \int_G \langle A, \tilde{A}\rangle d\lambda_G(g)\\ &= \langle A, \tilde{A}\rangle\end{aligned}$$

With this, we have that $\forall A \in \mathcal{L}$, $\forall \tilde{A} \in \mathcal{E}^G$:

$$\langle A - P(A), \tilde{A} \rangle = 0$$

and so, P is **the orthogonal projection onto** $\mathcal{E}^G$.

$\square$

C.10 Proofs for Section 4.1

Proof of Proposition 19. Indeed, let $\theta = \left(\begin{pmatrix} w_i \\ a_i \end{pmatrix}\right)_{i=1}^N \in (\mathbb{R}^D)^N$, we know that:

$$\Phi_\theta^N(x) := \frac{1}{N}\sum_{i=1}^N w_i \sigma(a_i^T x)$$

We can now write this in *block* form: $W = (w_1, \dots, w_N)$, and $A = \begin{pmatrix} a_1^T \\ \hline \vdots \\ \hline a_N^T \end{pmatrix}$ so we can actually write the network as:

$$\Phi_\theta^N(x) := \frac{1}{N} W\sigma(AX)$$

Now, we know that Φ_θ^N is G-equivariant **iff** $\overline{\rho}(g).(W, A) = (W, A)\ \forall g \in G$ (with the action defined above). Now, under this action, the conditions amounts to $\forall g \in G$:

$$\overline{\rho}_1(g).W = (\hat{\rho}_g w_1 \eta_g^T, \dots, \hat{\rho}_g w_n \eta_g^T) = (w_1, \dots, w_n) \ \ \textit{and}\ \ \overline{\rho}_0(g).A = \begin{pmatrix} \eta_g a_1^T \rho_g^T \\ \hline \vdots \\ \hline \eta_g a_N^T \rho_g^T \end{pmatrix} = \begin{pmatrix} a_1^T \\ \hline \vdots \\ \hline a_N^T \end{pmatrix}$$

which is equivalent to asking: $\forall g \in G,\ \forall i \in \{1, \dots, N\},\ \hat{\rho}_g w_i \eta_g^T = w_i$ and $\rho_g a_i \eta_g^T = a_i$. i.e. $\forall g \in G,\ \ \forall i \in \{1, \dots, N\},\ M_g\vec{\theta_i} = \theta_i$, or equivalently, $\forall i \in \{1, \dots, N\},\ \theta_i \in \mathcal{E}^G$ □

Proof of Proposition 21. Indeed, let $\theta = (\theta_i)_{i=1}^N \in (\mathcal{E}^G)^N$ $x \in \mathcal{X}$ and $g \in G$, we can see that:

$$\begin{aligned}\Phi_\theta^N(\rho_g.x) &= \frac{1}{N}\sum_{i=1}^N \sigma_*(\rho_g.x; \theta_i) = \frac{1}{N}\sum_{i=1}^N \sigma_*(\rho_g.x; M_g\theta_i) \\ &= \frac{1}{N}\sum_{i=1}^N \hat{\rho}_g.\sigma_*(x; \theta_i) = \hat{\rho}_g.\frac{1}{N}\sum_{i=1}^N \sigma_*(x; \theta_i) = \hat{\rho}_g.\Phi_\theta^N(x)\end{aligned}$$

Where we've respectively used the fact that $\forall i \in \{1, \dots, N\},\ \theta_i \in \mathcal{E}^G$, the joint equivariance of σ_* and the linearity of the representation.

For $\mu \in \mathcal{P}(\mathcal{E}^G)$, the proof is exactly the same, only that lemma 13 (as $\mu \in \mathcal{P}(\mathcal{E}^G) \subseteq \mathcal{P}^G(\mathcal{Z})$) and proposition 30 are used. Indeed, $\forall x \in \mathcal{X}$:

$$\langle \sigma_*(\rho_g.x; \cdot), \mu \rangle = \langle \sigma_*(\rho_g.x; \cdot), M_g\#\mu \rangle = \hat{\rho}_g \langle \sigma_*(x; \cdot), \mu \rangle$$

□

Proof of Proposition 22. Let $x \in \mathbb{R}^d$ and $\vec{\theta} = \begin{pmatrix} \vec{w} \\ \vec{a} \end{pmatrix} \in \mathbb{R}^D$, with $w \in \mathbb{R}^{c \times b}$, $a \in \mathbb{R}^{d \times b}$. Let

$g \in G$, we see that:

$$\begin{aligned}
\sigma_*(\rho_g x, M_g.\theta) &= \sigma_*\left(\rho_g x, \begin{pmatrix} \hat{\rho}_g w \eta_g^T \\ \rho_g a \eta_g^T \end{pmatrix}\right) \\
&= \hat{\rho}_g w \eta_g^T \sigma((\rho_g a \eta_g^T)^T \rho_g x) \\
&= \hat{\rho}_g w \eta_g^T \sigma(\eta_g a^T \rho_g^T \rho_g x) \\
&= \hat{\rho}_g w \eta_g^T \eta_g \sigma(a^T x) \\
&= \hat{\rho}_g w \sigma(a^T x) = \hat{\rho}_g \sigma_*(x, \theta)
\end{aligned}$$

We've used the fact that all actions are orthogonal, and also that the activation function σ is G-equivariant by hypothesis. So, as this is satisfied $\forall x \in \mathbb{R}^d$, $\forall \vec{\theta} = \begin{pmatrix} \vec{w} \\ \vec{a} \end{pmatrix} \in \mathbb{R}^D$, σ_* is G-equivariant (on both arguments).

□

C.11 Proofs for Section 4.2

Now we consider the proof of other Lemmas involved in this work:

Proof of Proposition 23. This object is well defined, as:

- $\forall B \in \mathcal{B}_{\mathcal{Z}},\ [g \in G \mapsto \mu(M_g^{-1}(B))] \in L^\infty(G)$ (because $\mu(\cdot) \leqslant 1$); so the integral is always convergent.
- It defines a measure, as:
  - $\forall B \in \mathcal{B}_{\mathcal{Z}},\ \mu^G(B) = \int_G \mu(M_g^{-1}(B))d\lambda_G \geqslant 0$ (because $\mu(\cdot) \geqslant 0$).
  - For $(B_n)_{n\in\mathbb{N}}$ a family of pairwise disjoint borel subsets of $\mathcal{Z}$:

    $$\mu^G(\dot{\bigcup_{n\in\mathbb{N}}} B_n) = \int_G \mu(M_g^{-1}(\dot{\bigcup_{n\in\mathbb{N}}} B_n))d\lambda_G = \int_G \mu(\dot{\bigcup_{n\in\mathbb{N}}} M_g^{-1}(B_n))d\lambda_G$$

    Where the union is still disjoint, as $M_g^{-1}(B_k) \cap M_g^{-1}(B_\ell) \neq \phi \implies B_k \cap B_\ell \neq \phi$ (as for $x \in M_g^{-1}(B_k) \cap M_g^{-1}(B_\ell)$, we would have $M_g x \in B_k \cap B_\ell$). So, as μ is a measure (and using Fubini's Theorem):

    $$\mu^G(\dot{\bigcup_{n\in\mathbb{N}}} B_n) = \int_G \sum_{n\in\mathbb{N}} \mu(M_g^{-1}(B_n))d\lambda_G = \sum_{n\in\mathbb{N}} \int_G \mu(M_g^{-1}(B_n))d\lambda_G = \sum_{n\in\mathbb{N}} \mu^G(B_n)$$

  - Finally, using the fact that $\forall g \in G, M_g^{-1}(\mathcal{Z}) = \mathcal{Z}$, and that μ is a probability measure over $\mathcal{Z}$, we get:

    $$\mu^G(\mathcal{Z}) = \int_G \mu(M_g^{-1}(\mathcal{Z}))d\lambda_G = \int_G \mu(\mathcal{Z})d\lambda_G = \int_G 1 d\lambda_G = 1$$

  So, μ^G is a probability measure over $\mathcal{Z}$.

As for the corresponding properties, notice that:

1. Let $g \in G$ and $\mu \in \mathcal{P}(\mathcal{Z})$. As for any measurable $f : \mathcal{Z} \to \mathbb{R}$ we have:

   $$\langle f, (M_g\#\mu)^G\rangle = \int_G \langle f, M_h\#(M_g\#\mu)\rangle d\lambda_G(h) = \int_G \langle f, (M_{hg})\#\mu\rangle d\lambda_G(h)$$

   Then, using the right invariance of the normalized Haar measure, we conclude that:

   $$\langle f, (M_g\#\mu)^G\rangle = \int_G \langle f, (M_{\tilde{h}})\#\mu\rangle d\lambda_G(\tilde{h}) = \langle f, \mu^G\rangle$$

   i.e. $(M_g\#\mu)^G = \mu^G$

2. Let $a, b \in \mathbb{R}$ and $\mu, \nu \in \mathcal{P}(\mathcal{Z})$. Consider any measurable $f : \mathcal{Z} \to \mathbb{R}$ and notice that (using the definition and also the linearity of the group representation):

$$\langle f, (a\mu + b\nu)^G \rangle = \int_G \langle f, M_h \#(a\mu + b\nu)\rangle d\lambda_G(h) = \int_G \langle f, aM_h\#\mu + bM_h\#\nu\rangle d\lambda_G(h)$$

Again, using the linearity of the integral and the corresponding definitions, we get:

$$\begin{aligned}\langle f, (a\mu + b\nu)^G \rangle &= a\int_G \langle f, M_h\#\mu\rangle d\lambda_G(h) + b\int_G \langle f, M_h\#\nu\rangle d\lambda_G(h) \\ &= a\langle f, \mu^G\rangle + b\langle f, \nu^G\rangle \\ &= \langle f, a\mu^G + b\nu^G\rangle\end{aligned}$$

□

Proof of Lemma 12. On the backward direction, it is clear that, as $\forall B \in \mathcal{B}_{\mathcal{Z}},\ \mu(B) = \mu(P_{\mathcal{E}}^{-1}(B))$, and thanks to the fact that $\mathcal{Z} = P_{\mathcal{E}}^{-1}(\mathcal{E})$, we have: $1 = \mu(\mathcal{Z}) = \mu(P_{\mathcal{E}}^{-1}(\mathcal{E})) = \mu(\mathcal{E})$.

For the reverse implication, as $\mu(\mathcal{E}) = 1$, we have that $\forall B \in \mathcal{B}_{\mathcal{Z}},\ \mu(B) = \mu(B \cap \mathcal{E})$. We can also see, by properties of the projection, that:

- $\forall B \in \mathcal{B}_{\mathcal{Z}},\ P_{\mathcal{E}}^{-1}(B \cap \mathcal{E}) = P_{\mathcal{E}}^{-1}(B)$. [4]
- $\forall \tilde{B} \in \mathcal{B}_{\mathcal{Z}},\ \tilde{B} \subseteq \mathcal{E},\ P_{\mathcal{E}}^{-1}(\tilde{B}) \cap \mathcal{E} = \tilde{B}$. [5]

So, letting $B \in \mathcal{B}_{\mathcal{Z}}$, we can check that (using all of the above properties, considering $\tilde{B} = B \cap \mathcal{E}$):

$$\mu(B) = \mu(B \cap \mathcal{E}) = \mu(P_{\mathcal{E}}^{-1}(B \cap \mathcal{E}) \cap \mathcal{E}) = \mu(P_{\mathcal{E}}^{-1}(B \cap \mathcal{E})) = \mu(P_{\mathcal{E}}^{-1}(B))$$

i.e we have that $\mu = P_{\mathcal{E}}\#\mu = \mu^{\mathcal{E}}$. □

Proof of Proposition 24. 1. Let $\lambda \in [0, 1]$ and $\nu, \eta \in \mathcal{P}_p(\mathcal{Z})$. We will study what happens for $\tilde{\nu} = \lambda\nu + (1 - \lambda)\eta$. Consider γ_1 and γ_2 the optimal couplings between ν and μ, and η and μ respectively. Then, by marginalizing, we get that $(\gamma_3)|_X = \lambda(\gamma_1)|_X + (1 - \lambda)(\gamma_2)|_X = \lambda\nu + (1-\lambda)\eta = \tilde{\nu}$ and also $(\gamma_3)|_Y = \lambda(\gamma_1)|_Y + (1-\lambda)(\gamma_2)|_Y = \lambda\mu + (1-\lambda)\mu = \mu$. That is, $\gamma_3 := \lambda\gamma_1 + (1-\lambda)\gamma_2$ is a coupling between $\tilde{\nu}$ and μ. We can then get, using the minimality of $W_p(\tilde{\nu}, \mu)$ and the fact that γ_1 and γ_2 are optimal couplings, that:

$$\begin{aligned}W_p(\tilde{\nu}, \mu) &\leqslant \mathbb{E}_{\tilde{\mathbb{P}}}[\|X - Y\|^p] \\ &= \lambda\mathbb{E}_{\gamma_1}[\|X - Y\|^p] + (1 - \lambda)\mathbb{E}_{\gamma_2}[\|X - Y\|^p] \\ &= \lambda W_p(\nu, \mu) + (1 - \lambda)W_p(\eta, \mu)\end{aligned}$$

[4] $\subseteq$: direct from $B \cap \mathcal{E} \subseteq B$; $\supseteq$: as $\forall x \in \mathcal{Z},\ Px \in \mathcal{E}$; if $x \in P_{\mathcal{E}}^{-1}(B)$, we have $Px \in B$, but also $Px \in \mathcal{E}$, so $x \in P_{\mathcal{E}}^{-1}(B \cap \mathcal{E})$

[5] $\supseteq$: if $x \in \tilde{B} \subseteq \mathcal{E}$, then $P_{\mathcal{E}}x = x$ (as $P_{\mathcal{E}}$ is surjective, $x = P_{\mathcal{E}}y$ for some y, and $P_{\mathcal{E}}x = P_{\mathcal{E}}^2 y = P_{\mathcal{E}}y = x$), and so $x \in P_{\mathcal{E}}^{-1}(\tilde{B}) \cap \mathcal{E}$. $\subseteq$: if $x \in P_{\mathcal{E}}^{-1}(\tilde{B}) \cap \mathcal{E}$, then $x \in \mathcal{E}$ (implying $P_{\mathcal{E}}x = x$) and $P_{\mathcal{E}}x \in \tilde{B}$, so $x \in \tilde{B}$

i.e. we have proven that:

$$\forall \lambda \in [0,1],\ \forall \nu, \eta \in \mathcal{P}_p(\mathcal{Z}),\ W_p(\lambda\nu + (1-\lambda)\eta, \mu) \leqslant \lambda W_p(\nu, \mu) + (1-\lambda)W_p(\eta, \mu)$$

and so, the function $\nu \mapsto W_p(\nu, \mu)$ is convex.

2. The continuity follows directly (as the topology of $\mathcal{P}_p(\mathcal{Z})$ is generated by W_p).

3. The strict convexity can be proven using a standard argument presented in Santambrogio [73]. We replicate it here for completitude:

   Suppose by contradiction that $\nu_0 \neq \nu_1$ and $t \in (0,1)$ are such that $W_p(\nu_t, \mu) = (1-t)W_p(\nu_0, \mu) + tW_p(\nu_1, \mu)$, where $\nu_t = (1-t)\nu_0 + t\nu_1$. Let γ_0 be **the** optimal transport plan in the transport from μ to ν_0 (as $\mu \lll \lambda$, it is a transport map); we write $\gamma_0 = (T_0, \mathrm{id})_{\#}\mu$. Analogously, take $\gamma_1 = (T_1, \mathrm{id})_{\#}\mu$ optimal from μ to ν_1. Define $\gamma_t = (1-t)\gamma_0 + t\gamma_1 \in \Pi(\nu_t, \mu)$ (as in our first point). Then, we have:

$$(1-t)W_p^p(\nu_0, \mu) + tW_p^p(\nu_1, \mu) = W_p^p(\nu_t, \mu) \leqslant \int |x-y|^p \, d\gamma_t = (1-t)W_p(\nu_0, \mu) + tW_p(\nu_1, \mu),$$

   which implies that γ_t is actually optimal in the transport from μ to ν_t. Yet γ_t is not induced from a transport map unless $T_0 = T_1$. This is a contradiction with $\nu_0 \neq \nu_1$ and proves the desired strict convexity.

4. For the G-invariance, consider $G \circlearrowright_M \mathcal{Z}$ orthogonally and let $\mu, \nu \in \mathcal{P}_p(\mathcal{Z})$. We will prove that, $\forall g \in G,\ W_p(M_g\#\mu, M_g\#\nu) = W_p(\mu, \nu)$. Indeed, let $g \in G$; it is not hard to check[6] that, for any $\gamma \in \Pi(\mu, \nu)$, $\tilde{\gamma} := (M_g, M_g)\#\gamma \in \Pi(M_g\#\mu, M_g\#\nu)$. In particular:

$$W_p^p(M_g\#\mu, M_g\#\nu) \leqslant \mathbb{E}_{\tilde{\gamma}}\left[\|X - Y\|^p\right]$$

   Now, if we take γ to be an **optimal coupling** between μ and ν, we see that:

$$\begin{aligned}
\mathbb{E}_{\tilde{\gamma}}\left[\|X - Y\|^p\right] &= \int_{\mathcal{Z}\times\mathcal{Z}} \|x-y\|^p d\tilde{\gamma}(x,y) \\
&= \int_{\mathcal{Z}\times\mathcal{Z}} \|x-y\|^p d(M_g, M_g)\#\gamma(x,y) \\
&= \int_{\mathcal{Z}\times\mathcal{Z}} \|M_g x - M_g y\|^p d\gamma(x,y) \\
&= \int_{\mathcal{Z}\times\mathcal{Z}} \|M_g(x-y)\|^p d\gamma(x,y) \\
&= \int_{\mathcal{Z}\times\mathcal{Z}} \|x-y\|^p d\gamma(x,y) \\
&= \mathbb{E}_{\gamma}\left[\|X - Y\|^p\right] = W_p^p(\mu, \nu)
\end{aligned}$$

   where we've used the definition of the pushforward measure as well as the linearity and orthogonality of M_g (and the optimality of γ). Thus, we conclude that:

$$W_p^p(M_g\#\mu, M_g\#\nu) \leqslant W_p^p(\mu, \nu)$$

[6] Suffices to see that $\pi_i\#\tilde{\gamma} = M_g\#(\pi_i\#\gamma)$ for $i = 1, 2$

Conversely, for any coupling $\tilde{\eta} \in \Pi(M_g\#\mu, M_g\#\nu)$, we can define $\eta := (M_g^{-1}, M_g^{-1})\#\tilde{\eta} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$ and easily check that it is a coupling between μ and ν. By an exactly analogous argument, if we take $\tilde{\eta} \in \Pi(M_g\#\mu, M_g\#\nu)$ to be **optimal**, we get:

$$W_p^p(\mu,\nu) \leqslant \mathbb{E}_\eta[\|X-Y\|k] = \mathbb{E}_{\tilde{\eta}}[\|X-Y\|k] = W_p^p(M_g\#\mu, M_g\#\nu)$$

Thus, we can conclude that, $\forall g \in G$:

$$W_p^p(M_g\#\mu, M_g\#\nu) = W_p^p(\mu,\nu)$$

and so $W_p : \mathcal{P}_p(\mathcal{Z}) \times \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}_+$ is **jointly** G-invariant (in the sense of Definition 4.4).

□

Proof of Proposition 25. We will study each point separately:

1. We'll start dealing with the convexity. Consider $\lambda \in [0,1]$ and:
   - For $\mu, \nu \in \mathcal{P}_p(\mathcal{Z})$, $\tilde{\mu} := \lambda\mu + (1-\lambda)\nu \in \mathcal{P}_p(\mathcal{Z})$, as by definition

     $$\int_{\mathcal{Z}} \|\theta\|^p d\tilde{\mu}(\theta) = \lambda \int_{\mathcal{Z}} \|\theta\|^p d\mu(\theta) + (1-\lambda) \int_{\mathcal{Z}} \|\theta\|^p d\nu(\theta) < +\infty$$

   - Let $\mu, \nu \in \mathcal{P}^{\mathcal{E}}(\mathcal{Z})$, then by definition:

     $$(\lambda\mu + (1-\lambda)\nu)(\mathcal{E}) = \lambda\mu(\mathcal{E}) + (1-\lambda)\nu(\mathcal{E}) = \lambda + (1-\lambda) = 1$$

     So, $(\lambda\mu + (1-\lambda)\nu) \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$, and as μ, ν and λ are arbitrary, $\mathcal{P}^{\mathcal{E}}(\mathcal{Z})$ is convex. It follows directly that $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z}) = \mathcal{P}^{\mathcal{E}}(\mathcal{Z}) \cap \mathcal{P}_p(\mathcal{Z})$ is convex.
   - Analogously, let $\mu, \nu \in \mathcal{P}^G(\mathcal{Z})$. For $g \in G$ and a measurable $f : \mathcal{Z} \to \mathbb{R}_+$, we have:

     $$\begin{aligned}\int_{\mathcal{Z}} f(M_g\theta) d(\lambda\mu + (1-\lambda)\nu)(\theta) &= \lambda \int_{\mathcal{Z}} f(M_g\theta) d\mu(\theta) + (1-\lambda) \int_{\mathcal{Z}} f(M_g\theta) d\nu(\theta) \\ &= \lambda \int_{\mathcal{Z}} f(\theta) d\mu(\theta) + (1-\lambda) \int_{\mathcal{Z}} f(\theta) d\nu(\theta) \\ &= \int_{\mathcal{Z}} f(\theta) d(\lambda\mu + (1-\lambda)\nu)(\theta)\end{aligned}$$

     where we've used the fact that $M_g\#\mu = \mu$ and $M_g\#\nu = \nu$. Since this argument works for any positive and measurable $f : \mathcal{Z} \to \mathbb{R}$, we get that: $M_g\#(\lambda\mu + (1-\lambda)\nu = \lambda\mu + (1-\lambda)\nu$; and as g was arbitrary, $(\lambda\mu + (1-\lambda)\nu) \in \mathcal{P}^G(\mathcal{Z})$. As μ, ν and λ were arbitrary, we conclude that $\mathcal{P}^G(\mathcal{Z})$ is convex. As before, it follows directly that $\mathcal{P}_p^G(\mathcal{Z}) = \mathcal{P}^G(\mathcal{Z}) \cap \mathcal{P}_p(\mathcal{Z})$ is convex.

   We now focus on the closedness under the weak topology. Notice that by definition of the topology induced by W_p, $\mathcal{P}_p(\mathcal{Z})$ is closed under it. We can also say that:

   - If $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_p(\mathcal{Z})$ and $\mu \in \mathcal{P}(\mathcal{Z})$ such that $W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$, clearly $\mu \in \mathcal{P}_p(\mathcal{Z})$, as $(\mathcal{P}_p(\mathcal{Z}), W_p)$ is a **Polish space** (thus complete). We conclude that $\mathcal{P}_p(\mathcal{Z})$ is closed under the topology induced by W_p.

It will be enough to check that $\mathcal{P}^{\mathcal{E}}(\mathcal{Z})$ and $\mathcal{P}^{G}(\mathcal{Z})$ are closed under the topology induced by W_p, as by intersection of closed sets, so will $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ and $\mathcal{P}_p^{G}(\mathcal{Z})$.

Let $\mu \in \mathcal{P}(\mathcal{Z})$ and $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathcal{Z})$ such that $W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$. Recall that this convergence is equivalent to having $\mu_n \xrightarrow[n\to\infty]{} \mu$ and $\int_{\mathcal{Z}} \|\theta\|^p d\mu_n(\theta) \xrightarrow[n\to\infty]{} \int_{\mathcal{Z}} \|\theta\|^p d\mu(\theta)$ (Ambrosio et al. [1]). Also recall that $\mu_n \xrightarrow[n\to\infty]{} \mu$ means that $\forall f \in C_b(\mathcal{Z})$, $\langle f, \mu_n\rangle \xrightarrow[n\to\infty]{} \langle f, \mu\rangle$

- If $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}^G(\mathcal{Z})$, consider $g \in G$, and $f \in C_b(\mathcal{Z})$; it is not hard to see that $\tilde{f} := f \circ M_g$ is also continuous and bounded (as a composition of continuous functions and using the boundedness of f). Thus, we get by weak convergence:

  $$\langle \tilde{f}, \mu_n\rangle \xrightarrow[n\to\infty]{} \langle \tilde{f}, \mu\rangle$$

  But as $\mu_n \in \mathcal{P}^G(\mathcal{Z})$ for all $n \in \mathbb{N}$, we know that:

  $$\langle \tilde{f}, \mu_n\rangle = \langle f \circ M_g, \mu_n\rangle = \langle f, M_g\#\mu_n\rangle = \langle f, \mu_n\rangle$$

  And also by weak convergence (as $f \in C_b^0(\mathcal{Z})$): $\langle f, \mu_n\rangle \xrightarrow[n\to\infty]{} \langle f, \mu\rangle$. By uniqueness of the limit, we must have that: $\langle f, \mu\rangle = \langle \tilde{f}, \mu\rangle = \langle f \circ M_g, \mu\rangle = \langle f, M_g\#\mu\rangle$. i.e. We've shown that $\forall f \in C_b(\mathcal{Z}), \int_{\mathcal{Z}} f(M_g\theta)d\mu(\theta) = \int_{\mathcal{Z}} f(\theta)d\mu(\theta)$, by **density of $C_b(\mathcal{Z})$ in $L^\infty(\mathcal{Z})$**, we get that $\mu = M_g\#\mu$; and as $g \in G$ was arbitrary, we get $\mu \in \mathcal{P}^G(\mathcal{Z})$. Thus, we conclude that $\mathcal{P}^G(\mathcal{Z})$ is closed for the topology induced by W_p.

- Analogously, if $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}^{\mathcal{E}}(\mathcal{Z})$, we use an identical argument to prove that $\mu \in \mathcal{P}^{\mathcal{E}}$. For $f \in C_b(\mathcal{Z})$ we consider $\tilde{f} := f \circ P_{\mathcal{E}}$, which is also a continuous and bounded function (as $P_{\mathcal{E}}$ is continuous). Thus, we get by weak convergence:

  $$\langle \tilde{f}, \mu_n\rangle \xrightarrow[n\to\infty]{} \langle \tilde{f}, \mu\rangle$$

  But as $\mu_n \in \mathcal{P}^{\mathcal{E}}(\mathcal{Z})$ for all $n \in \mathbb{N}$, we know that:

  $$\langle \tilde{f}, \mu_n\rangle = \langle f \circ P_{\mathcal{E}}, \mu_n\rangle = \langle f, P_{\mathcal{E}}\#\mu_n\rangle = \langle f, \mu_n\rangle$$

  And also by weak convergence: $\langle f, \mu_n\rangle \xrightarrow[n\to\infty]{} \langle f, \mu\rangle$. By uniqueness of the limit, we must have that: $\langle f, \mu\rangle = \langle \tilde{f}, \mu\rangle = \langle f \circ P_{\mathcal{E}}, \mu\rangle = \langle f, P_{\mathcal{E}}\#\mu\rangle$. i.e. We've shown that $\forall f \in C_b(\mathcal{Z}), \int_{\mathcal{Z}} f(P_{\mathcal{E}}\theta)d\mu(\theta) = \int_{\mathcal{Z}} f(\theta)d\mu(\theta)$, by the same **density argument as before**, we get that $\mu = P_{\mathcal{E}}\#\mu$; and so $\mu \in \mathcal{P}^{\mathcal{E}}(\mathcal{Z})$. Thus, $\mathcal{P}^{\mathcal{E}}(\mathcal{Z})$ is closed for the topology induced by W_p.

In conclusion, by intersection of closed sets, we get that both $\mathcal{P}_p^{G}(\mathcal{Z})$ and $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ are **closed** under the topology induced by W_p.

2. We will study each case separately:

   - First, notice that for all $B \in \mathcal{B}_{\mathcal{Z}}$, $(P_{\mathcal{E}}\#\mu)(B) = \mu(P_{\mathcal{E}}^{-1}(B))$, in particular, as $P_{\mathcal{E}}^{-1}(\mathcal{E}) = \mathcal{Z}$, we have:

     $$(P_{\mathcal{E}}\#\mu)(\mathcal{E}) = \mu(P_{\mathcal{E}}^{-1}(\mathcal{E})) = \mu(\mathcal{Z}) = 1$$

On the other hand,

$$\begin{aligned}\int_{\mathcal{Z}} \|\theta\|^p (P_{\mathcal{E}}\#\mu)(d\theta) &= \int_{\mathcal{E}} \|\theta\|^p (P_{\mathcal{E}}\#\mu)(d\theta) \\ &= \int_{\mathcal{E}} \|P_{\mathcal{E}}\theta\|^p \mu(d\theta) \\ \textit{as } P_{\mathcal{E}} \textit{ fixes points in } \mathcal{E} &= \int_{\mathcal{E}} \|\theta\|^p \mu(d\theta) \\ &\leqslant \int_{\mathcal{Z}} \|\theta\|^p \mu(d\theta) < +\infty\end{aligned}$$

So, we have succesfully proven that $P_{\mathcal{E}}\#\mu \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$

- Let $h \in G$ and $B \in \mathcal{B}_{\mathcal{Z}}$, we note that (using the properties of the representation M and those of the Haar measure λ_G):

$$\begin{aligned}\mu^G(M_h^{-1}(B)) &= \int_G \mu(M_g^{-1}(M_h^{-1}(B)))d\lambda_G(g) \\ &= \int_G \mu((M_h M_g)^{-1}(B))d\lambda_G(g) \\ &= \int_G \mu(M_{hg}^{-1}(B))d\lambda_G(g) \\ &= \int_G \mu(M_{\tilde{g}}^{-1}(B))d\lambda_G(\tilde{g}) \\ &= \mu^G(B)\end{aligned}$$

So, $\forall g \in G,\ \mu^G = M_g\#\mu^G$. As, by definition of μ^G, for any positive and measurable $f : \mathcal{Z} \to \mathbb{R}$ we have:

$$\langle f, \mu^G \rangle = \int_G \langle f \circ M_g, \mu \rangle d\lambda_G(g)$$

It follows that (by Fubini's theorem):

$$\begin{aligned}\int_{\mathcal{Z}} \|\theta\|^p d\mu^G(\theta) &= \int_G \int_{\mathcal{Z}} \|M_g\theta\|^p d\mu(\theta) d\lambda_G(g) \\ &\leqslant \int_G \int_{\mathcal{Z}} \|M_g\|^p \|\theta\|^p d\mu(\theta) d\lambda_G(g) \\ &= \int_{\mathcal{Z}} \|\theta\|^p d\mu(\theta) \int_G \|M_g\|^p d\lambda_G(g)\end{aligned}$$

For an **orthogonal representation** M, we have that $\|M_g\| \leqslant 1,\ \forall g \in G$.

$$\int_{\mathcal{Z}} \|\theta\|^p d\mu^G(\theta) = \int_{\mathcal{Z}} \|\theta\|^p d\mu(\theta) \int_G \|M_g\|^p d\lambda_G(g) \leqslant \int_{\mathcal{Z}} \|\theta\|^p d\mu(\theta) < +\infty$$

So, $\mu^G \in \mathcal{P}_p^G(\mathcal{Z})$.

3. Let $\mu \in \mathcal{P}_p(\mathcal{Z})$. We will prove that $\mu^{\mathcal{E}}$ minimizes $W_p(\mu, \cdot)$ over $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$. For this purpose, recall that, by definition of the projection onto $\mathcal{E}$, $\forall z \in \mathcal{Z},\ \forall w \in \mathcal{E},\ \|z - P_{\mathcal{E}} z\| \leqslant \|z - w\|$. Let $\nu \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$, $\Pi(\mu, \nu)$ the set of couplings between μ and ν and consider $\gamma \in \Pi(\mu, \nu)$. We can prove[7] that $\gamma(\mathcal{Z} \times \mathcal{E}) = 1$, thus:

$$\mathbb{E}_\gamma \left[\|X - Y\|^p\right] = \int_{\mathcal{Z} \times \mathcal{Z}} \|z - w\| d\gamma(z, w) = \int_{\mathcal{Z} \times \mathcal{E}} \|z - w\| d\gamma(z, w)$$

By the **disintegration theorem** (Kallenberg [43]) we get that, for any measurable $f : \mathcal{Z} \times \mathcal{E} \to \mathbb{R}$ we can write:

$$\begin{aligned} \int_{\mathcal{Z} \times \mathcal{E}} f(z, w) d\gamma(z, w) &= \int_{\mathcal{Z}} \left(\int_{\mathcal{E}} f(z, w) d\gamma(w|z) \right) d(\pi_1 \# \gamma)(z) \\ \textit{by the definition of}\, \Pi(\mu, \nu) \quad &= \int_{\mathcal{Z}} \left(\int_{\mathcal{E}} f(z, w) d\gamma(w|z) \right) d\mu(z) \end{aligned}$$

In particular, for $f(z, w) = \|z - w\|^p$, we get:

$$\mathbb{E}_\gamma \left[\|X - Y\|^p\right] = \int_{\mathcal{Z}} \left(\int_{\mathcal{E}} \|z - w\|^p d\gamma(w|z) \right) d\mu(z)$$

Now, from the fact that $\forall z \in \mathcal{Z},\ \forall w \in \mathcal{E},\ \|z - P_{\mathcal{E}} z\|^p \leqslant \|z - w\|^p$, we get that: $\forall z \in \mathcal{Z}$:

$$\int_{\mathcal{E}} \|z - w\|^p d\gamma(w|z) \geqslant \int_{\mathcal{E}} \|z - P_{\mathcal{E}} z\|^p d\gamma(w|z)$$

So, in particular (as the term inside the integral on the RHS doesn't depend on w):

$$\begin{aligned} \int_{\mathcal{Z}} \left(\int_{\mathcal{E}} \|z - w\|^p d\gamma(w|z) \right) d\mu(z) &\geqslant \int_{\mathcal{Z}} \left(\int_{\mathcal{E}} \|z - P_{\mathcal{E}} z\|^p d\gamma(w|z) \right) d\mu(z) \\ &= \int_{\mathcal{Z}} \|z - P_{\mathcal{E}} z\|^p \left(\int_{\mathcal{E}} d\gamma(w|z) \right) d\mu(z) \\ &= \int_{\mathcal{Z}} \|z - P_{\mathcal{E}} z\|^p d\mu(z) \end{aligned}$$

Now, consider $\tilde{\gamma} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$, given by $\tilde{\gamma} = \mathbf{Law}((X, P_{\mathcal{E}} X))$, where $X \sim \mu$. This probability law is naturally a coupling $\tilde{\gamma} \in \Pi(\mu, \mu^{\mathcal{E}})$, as: $\pi_1 \# \tilde{\gamma} = \mathbf{Law}(X) = \mu$ and $\pi_2 \# \tilde{\gamma} = \mathbf{Law}(P_{\mathcal{E}} X) = P_{\mathcal{E}} \# \mu$ by known property. It also satisfies:

$$\mathbb{E}_{\tilde{\gamma}} \left[\|X - Y\|^p\right] = \int_{\mathcal{Z}} \|z - P_{\mathcal{E}} z\|^p d\mu(z)$$

So, by our previous calculation:

$$\mathbb{E}_\gamma \left[\|X - Y\|^p\right] \geqslant \mathbb{E}_{\tilde{\gamma}} \left[\|X - Y\|^p\right] \geqslant W_p^p(\mu, \mu^{\mathcal{E}})$$

where we recall that $\tilde{\gamma} \in \Pi(\mu, \mu^{\mathcal{E}})$ and that $W_p^p(\mu, \mu^{\mathcal{E}})$ is defined as an infimum over all such couplings. As we considered an arbitrary $\gamma \in \Pi(\mu, \nu)$, we get (by taking the infimum) that:

$$W_p^p(\mu, \mu^{\mathcal{E}}) \leqslant W_p^p(\mu, \nu)$$

[7] Indeed, $\gamma(\mathcal{Z} \times \mathcal{E}) = \mathbb{P}_\gamma(\pi_2((X, Y)) \in \mathcal{E}) = \pi_2 \# \gamma(\mathcal{E}) = \nu(\mathcal{E}) = 1$

which is valid $\forall \nu \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$. In particular, this means that

$$\mu^{\mathcal{E}} \in \arg \min_{\nu \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})} W_p(\mu, \nu)$$

Whenever $\mathcal{Z} = \mathbb{R}^D$, $\mu \lll \lambda$ and $p > 1$, we use the strict convexity given by proposition 24 to conclude it must be the **unique** global minimum.

4. For the last point notice that, from lemma 12, the assertion for $\mu^{\mathcal{E}}$ is direct. On the other hand, we know by the previous points that: $\mu = \mu^G \implies \mu \in \mathcal{P}_p^G(\mathcal{Z})$ and conversely, if $\mu \in \mathcal{P}_p^G(\mathcal{Z})$ then for any measurable $f : \mathcal{Z} \to \mathbb{R}$:

$$\langle f, \mu^G \rangle = \int_G \langle f, M_g \# \mu \rangle d\lambda_G(g) = \int_G \langle f, \mu \rangle d\lambda_G(g) = \langle f, \mu \rangle$$

so that $\mu = \mu^G$

□

Proof of Lemma 13. We can see that:

- Let $\mu \in \mathcal{P}^{\mathcal{E}^G}(\mathcal{Z})$ $g \in G$ and consider an arbitrary positive measurable $f : \mathcal{Z} \to \mathbb{R}$, we see that (using the fact that $\mu(\mathcal{E}^G) = 1$ and the definition of $\mathcal{E}^G$):

$$\int_{\mathcal{Z}} f(M_g\theta)\mu(d\theta) = \int_{\mathcal{E}^G} f(M_g\theta)\mu(d\theta) = \int_{\mathcal{E}^G} f(\theta)\mu(d\theta) = \int_{\mathcal{Z}} f(\theta)\mu(d\theta)$$

So, we conclude that $\mu = M_g\#\mu$ for an arbitrary $g \in G$, and so $\mu \in \mathcal{P}^G(\mathcal{Z})$.

- Analogously, let $\mu \in \mathcal{P}(\mathcal{Z})$ and $B \in \mathcal{B}_{\mathcal{Z}}$ be a borel set; we see that:

$$(\mu^G)^{\mathcal{E}^G}(A) = \mu^G(P_{\mathcal{E}^G}^{-1}(A)) = \int_G \mu(M_g^{-1}P_{\mathcal{E}^G}^{-1}(A))d\lambda_G(g)$$

And we can see that[8]: $\forall g \in G$, $\forall \theta \in \mathcal{Z}$ $P_{\mathcal{E}^G}M_g\theta = P_{\mathcal{E}^G}\theta$. Thus, $M_g^{-1}P_{\mathcal{E}^G}^{-1}(A) = (P_{\mathcal{E}^G}M_g)^{-1}(A) = (P_{\mathcal{E}^G})^{-1}(A)$, so that:

$$(\mu^G)^{\mathcal{E}^G}(A) = \int_G \mu(M_g^{-1}P_{\mathcal{E}^G}^{-1}(A))d\lambda_G(g) = \int_G \mu(P_{\mathcal{E}^G}^{-1}(A))d\lambda_G(g) = \mu(P_{\mathcal{E}^G}^{-1}(A)) = \mu^{\mathcal{E}^G}(A)$$

For the last equality, we notice that $\mu^{\mathcal{E}^G} \in \mathcal{P}^{\mathcal{E}^G}(\mathcal{Z}) \subseteq \mathcal{P}^G(\mathcal{Z})$, and so $(\cdot)^G$ leaves it as it was: $(\mu^{\mathcal{E}^G})^G = \mu^{\mathcal{E}^G}$.

□

Proof of Proposition 26. First of all, the first point, concerning $\mathcal{P}^{\mathcal{E}^G}(\mathcal{Z})$, we can see that:

[8] It's enough to notice that $P_{\mathcal{E}^G}M_g\theta = \int_G M_h(M_g\theta)d\lambda_G(h) = \int_G (M_hM_g\theta)d\lambda_G(h) = \int_G M_{\tilde{h}}\theta)d\lambda_G(\tilde{h}) = P_{\mathcal{E}^G}\theta$, as the Haar measure is G-invariant.

- For notational convenience, we will denote $\mathcal{E} := \mathcal{E}^G$ (though, in reality, nothing's special about $\mathcal{E}^G$ in the proof that follows).
- Given any positive and measurable $f : \mathcal{E} \to \mathbb{R}$, we see that:

$$\langle f, \mu^{\mathcal{E}} \rangle = \int_{\mathcal{Z}} f(P_{\mathcal{E}}.z)d\mu(z) = \int_{\mathcal{Z}} f(P_{\mathcal{E}}.z)u(z)d\lambda(z)$$

  By the disintegration theorem (see Kallenberg [43], as λ is σ-finite), we know that $\exists \varphi : \mathcal{E} \to \mathcal{E}^{\perp}$ a *measurable kernel* (consisting of probability measures) such that for every positive Borel measurable $g : \mathcal{Z} \cong \mathcal{E} \times \mathcal{E}^{\perp} \to \mathbb{R}$:

$$\int_{\mathcal{Z}} g(z)d\lambda(z) = \int_{\mathcal{E}} \int_{\mathcal{E}^{\perp}} g((x,y))d\varphi_x(y)dP_{\mathcal{E}}\#\lambda(x)$$

  In particular, as $\mathcal{Z} \cong \mathcal{E} \oplus \mathcal{E}^{\perp}$, we get:

$$\langle f, \mu^{\mathcal{E}} \rangle = \int_{\mathcal{Z}} f(P_{\mathcal{E}}.z)u(z)d\lambda(z) = \int_{\mathcal{E}} \int_{\mathcal{E}^{\perp}} f(P_{\mathcal{E}}.(x+y))u(x+y)d\varphi_x(y)dP_{\mathcal{E}}\#\lambda(x)$$

  Now, as the projection satisfies: $\forall x \in \mathcal{E}, \forall y \in \mathcal{E}^{\perp},\ P_{\mathcal{E}}(x+y) = x$, we get:

$$\begin{aligned}\langle f, \mu^{\mathcal{E}} \rangle &= \int_{\mathcal{E}} \int_{\mathcal{E}^{\perp}} f(P_{\mathcal{E}}.(x+y))u(x+y)d\varphi_x(y)d\lambda_{\mathcal{E}}(x) \\ &= \int_{\mathcal{E}} \int_{\mathcal{E}^{\perp}} f(x)u(x+y)d\varphi_x(y)d\lambda_{\mathcal{E}}(x) \\ &= \int_{\mathcal{E}} f(x) \int_{\mathcal{E}^{\perp}} u(x+y)d\varphi_x(y)d\lambda_{\mathcal{E}}(x) \\ &= \int_{\mathcal{E}} f(x)u^{\mathcal{E}}(x)d\lambda_{\mathcal{E}}(x)\end{aligned}$$

  Where we've defined $u^{\mathcal{E}}(x) = \int_{\mathcal{E}^{\perp}} u(x+y)d\varphi_x(y) \geqslant 0$ and, in particular, it satisfies:

$$\int_{\mathcal{E}} u^{\mathcal{E}}(x)d\lambda_{\mathcal{E}}(x) = \int_{\mathcal{E}} \int_{\mathcal{E}^{\perp}} u(x+y)d\varphi_x(y)d\lambda_{\mathcal{E}}(x) = \int_{\mathcal{Z}} u(z)d\lambda(z) = 1$$

  Thus, $\mu^{\mathcal{E}}$ has a density with respect to $\lambda_{\mathcal{E}} := P_{\mathcal{E}}\#\lambda$ (restricted to $\mathcal{E}$).
- As a direct consequence, if $\mathcal{E}$ is a **strict** linear subspace of $\mathcal{Z} = \mathbb{R}^D$, then any measure **concentrated** on it, **can't have** a density with respecto to λ of $\mathbb{R}^D$. This is the case of $\mathcal{E}^G$ as stated.

Concerning the case of $\mathcal{P}^G(\mathcal{Z})$, we can see that:

- As M is orthogonal, we have $\forall g \in G,\ \lambda = M_g\#\lambda$ by Lemma 14. i.e. $\lambda \in \mathcal{M}_{\mathcal{Z}}^G$

- Consider any positive and measurable $f : \mathcal{Z} \to \mathbb{R}$. We know that $\langle f, \mu \rangle = \int_{\mathcal{Z}} f(z)u(z)d\lambda(z)$. In particular, we can compute:

$$\begin{aligned}
\langle f, \mu^G \rangle &= \int_G \langle f, M_g \# \mu \rangle d\lambda_G(g) \\
&= \int_G \int_{\mathcal{Z}} f(M_g.\theta)d\mu(\theta)d\lambda_G(g) \\
&= \int_G \int_{\mathcal{Z}} f(M_g.z)u(z)d\lambda(z)d\lambda_G(g) \\
&= \int_G \int_{\mathcal{Z}} f(z)u(M_g^T.z)dM_g\#\lambda(z)d\lambda_G(g) \\
\textit{λ is G-invariant } &= \int_G \int_{\mathcal{Z}} f(z)u(M_g^T.z)d\lambda(z)d\lambda_G(g) \\
\textit{Fubini } &= \int_{\mathcal{Z}} f(z) \int_G u(M_{g^{-1}}.z)d\lambda_G(g)d\lambda(z) \\
\textit{λ_G invariant to inversion } &= \int_{\mathcal{Z}} f(z) \int_G u(M_g.z)d\lambda_G(g)d\lambda(z) \\
\textit{by definition } &= \int_{\mathcal{Z}} f(z)u^G(z)d\lambda(z)
\end{aligned}$$

  We also notice that, clearly $u^G \geqslant 0$ and

$$\begin{aligned}
\int_{\mathcal{Z}} u^G(z)d\lambda(z) &= \int_{\mathcal{Z}} \int_G u(M_g.z)d\lambda_G(g)d\lambda(z) = \int_G \int_{\mathcal{Z}} u(M_g.z)d\lambda(z)d\lambda_G(g) \\
&= \int_G \int_{\mathcal{Z}} u(z)dM_g\#\lambda(z)d\lambda_G(g) = \int_G \int_{\mathcal{Z}} u(z)d\lambda(z)d\lambda_G(g) = 1
\end{aligned}$$

  In particular, u^G is the (λ-a.s. unique) density of μ^G.

  As for the equivalence:

$\Longrightarrow$: By Theorem 12, as $\mu, \lambda \in \mathcal{M}_{\mathcal{Z}}^G$ and $\mu \lll \lambda$, we know that $\exists h : \mathcal{Z} \to \mathbb{R}_+$ measurable and G-invariant such that h is the density of μ.

We know that the **density** is λ-a.s. unique, so that λ-a.s. $u = h$ (and so, u is a.s. G-invariant).

$\Longleftarrow$: If u is λ-a.s. G-invariant (and measurable), we know that λ-*a.s.* $\forall z \in \mathcal{Z}$, $\forall g \in G$, $u(M_g.z) = u(z)$. In particular, let $\Omega_G := \{z \in \mathcal{Z} \ : \ \forall g \in G,\ u(z) = u(M_g.z)\}$ (such that $\lambda(\Omega_G^c) = 0$). We notice that, for every $z \in \Omega_G$, $u^G(z) = \int_G u(M_g.z)d\lambda_G(g) = \int_G u(z)d\lambda_G(g) = u(z)$. This means that λ-a.s. $u^G = u$, and therefore for any positive and measurable $\forall f : \mathcal{Z} \to \mathbb{R}$,

$$\langle f, \mu^G \rangle = \int_{\mathcal{Z}} f(z)u^G(z)d\lambda(z) = \int_{\mathcal{Z}} f(z)u(z)d\lambda(z) = \langle f, \mu \rangle$$

which allows us to conclude thanks to proposition 25.

□

Proof of Proposition 27. Following the standard *first isomorphism theorem* argument, notice that $\overline{\varphi} : G\backslash\mathcal{Z} \to \mathcal{Z}$ defined as $Gx \in G\backslash\mathcal{Z} \mapsto \varphi_x \in \mathcal{M}(\mathcal{Z})$ is well defined, as for any $x, y \in \mathcal{Z}$ such that $Gx = Gy$ we have $\varphi_x = \varphi_y$. It then clearly satisfies $\overline{\varphi}_{p(x)} = \varphi_x$ by definition; and thus it is unique as well. This function is measurable, as for any measurable $B \subseteq \mathcal{M}(\mathcal{Z})$, to check that $\overline{\varphi}^{-1}(B) \in \mathcal{B}_{G\backslash\mathcal{Z}}$, we need to check that $p^{-1}(\overline{\varphi}^{-1}(B)) \in \mathcal{B}_{\mathcal{Z}}$, notice, however, that $p^{-1}(\overline{\varphi}^{-1}(B)) = (\overline{\varphi} \circ p)^{-1}(B) = \varphi^{-1}(B) \in \mathcal{B}_{\mathcal{Z}}$. So, we get that $\overline{\varphi}$ is measurable and so it constitutes a kernel. □

Proof of Corollary 5. Given $\nu \in \mathcal{M}^G(\mathcal{Z})$, we know that there exists a unique $\mu \in \mathcal{M}(\varphi(\mathcal{Z}))$ such that $\nu = \int_{\mathcal{Z}} \varphi_x(\cdot) d\nu(x) = \int_{\varphi(\mathcal{Z})} m d\mu(m)$. Let $A \in \mathcal{B}_{\mathcal{Z}}$, we notice that, by definition of $\overline{\varphi}$: $\nu(A) = \int_{\mathcal{Z}} \varphi_x(A) d\nu(x) = \int_{\mathcal{Z}} \overline{\varphi}_{p(x)}(A) d\nu(x)$. Now, by the change of variables property (see lemma 6), we get that: $\nu(A) = \int_{\mathcal{Z}} \overline{\varphi}_{p(x)}(A) d\nu(x) = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(A) d\overline{\nu}(\overline{x})$ (where we denote $\overline{\nu} = p_{\#}\nu$).

We see that this measure is **unique** because, if there was another $\overline{\eta} \in \mathcal{M}(G\backslash\mathcal{Z})$ such that $\nu = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(\cdot) d\overline{\eta}(\overline{x})$, we could define, using our measurable cross-section, $\eta = (s\#\overline{\eta})^G \in \mathcal{M}^G(\mathcal{Z})$, and we would have $p\#\eta = \overline{\eta}$ (as $p \circ s = \mathrm{Id}_{G\backslash\mathcal{Z}}$ and $\forall g \in G, p \circ M_g = p$)). From here, we could define $\tilde{\mu} \in \mathcal{M}(\varphi(\mathcal{Z}))$ by $\int f d\tilde{\mu} = \int_{\mathcal{Z}} f(\varphi_z) d\eta(z)$. This would clearly imply (by the change of variable theorem) that $\int f d\tilde{\mu} = \int_{G\backslash\mathcal{Z}} f(\overline{\varphi}_{\overline{x}}) d\overline{\eta}(\overline{x})$. So, in particular, $\nu = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(\cdot) d\overline{\eta}(\overline{x}) = \int m d\tilde{\mu}(m)$. By uniqueness of μ in theorem 10, we get that $\mu = \tilde{\mu}$; thus, $\int f d\tilde{\mu} = \int_{\mathcal{Z}} f(\varphi_z) d\eta(z) = \int_{\mathcal{Z}} f(\varphi_z) d\nu(z) = \int f d\mu$. Now, as $\eta \in \mathcal{M}^G(\mathcal{Z})$, in particular $\eta = \int_{\mathcal{Z}} \varphi_z d\nu(z)$ and so replacing in the last equality, we get $\eta = \nu$ which naturally leads to the desired $\overline{\eta} = \overline{\nu}$. □

Proof of Proposition 28. Ψ is naturally well defined (as $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ is measurable). We can see that:

- Ψ **is injective**: Let $\mu_1, \mu_2 \in \mathcal{P}^G(\mathcal{Z})$ be such that $p\#\mu_1 = p\#\mu_2$. Then, we notice that, by corollary 5, considering the **orbit measure kernel** $\overline{\varphi} : G\backslash\mathcal{Z} \to \mathcal{Z}$, we have (for $i = 1, 2$): $\mu_i = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(\cdot) dp\#\mu_i(\overline{x})$. But as $p\#\mu_1 = p\#\mu_2$, we have that:

$$\mu_1 = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(\cdot) dp\#\mu_1(\overline{x}) = \int_{G\backslash\mathcal{Z}} \overline{\varphi}_{\overline{x}}(\cdot) dp\#\mu_2(\overline{x}) = \mu_2$$

  which proves the desired injectivity.

- Ψ **is surjective**: Let $\overline{\mu} \in \mathcal{P}(G\backslash\mathcal{Z})$, we define $\mu = (s\#\overline{\mu})^G$ (which is well defined as $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ is measurable; and the symmetrization is also well defined). We have that $p\#\mu = \overline{\mu}$, as for any $f \in \mathcal{M}(G\backslash\mathcal{Z}, \mathbb{R})$:

$$\int_{\mathcal{Z}} f(p(z)) d(s\#\overline{\mu})^G(z) = \int_G \int_{\mathcal{Z}} f(p(M_g.z)) d(s\#\overline{\mu})(z) d\lambda_G(g)$$

but, for any $g \in G$, $p(M_g.z) = p(z)$, therefore:

$$\begin{aligned}\int_{\mathcal{Z}} f(p(z))d(s\#\overline{\mu})^G(z) &= \int_G \int_{\mathcal{Z}} f(p(z))d(s\#\overline{\mu})(z)d\lambda_G(g)\\ &= \int_{\mathcal{Z}} f(p(z))d(s\#\overline{\mu})(z)\\ &= \int_{G\backslash\mathcal{Z}} f(p(s(\overline{z})))d(\overline{\mu})(\overline{z})\\ &= \int_{G\backslash\mathcal{Z}} f(\overline{z})d(\overline{\mu})(\overline{z})\end{aligned}$$

Where we've used the identity $p \circ s = \mathrm{Id}_{G\backslash\mathcal{Z}}$.

Therefore, Ψ is a bijection and its (unique) inverse is $\overline{\mu} \to (s\#\overline{\mu})^G$.

We can check that $\Psi : \mu \to p\#\mu$ is continuous. We endow $\mathcal{P}(G\backslash\mathcal{Z})$ with the *weak-* topology*. However, as $(G\backslash\mathcal{Z}, \mathcal{T}_{G\backslash\mathcal{Z}})$ isn't necessarily metrizable, we have to argue using *nets* (as it is known that for any topological space $\mathcal{X}$, $(\mu_\alpha)_{\alpha\in A}$ converges in $\mathcal{P}(\mathcal{X})$ **iff** $\forall f \in C_b(\mathcal{X})$, $(\langle f, \mu_\alpha\rangle)_{\alpha\in A}$ converges in $\mathbb{R}$). Indeed, to check continuity it's enough to take a net $(\mu_\alpha)_{\alpha\in A}$ converging to $\mu \in \mathcal{P}(\mathcal{Z})$ and notice that:

$$\langle f, \Psi(\mu_\alpha)\rangle = \langle f \circ p, \mu_\alpha\rangle \to \langle f \circ p, \mu\rangle = \langle f, \Psi(\mu)\rangle$$

where we've used the continuity of p and the fact that $\mu_\alpha \to \mu$ in $\mathcal{P}(\mathcal{Z})$. We conclude that for every net $(\mu_\alpha)_{\alpha\in A}$ converging to $\mu \in \mathcal{P}(\mathcal{Z})$, $(\Psi(\mu_\alpha))_{\alpha\in A}$ converges to $\Psi(\mu)$ in $\mathcal{P}(\mathcal{Z})$ (thus proving continuity). In particular, Ψ is also measurable.

It is also not hard to see that $(\cdot)^G : \mathcal{P}(\mathcal{Z}) \to \mathcal{P}(\mathcal{Z})$ is continuous, as for every weakly converging sequence $\mu_n \to \mu$, we have, for every $f : \mathcal{Z} \to \mathbb{R}$ bounded and continuous:

$$\langle f, \mu_n^G\rangle = \int_G \langle f \circ M_g, \mu_n\rangle d\lambda_G(g) \to \int_G \langle f \circ M_g, \mu\rangle d\lambda_G(g) = \langle f, \mu^G\rangle$$

where we've used the continuity of M_g, $\forall g \in G$, the weak convergence of the sequence and the dominated convergence theorem (as f is bounded). Therefore, to establish the desired *measurability of the inverse map* $\Psi(\mu)$, it suffices to check that $\overline{\mu} \mapsto s\#\overline{\mu}$ is a **measurable map**. This is a standard result, thanks to the measurability of s.

Further assuming that $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ is **continuous** is quite strong. In particular, as s is injective (it must be, as $p \circ s = \mathrm{Id}_{G\backslash\mathcal{Z}}$), it allows us to define a **metric** on $G\backslash\mathcal{Z}$ (via $d_{G\backslash\mathcal{Z}}(\overline{x}, \overline{y}) = d_{\mathcal{Z}}(s(\overline{x}), s(\overline{y}))$) that generates the topology of $G\backslash\mathcal{Z}$. In particular, weak convergence can be defined in $\mathcal{P}(G\backslash\mathcal{Z})$ and a direct computation (similar to what we've already done) shows that both Ψ and Ψ^{-1} are *weak* continuous. □

C.12 Proofs for Section 4.3

Proof of Proposition 29. Indeed, we know that $\forall z \in \mathcal{Z}$ $D_x f(\cdot, z)$ is the unique function that satisfies, $\forall \tilde{x} \in \mathcal{X}$:

$$\lim_{h \to 0} \frac{\|f(\tilde{x} + h, z) - f(\tilde{x}) - D_x f(\tilde{x}, z)h\|}{\|h\|} = 0$$

Now, we want to prove that $\forall \tilde{x} \in \mathcal{X}, \forall z \in \mathcal{Z}, \ \forall g \in G : D_x f(\chi_g.\tilde{x}, \check{\chi}_g.z) = \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1}$, for this, it will be enough to check that:

$$\lim_{h \to 0} \frac{\|f(\chi_g.\tilde{x} + h, \check{\chi}_g z) - f(\chi_g.\tilde{x}, \check{\chi}_g.z) - \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1} h\|}{\|h\|} = 0$$

since by uniqueness this will imply that $D_x f(\chi_g.\tilde{x}, \check{\chi}_g.z) = \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1}$. Now, thanks to the property satisfied by f, we have $\forall h \neq 0$:

$$\begin{aligned}
&\frac{\|f(\chi_g.\tilde{x} + h, \check{\chi}_g z) - f(\chi_g.\tilde{x}, \check{\chi}_g.z) - \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1} h\|}{\|h\|} \\
&= \frac{\|f(\chi_g.(\tilde{x} + \chi_g^{-1}.h), \check{\chi}_g z) - f(\chi_g.\tilde{x}, \check{\chi}_g.z) - \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1} h\|}{\|h\|} \\
&= \frac{\|\check{\chi}_g.f(\tilde{x} + \chi_g^{-1}.h, z) - \check{\chi}_g.f(\tilde{x}, z) - \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1} h\|}{\|h\|} \\
&= \frac{\|\check{\chi}_g.\left[f(\tilde{x} + \chi_g^{-1}.h, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)(\chi_g^{-1} h)\right]\|}{\|\chi_g.\chi_g^{-1}.h\|}
\end{aligned}$$

Now, recall that for every $g \in G$, the operator $\check{\chi}_g$ is **bounded**, i.e. it has a **finite operator norm** $0 < \|\check{\chi}_g\| < \infty$ (notice it is non-zero as $\check{\chi}_g$ is invertible). Now, notice that, defining $\tilde{h} := \chi_g^{-1}.h$

$$\frac{\|\check{\chi}_g.\left[f(\tilde{x} + \tilde{h}, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)\tilde{h}\right]\|}{\|\chi_g \tilde{h}\|} \leqslant \frac{\|\check{\chi}_g\| \|f(\tilde{x} + \tilde{h}, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)\tilde{h}\|}{\|\chi_g.\tilde{h}\|}$$

We can rewrite the last term to get:

$$\begin{aligned}
&\|\check{\chi}_g\| \cdot \frac{\|f(\tilde{x} + \tilde{h}, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)\tilde{h}\|}{\|\tilde{h}\|} \cdot \frac{\|\chi_g^{-1}\chi_g \tilde{h}\|}{\|\chi_g.\tilde{h}\|} \\
&\qquad \leqslant \|\check{\chi}_g\| \|\chi_g^{-1}\| \cdot \frac{\|f(\tilde{x} + \tilde{h}, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)\tilde{h}\|}{\|\tilde{h}\|}
\end{aligned}$$

Now, as χ_g and χ_g^{-1} are bounded operators, we have that: $h \to 0 \iff \tilde{h} = \chi_g^{-1} h \to 0$. Also, both $\|\check{\chi}_g\|$ and $\|\chi_g^{-1}\|$ are finite numbers, and so:

$$\begin{aligned}
&\lim_{h \to 0} \frac{\|f(\chi_g.\tilde{x} + h, \check{\chi}_g z) - f(\chi_g.\tilde{x}, \check{\chi}_g.z) - \check{\chi}_g D_x f(\tilde{x}, z)\chi_g^{-1} h\|}{\|h\|} \\
&\leqslant \lim_{h \to 0} \|\check{\chi}_g\| \|\chi_g^{-1}\| \cdot \frac{\|f(\tilde{x} + \tilde{h}, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)\tilde{h}\|}{\|\tilde{h}\|} \\
&\leqslant \|\check{\chi}_g\| \|\chi_g^{-1}\| \cdot \lim_{\tilde{h} \to 0} \frac{\|f(\tilde{x} + \tilde{h}, z) - f(\tilde{x}, z) - D_x f(\tilde{x}, z)\tilde{h}\|}{\|\tilde{h}\|} = 0
\end{aligned}$$

This concludes the proof.

□

Proof of Corollary 7. It suffices to see that the proof of Proposition 29 works for $f : \mathcal{X} \to \mathcal{Z}$; we could see f as a function $\hat{f} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, as $\hat{f}(x, z) = f(x)$, taking the trivial action on $\mathcal{Z}$. Also, we could notice that the property of f being G-equivariant, corresponds to $\hat{f}$ satisfying: $\forall g \in G,\ \forall x \in \mathcal{X},\ z \in \mathcal{Z},\ \hat{f}(\chi_g.x, \mathrm{Id}.z) = f(\chi_g.x) = \check{\chi}_g f(x) = \check{\chi}_g \hat{f}(x, z)$, which is exactly what we need to conclude the proof (following the same ideas of Proposition 29). □

Proof of Proposition 30. Let $\mu \in \mathcal{P}(\mathcal{Z})$ and f be as stated. Notice that, $\forall x \in \mathcal{X},\ \forall g \in G$, we have:

$$\check{\chi}_g \langle f(x, \cdot), \mu \rangle = \check{\chi}_g \int_{\mathcal{Z}} f(x, \theta) d\mu(\theta) = \int_{\mathcal{Z}} \check{\chi}_g f(x, \theta) d\mu(\theta)$$

Where we've used the **linearity of the Bochner integral** under **continuous linear operators** (such as is $\check{\chi}_g$). It follows, from the joint G-equivariance of f and the definition of the *pushforward measure*, that:

$$\int_{\mathcal{Z}} \check{\chi}_g f(x, \theta) d\mu(\theta) = \int_{\mathcal{Z}} f(\chi_g x, \check{\chi}_g \theta) d\mu(\theta) = \int_{\mathcal{Z}} f(\chi_g x, \tilde{\theta}) d(\check{\chi}_g \# \mu)(\tilde{\theta})$$

With this, we've succesfully proven that:

$$\check{\chi}_g \langle f(x, \cdot), \mu \rangle = \langle f(\chi_g x, \cdot), \check{\chi}_g \# \mu \rangle$$

which allows us to conclude as desired. □

C.13 Proofs for Section 4.4

Proof of Proposition 31. To prove this, recall that the linear functional derivative of R is the only function $\frac{\partial R}{\partial \mu} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathbb{R}$ satisfying $\forall \mu, \nu \in \mathcal{P}(\mathcal{Z})$:

$$\lim_{h\to 0} \frac{R((1-h)\mu + h\nu) - R(\mu)}{h} = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu,\theta) d(\nu - \mu)(\theta) \text{ and } \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu,\theta) d\mu(\theta) = 0$$

In particular, as R is G-invariant, we can write $\forall \mu, \nu \in \mathcal{P}(\mathcal{Z})$, $\forall h \neq 0$ and $g \in G$:

$$\begin{aligned}\frac{R((1-h)\mu + h\nu) - R(\mu)}{h} &= \frac{R(M_g\#((1-h)\mu + h\nu)) - R(M_g\#\mu)}{h} \\ &= \frac{R((1-h)(M_g\#\mu + h(M_g\#\nu) - R(M_g\#\mu)}{h}\end{aligned}$$

Taking the limit, we get:

$$\begin{aligned}\lim_{h\to 0} \frac{R((1-h)\mu + h\nu) - R(\mu)}{h} &= \lim_{h\to 0} \frac{R((1-h)(M_g\#\mu + h(M_g\#\nu) - R(M_g\#\mu)}{h} \\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, \theta) d(M_g\#\nu - M_g\#\mu)(\theta) \\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, \theta) dM_g\#(\nu - \mu)(\theta) \\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta) d(\nu - \mu)(\theta)\end{aligned}$$

Finally, we can also check that:

$$\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta) d\mu(\theta) = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, \theta) dM_g\#\mu(\theta) = 0$$

So, by the uniqueness of the linear functional derivative, we get $\forall g \in G$:

$$\frac{\partial R}{\partial \mu}(\mu, \theta) = \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta)$$

Finally, from proposition 29, as $\frac{\partial R}{\partial \mu}$ is jointly G-invariant, we get $\forall g \in G$:

$$D_\theta\left(\frac{\partial R}{\partial \mu}\right)(M_g.\mu, M_g.\theta) = D_\theta\left(\frac{\partial R}{\partial \mu}\right)(\mu, \theta).M_g^T$$

so that by considering the *gradient*, we get the desired result:

$$D_\mu R(M_g\#\mu, M_g.\theta) = M_g.D_\mu R(\mu, \theta)$$

□

Proof of Proposition 32. We know from lemma 4 that whenever R is convex and of class $\mathcal{C}^1$ the following inequality holds $\forall \mu_1, \mu_2 \in \mathcal{P}(\mathcal{Z})$:

$$R(\mu_1) \geqslant R(\mu_2) + \int \frac{\partial R}{\partial \mu}(\mu_2, z) d(\mu_1 - \mu_2)(z)$$

Let $\tilde{s} \in S$ be arbitrary and consider $\mu_2 = \tilde{\mu} := \int \mu_s d\lambda(s)$; and $\mu_1 = \mu_{\tilde{s}}$. Then:

$$R(\left(\int \mu_s d\lambda(s)\right) \leqslant R(\mu_{\tilde{s}}) - \int \frac{\partial R}{\partial \mu}\left(\int \mu_s d\lambda(s), z\right) d\left(\mu_{\tilde{s}} - \int \mu_s d\lambda(s)\right)(z)$$

Integrating the inequality with respect to λ (on $\tilde{s}$):

$$\int_S R\left(\int \mu_s d\lambda(s)\right) d\lambda(\tilde{s}) \leqslant \int_S R(\mu_{\tilde{s}}) d\lambda(\tilde{s}) - \int_S \left(\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int_S \mu_s d\lambda(s), \cdot\right) d\left(\mu_{\tilde{s}} - \int_S \mu_s d\lambda(s)\right)\right) d\lambda(\tilde{s}) \tag{C.2}$$

We notice that the LHS doesn't depend on $\tilde{s}$, so that $\int_S R\left(\int \mu_s d\lambda(s)\right) d\lambda(\tilde{s}) = R\left(\int \mu_s d\lambda(s)\right)$. On the other hand, the right-most term can be developed as:

$$\begin{aligned}
\star &= \int_S \left(\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int_S \mu_s d\lambda(s), \cdot\right) d\left(\mu_{\tilde{s}} - \int_S \mu_s d\lambda(s)\right)\right) d\lambda(\tilde{s}) \\
&= \int_S \left(\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int_S \mu_s d\lambda(s), \cdot\right) d\mu_{\tilde{s}} - \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int_S \mu_s d\lambda(s), \cdot\right) d\left(\int_S \mu_s d\lambda(s)\right)\right) d\lambda(\tilde{s}) \\
&= \int_S \left(\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int \mu_s d\lambda(s), \cdot\right) d(\mu_{\tilde{s}})\right) d\lambda(\tilde{s}) - \int_S \left(\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int_S \mu_s d\lambda(s), \cdot\right) d\left(\int_S \mu_s d\lambda(s)\right)\right) d\lambda(\tilde{s})
\end{aligned}$$

Notice that the **linear functional derivative** is chosen in such a way so that it satisfies $\forall \nu \in \mathcal{P}(\mathcal{Z})$, $\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\nu, \theta) d\nu(\theta) = 0$. In particular, the second term of the previous expression vanishes. We get that

$$\star = \int_S \left(\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int_S \mu_s d\lambda(s), \theta\right) d\mu_{\tilde{s}}(\theta)\right) d\lambda(\tilde{s})$$

But, by definition: $\forall f : \mathcal{Z} \to \mathbb{R}$ measurable,

$$\langle f, \int \mu_s d\lambda(s)\rangle = \int \langle f, \mu_s\rangle d\lambda(s) = \int_S \left(\int_{\mathcal{Z}} f(z) d\mu_s(z)\right) d\lambda(s)$$

So this is:

$$\star = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}\left(\int \mu_s, \theta\right) d\left(\int \mu_{\tilde{s}} d\lambda(\tilde{s})\right)(\theta) = 0$$

(by the same convention on the definition of the linear functional derivative). With this, we conclude that equation (C.2) turns into:

$$R(\tilde{\mu}) = R\left(\int_S \mu_s d\lambda(s)\right) \leqslant \int_S R(\mu_s) d\lambda(s)$$

which is what we wanted to prove. □

Proof of Proposition 33. Evidently, $\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) \geqslant \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$.

For the other inequality, take $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathcal{Z})$ an *infimizing sequence* for R (such that $R(\mu_n) \geqslant R(\mu_{n+1})$ and $R(\mu_n) \xrightarrow[n\to\infty]{} \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$); such a sequence always exists. By corollary 8, we have $\forall n \in \mathbb{N},\ R(\mu_n^G) \leqslant R(\mu_n)$; thus, $\forall n \in \mathbb{N}$:

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) \leqslant R(\mu_n^G) \leqslant R(\mu_n)$$

Which allows us to infer, by taking $n \to \infty$, that: $\inf_{\mu\in P^G(\mathcal{Z})} R(\mu) \leqslant \inf_{\mu\in P(\mathcal{Z})} R(\mu)$ Therefore, we conclude what we wanted:

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$$

The assertion about minima follow directly from corollary 8. □

Proof of Proposition 34. Thanks to the assumed conditions, we can establish the bijection given by Ψ in proposition 28. Then:

On the one hand, we immediately have that: $\forall \overline{\mu} \in \mathcal{P}(G\backslash\mathcal{Z}),\ \Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G \in \mathcal{P}^G(\mathcal{Z})$ Thus, $R((s\#\overline{\mu})^G) \geqslant \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu)$.

On the other hand, we might take $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}^G(\mathcal{Z})$ an *infimizing sequence*, such that $R(\mu_n) \xrightarrow[n\to\infty]{} \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu)$. Consider the *image through* Ψ of these measures, $(\overline{\mu_n})_{n\in\mathbb{N}} = (\Psi(\mu_n))_{n\in\mathbb{N}} \subseteq \mathcal{P}(G\backslash\mathcal{Z})$. We know that $\forall n \in \mathbb{N},\ \mu_n = \Psi^{-1}(\overline{\mu}_n) = (s\#\overline{\mu}_n)^G$ and thus:

$$\forall n \in \mathbb{N},\ \inf_{\overline{\mu}\in\mathcal{P}(G\backslash\mathcal{Z})} R((s\#\overline{\mu})^G) \leqslant R((s\#\overline{\mu}_n)^G) = R(\mu_n) \xrightarrow[n\to\infty]{} \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu)$$

So, by taking the limit with $n \to \infty$ we get:

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\overline{\mu}\in\mathcal{P}(G\backslash\mathcal{Z})} R((s\#\overline{\mu})^G)$$

□

Counterexample of an explicitly G-equivariant optimum (Proposition 35). Consider the group $G = C_4$ acting on $\mathbb{R}^2$ via 90° rotations. Let $K = B(0,1) \subseteq \mathbb{R}^2$ be a compact set. Consider a random variable $X \sim \mathcal{N}(0, \mathrm{Id}_2)|_K$ (i.e. given by $X = Z\mathbb{1}_{Z\in K}$ for $Z \sim \mathcal{N}(0, \mathrm{Id}_2)$) and the constant r.v. $Y \equiv 1$.

Clearly G is finite (thus compact) and it can be seen as its ortogonal representation:

$$\rho_G = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right\} \subseteq O(2)\ ,\ \hat{\rho}_G = \{\mathrm{Id}_1\} \subseteq O(1) \text{ (trivial repr.)}$$

By the definition of our r.v.s, it is clear that:

- $X \overset{(d)}{=} \rho_g X \ \ \forall g \in G$ because $X \sim N(0, \mathrm{Id}_2)$

- $Y \overset{(d)}{=} \hat{\rho}_g Y \;\; \forall g \in G$ and[9] $Y \perp X$; so, in particular, $\forall g \in G, \; (X,Y) \overset{(d)}{=} (\rho_g X, \hat{\rho}_g Y)$

Therefore, $\pi = \text{Law}(X,Y)$ is G-invariant (and also compactly supported). Consider a *shallow NN* given by: $\Phi_\theta^N : \mathbb{R}^2 \longrightarrow \mathbb{R}^{N\times b} \longrightarrow \mathbb{R}$ as: $\Phi_\theta^N(x) = \frac{1}{N}\sum_{i=1}^N w_i \sigma(a_i^T x), \quad \forall x \in \mathbb{R}^d$; where $\theta_i = \begin{pmatrix} w_i \\ a_i \end{pmatrix} \in \mathbb{R}^{3\times b} := \mathbb{R}^D$ with $b \in \mathbb{N}$ (also, considering some action $G \circlearrowright_\eta \mathbb{R}^b$). We let $G \circlearrowright_M \mathbb{R}^D$ as described in section 4.1.2, i.e.

$$M_g \theta_i = \begin{pmatrix} \hat{\rho}_g w_i \; \eta_g^T \\ \rho_g \; a_i \; \eta_g^T \end{pmatrix} = \begin{pmatrix} w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$

Assume, for instance, that $\eta_g \equiv \text{Id}_1$ (thus no condition is required for σ_* to be jointly G-equivariant) and recall that: $\theta_i \in \mathcal{E}^G \iff \forall g \in G, \;\; M_g\theta_i = \theta_i$. However, if we assume that: $\forall g \in G, \; \rho_g a_i = a_i$, then, in particular: $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} a_i^1 \\ a_i^2 \end{pmatrix} = \begin{pmatrix} a_i^1 \\ a_i^2 \end{pmatrix}$. This in turn implies, as $a_i^1 = -a_i^1$ and $a_i^2 = -a_i^2$ that $a_i^1 = a_i^2 = 0$. i.e. $a_i \equiv 0$. Thus, any $\theta_i = \begin{pmatrix} w_i \\ a_i \end{pmatrix} \in \mathcal{E}^G$ has $a_i = 0$. Therefore, if we choose any σ such that $\sigma(0) = 0$ (e.g $\sigma = \tanh$, that is $\mathcal{C}^\infty$ and bounded) and we choose $\theta_i \in \mathcal{E}^G \; \forall \; i = 1,...,N$; then:

$$\forall x \in \mathbb{R}^2, \; \Phi_\theta^{\mathcal{E}^G}(x) = \frac{1}{N}\sum_{1=1}^N w_i \sigma(0^T \cdot x) = \frac{1}{N}\sum_{i=1}^N w_i 0 = 0$$

i.e. **any equivariant architecture in this context satisfies** $\Phi_\theta^{\mathcal{E}^G}(x) \equiv 0$ (whereas $Y \equiv 1$). In particular: $\mathbb{E}[\|Y - \Phi_\theta^{\mathcal{E}^G}(x)\|^2] = 1, \; \forall N \in \mathbb{N}, \; \forall \theta \in (\mathcal{E}^G)^N$, and thus:

$$\inf_{\substack{\theta_i \in \mathcal{E}^G \\ i=1...N \\ N\in\mathbb{N}}} R(\Phi_\theta^{\mathcal{E}^G}) = 1 = \inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu)$$

On the other hand, consider a fully conected neuronal network. By the **universal approximation theorem** (with $\sigma = \tanh$), as π is **compactly supported** (in particular, $\pi_x(K) = 1$); we consider the **parameters** that approximate the function $f(x) \equiv 1$ in $K = B(0,1)$ to precision $\varepsilon > 0$. i.e. For $\varepsilon \in (0, \frac{1}{2})$, we know: $\implies \exists N \in \mathbb{N}, \; \exists a_1,...,a_N \in \mathbb{R}^2, \; \exists w_1,...,w_N \in \mathbb{R}^1$ such that:

$$\|\Phi_\theta - f\|_{\infty,K} = \sup_{x\in K} |\Phi_\theta(x) - f(x)| < \varepsilon < \frac{1}{2}$$

Then:

$$\mathbb{E}[|1 - \Phi_\theta(x)|^2] \leqslant \mathbb{E}[(\sup_{x\in K}|\Phi_\theta(x) - 1|)^2] < \mathbb{E}[\frac{1}{4}] = \frac{1}{4}$$

But, in particular, $\exists \nu_\theta^N \in \mathcal{P}(\mathcal{Z})$ such that:

$$\mathbb{E}[|Y - \Phi_\theta(x)|^2] < \frac{1}{4} < 1$$

[9] The independence comes from the fact that for any measurable f and h the following holds:

$$\mathbb{E}[f(X)h(Y)] = \mathbb{E}[f(X)h(1)] = h(1)\mathbb{E}[f(X)] = \mathbb{E}[f(X)]\mathbb{E}[h(Y)]$$

and so:

$$\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) \leqslant \inf_{\theta \in \mathcal{Z}^N} \mathbb{E}[|Y - \Phi_\theta(x)|^2] < \inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu)$$

In particular, we **can't expect** an optimum of the learning problem to be achieved within the space $\mathcal{P}(\mathcal{E}^G)$. $\square$

C.14 Proofs for Section 5.1

Proof of Proposition 38. By definition 3.8, we know that $\forall x \in \mathcal{X}$

$$(\mathcal{Q}f_\mu)(x) = \int_G \hat{\rho}_{g^{-1}} f_\mu(\rho_g x) d\lambda_G(g)$$

In particular:

$$\begin{aligned}
(\mathcal{Q}f_\mu)(x) &= \int_G \hat{\rho}_{g^{-1}} f_\mu(\rho_g x) d\lambda_G(g) \\
&= \int_G \hat{\rho}_{g^{-1}} \langle \sigma_*(\rho_g x, \cdot), \mu \rangle d\lambda_G(g) \\
&= \int_G \langle \hat{\rho}_{g^{-1}} \sigma_*(\rho_g x, \cdot), \mu \rangle d\lambda_G(g) \\
&= \langle \int_G \hat{\rho}_{g^{-1}} \sigma_*(\rho_g x, \cdot) d\lambda_G(g), \mu \rangle
\end{aligned}$$

Where we've used the linearity of the integral and Fubini's theorem. We have obtained that: $(\mathcal{Q}f_\mu)(x) = \langle \sigma_*^G(x, \cdot), \mu \rangle$.

Further assuming σ_* to be G-equivariant, as M_g is invertible $\forall g \in G$, $\langle \sigma_*(\rho_g x, \cdot), \mu \rangle = \langle \sigma_*(\rho_g x, \cdot), M_g \#(M_g^{-1} \# \mu) \rangle = \hat{\rho}_g \langle \sigma_*(x, \cdot), M_g^{-1} \# \mu \rangle$, where we've also used proposition 30 (because σ_* is jointly G-equivariant). So, we can write:

$$\begin{aligned}
(\mathcal{Q}f_\mu)(x) &= \int_G \hat{\rho}_{g^{-1}} \hat{\rho}_g \langle \sigma_*(x, \cdot), M_g^{-1} \# \mu \rangle d\lambda_G(g) \\
&= \int_G \langle \sigma_*(x, \cdot), M_g^{-1} \# \mu \rangle d\lambda_G(g) \\
\lambda_G \text{ is invariant to inversion} &= \int_G \langle \sigma_*(x, \cdot), M_g \# \mu \rangle d\lambda_G(g) \\
\textit{by definition } &= \langle \sigma_*(x, \cdot), \mu^G \rangle = f_{\mu^G}(x)
\end{aligned}$$

□

C.15 Proofs for Section 5.2

Proof of Proposition 39. Everything simply reduces to showing that the population risk $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is G-invariant as, by proposition 4, corollary 2, corollary 8 and proposition 33, we can conclude the rest of the proposition. Indeed, $\forall g \in G$ and $\forall \mu \in \mathcal{P}(\mathcal{Z})$,

$$\begin{aligned}
R(M_g\#\mu) &= \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), M_g\#\mu\rangle, Y)\right] \\
\textit{by proposition 30} \quad &= \mathbb{E}_\pi\left[\ell(\hat{\rho}_g\langle \sigma_*(\rho_g^{-1}X;\cdot), \mu\rangle, \hat{\rho}_g\hat{\rho}_g^{-1}Y)\right] \\
\textit{G-invariance of } \ell \quad &= \mathbb{E}_\pi\left[\ell(\langle \sigma_*(\rho_g^{-1}X;\cdot), \mu\rangle, \hat{\rho}_g^{-1}Y)\right] \\
\textit{G-invariance of } \pi \quad &= \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right] = R(\mu)
\end{aligned}$$

□

Proof of Proposition 40. From the fact that σ_* **is jointly G-equivariant**, following proposition 38, we get that $\forall x \in \mathcal{X}$: $\langle \sigma_*^G(x,\cdot), \mu\rangle = \langle \sigma_*(x,\cdot), \mu^G\rangle$. In particular, this means that:

$$R^{FA}(\mu) = \mathbb{E}_\pi\left[\ell\left(\langle \sigma_*^G(X,\cdot), \mu\rangle, Y\right)\right] = \mathbb{E}_\pi\left[\ell\left(\langle \sigma_*(X,\cdot), \mu^G\rangle, Y\right)\right] = R(\mu^G)$$

Even more directly, $R^{EA}(\mu) = \mathbb{E}_\pi\left[\ell\left(\langle \sigma_*(X,\cdot), \mu^{\mathcal{E}^G}\rangle, Y\right)\right] = R(\mu^{\mathcal{E}^G})$.

On the other hand, using the joint G-equivariance of σ_* (via proposition 30) and the G-invariance of ℓ one can easily compute that $\forall x \in \mathcal{X},\ \forall y \in \mathcal{Y},\ \forall g \in G$:

$$\begin{aligned}
L_{\rho_g.x,\hat{\rho}_g.y}(M_g\#\mu) &= \ell\left(\langle \sigma_*(\rho_g.x,\cdot), M_g\#\mu\rangle, \hat{\rho}_g.y\right) \\
&= \ell\left(\hat{\rho}_g.\langle \sigma_*(x,\cdot), \mu\rangle, \hat{\rho}_g.y\right) \\
&= \ell\left(\langle \sigma_*(x,\cdot), \mu\rangle, y\right) = L_{x,y}(\mu)
\end{aligned}$$

$$\text{i.e. } L_{\rho_g.x,\hat{\rho}_g.y}(M_g\#\mu) = L_{x,y}(\mu)$$

This is equivalent to stating that the map $L : \mathcal{P}(\mathcal{Z}) \to L^2(\mathcal{X}\times\mathcal{Y},\pi)$ given by $L(\mu) \mapsto [(x,y) \mapsto L_{x,y}(\mu)]$ is G-equivariant under the *correct* G-actions[10].

Notice then that, using the Haar's measure invariance under *inversion*, we get:

$$\begin{aligned}
R^G(\mu) &= \int_G \mathbb{E}_\pi\left[L_{\rho_g.X,\hat{\rho}_g.Y}(\mu)\right] d\lambda_G(g) \\
&= \int_G \mathbb{E}_\pi\left[L_{\rho_g.X,\hat{\rho}_g.Y}(M_g\#M_g^{-1}\#\mu)\right] d\lambda_G(g) \\
&= \int_G \mathbb{E}_\pi\left[L_{X,Y}(M_{g^{-1}}\#\mu)\right] d\lambda_G(g) \\
&= \int_G \mathbb{E}_\pi\left[L_{X,Y}(M_g\#\mu)\right] d\lambda_G(g) = \int_G R(M_g\#\mu) d\lambda_G(g)
\end{aligned}$$

Finally, for the G-invariance, notice that, for $g \in G$ and $\mu \in \mathcal{P}(\mathcal{Z})$:

$$R^{FA}(M_g\#\mu) = R((M_g\#\mu)^G) = R(\mu^G) = R^{FA}(\mu)$$

[10] Where $\forall g \in G$, we consider $g.\mu = M_g\#\mu$ and $\forall x \in \mathcal{X},\ \forall y \in \mathcal{Y},\ g.(f(x,y)) = f(g^{-1}.x, g^{-1}.y)$

Where we've used the fact that: $\forall g \in G,\ (M_g\#\mu)^G = \mu^G$ (See proposition 23). Analogously, $R^{EA}(M_g\#\mu) = R((M_g\#\mu)^{\mathcal{E}^G}) = R((P_{\mathcal{E}^G} \circ M_g)\#\mu) = R(P_{\mathcal{E}^G}\#\mu) = R^{EA}(\mu)$.

Finally,

$$\begin{aligned} R^G(M_g\#\mu) &= \int_G R((M_h\#(M_g\#\mu))d\lambda_G(h) \\ &= \int_G R((M_{hg}\#\mu))d\lambda_G(h) \\ &= \int_G R((M_{\tilde{h}}\#\mu))d\lambda_G(\tilde{h}) = R^G(\mu) \end{aligned}$$

Where we've used the right-invariance of the Haar Measure. □

Proof of Proposition 41. For the convexity, it is not hard to notice that, for $\lambda \in [0,1],\ \mu, \nu \in \mathcal{P}(\mathcal{Z})$:

$$\begin{aligned} R^G(\lambda\mu + (1-\lambda)\nu) &= \int_G R(M_g\#(\lambda\mu + (1-\lambda)\nu))d\lambda_G(g) \\ &= \int_G R(\lambda M_g\#\mu + (1-\lambda)M_g\#\nu)d\lambda_G(g) \\ &\leqslant \int_G \lambda R(M_g\#\mu) + (1-\lambda)R(M_g\#\nu)d\lambda_G(g) \\ &\leqslant \int_G \lambda R(M_g\#\mu)d\lambda_G(g) + \int_G (1-\lambda)R(M_g\#\nu)d\lambda_G(g) \\ &\leqslant \lambda R^G(\mu) + (1-\lambda)R^G(\nu) \end{aligned}$$

And similarly:

$$\begin{aligned} R^{FA}(\lambda\mu + (1-\lambda)\nu) &= R((\lambda\mu + (1-\lambda)\nu)^G) = R(\lambda\mu^G + (1-\lambda)\nu^G) \\ &\leqslant \lambda R(\mu^G) + (1-\lambda)R(\nu^G) = \lambda R^{FA}(\mu) + (1-\lambda)R^{FA}(\nu) \end{aligned}$$

An analogous argument justifies the convexity of R^{EA} (as $P_{\mathcal{E}^G}$ is linear).

As for the expressions of the linear functional derivatives, we can formally determine, by definition, that, given $\mu, \nu \in \mathcal{P}(\mathcal{Z})$:

$$\begin{aligned} \lim_{h\to 0} \frac{R^G((1-h)\mu + h\nu) - R^G(\mu)}{h} &= \lim_{h\to 0} \frac{\int_G R(M_g\#((1-h)\mu + h\nu))d\lambda_G(g) - \int_G R(M_g\#\mu)d\lambda_G(g)}{h} \\ &= \lim_{h\to 0} \int_G \frac{R((1-h)M_g\#\mu + hM_g\#\nu) - R(M_g\#\mu)}{h} d\lambda_G(g) \\ &= \int_G \lim_{h\to 0} \frac{R((1-h)M_g\#\mu + hM_g\#\nu) - R(M_g\#\mu)}{h} d\lambda_G(g) \\ &= \int_G \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, \theta)d(M_g\#\nu - M_g\#\mu)(\theta)d\lambda_G(g) \\ &= \int_G \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta)d(\nu - \mu)(\theta)d\lambda_G(g) \\ &= \int_{\mathcal{Z}} \int_G \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta)d\lambda_G(g)d(\nu - \mu)(\theta) \end{aligned}$$

Where we have used *Fubini's theorem*, which is applicable[11] thanks to the fact that R is of class $\mathcal{C}^1$, and we've used the definition of the linear functional derivative for R. Also, we see that (using Fubini's Theorem once again, as well as the definition of the linear functional derivative of R):

$$\begin{aligned}\int_{\mathcal{Z}}\int_G \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta)d\lambda_G(g)d\mu(\theta) &= \int_G\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta)d\mu(\theta)d\lambda_G(g)\\ &= \int_G\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(M_g\#\mu, \theta)d(M_g\#\mu)(\theta)d\lambda_G(g)\\ &= \int_G 0 d\lambda_G(g) = 0\end{aligned}$$

So, by the definition of the lineal functional derivative, we can identify that:

$$\frac{\partial R^G}{\partial \mu}(\mu,\theta) = \int_G \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.\theta)d\lambda_G(g)$$

Thus, formally taking the gradient we also get:

$$D_\mu R^G(\mu,\theta) = \int_G M_g^T.D_\mu R(M_g\#\mu, M_g.\theta)d\lambda_G(g)$$

Analogously, we calcuate the expression for the l.f.d. of R^{FA}; let $\mu,\nu \in \mathcal{P}(\mathcal{Z})$:

$$\begin{aligned}\lim_{h\to 0}\frac{R^{FA}((1-h)\mu + h\nu) - R^{FA}(\mu)}{h} &= \lim_{h\to 0}\frac{R(((1-h)\mu + h\nu)^G) - R(\mu^G)}{h}\\ &= \lim_{h\to 0}\frac{R((1-h)\mu^G + h\nu^G) - R(\mu^G)}{h}\\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^G,\theta)d(\nu^G - \mu^G)(\theta)\\ &= \int_G\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^G, M_g.\theta)d(\nu-\mu)(\theta)d\lambda_G(g)\\ \text{Fubini } &= \int_{\mathcal{Z}}\int_G \frac{\partial R}{\partial \mu}(\mu^G, M_g.\theta)d\lambda_G(g)d(\nu-\mu)(\theta)\end{aligned}$$

And also:

$$\begin{aligned}\int_{\mathcal{Z}}\int_G \frac{\partial R}{\partial \mu}(\mu^G, M_g.\theta)d\lambda_G(g)d\mu(\theta) &= \int_G\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^G, M_g.\theta)d\mu(\theta)d\lambda_G(g)\\ &= \int_G\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^G, \theta)d(M_g\#\mu)(\theta)d\lambda_G(g)\\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^G,\theta)d\mu^G(\theta) = 0\end{aligned}$$

[11] In particular, as for any fixed $\mu \in \mathcal{P}(\mathcal{Z})$ the function $g \in G \mapsto M_g\#\mu$ is continuous (thus, of **compact** image), then the function $(g,z) \in G \times \mathcal{Z} \mapsto \frac{\partial R}{\partial \mu}(M_g\#\mu, M_g.z)$ is **bounded**

So, by the definition of the lineal functional derivative, we can identify that:

$$\frac{\partial R^{FA}}{\partial \mu}(\mu, \theta) = \int_G \frac{\partial R}{\partial \mu}(\mu^G, M_g.\theta) d\lambda_G(g)$$

Thus, formally taking the gradient we also get:

$$D_\mu R^{FA}(\mu, \theta) = \int_G M_g^T.D_\mu R(\mu^G, M_g.\theta) d\lambda_G(g)$$

Lastly, the l.f.d. of R^{EA} is calculated similarly, noticing that:

$$\begin{aligned}\lim_{h\to 0} \frac{R^{EA}((1-h)\mu + h\nu) - R^{EA}(\mu)}{h} &= \lim_{h\to 0} \frac{R(P_{\mathcal{E}^G}\#((1-h)\mu + h\nu)) - R(P_{\mathcal{E}^G}\mu)}{h} \\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^{\mathcal{E}^G}, \theta) d(P_{\mathcal{E}^G}\#\nu - P_{\mathcal{E}^G}\#\mu)(\theta) \\ &= \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^{\mathcal{E}^G}, P_{\mathcal{E}^G}\theta) d(\nu - \mu)(\theta)\end{aligned}$$

And that:

$$\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^{\mathcal{E}^G}, P_{\mathcal{E}^G}.\theta) d\mu(\theta) = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu^{\mathcal{E}^G}, \theta) d(\mu^{P_{\mathcal{E}^G}})(\theta) = 0$$

So that we can identify:

$$\frac{\partial R^{EA}}{\partial \mu}(\mu, \theta) = \frac{\partial R}{\partial \mu}(\mu^{\mathcal{E}^G}, P_{\mathcal{E}^G}.\theta) \quad \text{and} \quad D_\mu R^{EA}(\mu, \theta) = P_{\mathcal{E}^G}^T.D_\mu R(\mu^{\mathcal{E}^G}, P_{\mathcal{E}^G}.\theta)$$

Finally, from the fact that R is **convex** and C^1, we know by proposition 32 that $\forall \mu \in \mathcal{P}(\mathcal{Z})$:

$$R^{FA}(\mu) = R(\mu^G) \leqslant \int_G R(M_g\#\mu) d\lambda_G(g) = R^G(\mu)$$

□

Proof of Proposition 43. Naturally, if $R = R^G$, from proposition 40, we get that R is G-invariant.

For the converse, whenever R is G-invariant, we have that $\forall \mu \in \mathcal{P}(\mathcal{Z}),\ \forall g \in G, R(M_g\#\mu) = R(\mu)$, so that:

$$\forall \mu \in \mathcal{P}(\mathcal{Z}),\ \ R^G(\mu) = \int_G R(M_g\#\mu) d\lambda_G(g) = \int_G R(\mu) d\lambda_G(g) = R(\mu)$$

□

Proof of Proposition 44. First of all, notice that for $\mu \in \mathcal{P}(\mathcal{Z})$, by definition:

$$\begin{aligned}|R(\mu) - R^G(\mu)| &= \left|\mathbb{E}_\pi[L^\mu(X, Y)] - \int_G \mathbb{E}_\pi[L^\mu(\rho_g.X, \hat{\rho}_g.Y)] d\lambda_G(g)\right| \\ &= \left|\int_G \mathbb{E}_\pi\left[L^\mu(X, Y) - L^\mu(\rho_g.X, \hat{\rho}_g.Y)\right] d\lambda_G(g)\right|\end{aligned}$$

For the first part, notice, as in Chen et al. [14], that by a known characterization of the norms generated by inner products, we can say that, for any $a \in \mathbb{R}$: $|a| = \sup_{|v|\leqslant 1}\langle v, a\rangle$, in particular:

$$\begin{aligned}
|R(\mu) - R^G(\mu)| &= \sup_{|v|\leqslant 1} \left\langle v, \int_G \mathbb{E}_\pi \left[L^\mu(X,Y) - L^\mu(\rho_g.X, \hat{\rho}_g.Y)\right] d\lambda_G(g) \right\rangle \\
&= \sup_{|v|\leqslant 1} \int_G \mathbb{E}_\pi \left[\langle v, L^\mu(X,Y) - L^\mu(\rho_g.X, \hat{\rho}_g.Y)\rangle\right] d\lambda_G(g) \\
&= \sup_{|v|\leqslant 1} \int_G \mathbb{E}_\pi \left[\langle v, L^\mu(X,Y)\rangle - \langle v, L^\mu(\rho_g.X, \hat{\rho}_g.Y)\rangle\right] d\lambda_G(g) \\
&= \sup_{|v|\leqslant 1} \int_G \mathbb{E}_\pi \left[\langle v, L^\mu(X,Y)\rangle\right] - \mathbb{E}_\pi \left[\langle v, L^\mu(\rho_g.X, \hat{\rho}_g.Y)\rangle\right] d\lambda_G(g)
\end{aligned}$$

Now, for any $v \in \mathbb{R}$ with $|v| \leqslant 1$, we define $\varphi_v(\cdot) = \langle v, \cdot\rangle$, which is clearly a 1-Lipschitz function (by the Cauchy-Schwartz inequality).

By the Kantorovich-Rubinstein characterization of the W_1 distance, we have, for any $\nu_1, \nu_2 \in \mathcal{P}(\mathbb{R})$: $W_1(\nu_1, \nu_2) = \sup\left\{\left|\int \varphi d\nu_1 - \int \varphi d\nu_2\right| \ : \ \varphi \text{ 1-Lipschitz}\right\}$.

So, in particular, considering $\nu_1 = L^\mu\#\pi$ and $\nu_2 = L^\mu\#g\#\pi$ we have, for any $v \in \mathbb{R}$:

$$\begin{aligned}
W_1(L^\mu\#\pi, L^\mu\#g\#\pi) &\geqslant \left|\int \varphi_v dL^\mu\#\pi - \int \varphi_v dL^\mu\#g\#\pi\right| \\
&\geqslant \left|\int (\varphi_v \circ L^\mu) d\pi - \int (\varphi_v \circ L^\mu \circ g) d\pi\right| \\
&\geqslant \mathbb{E}_\pi \left[\langle v, L^\mu(X,Y)\rangle\right] - \mathbb{E}_\pi \left[\langle v, L^\mu(\rho_g.X, \hat{\rho}_g.Y)\rangle\right]
\end{aligned}$$

We can integrate this inequality to get:

$$\int_G W_1(L^\mu\#\pi, L^\mu\#g\#\pi) d\lambda_G(g) \geqslant \int_G \left(\mathbb{E}_\pi \left[\langle v, L^\mu(X,Y)\rangle\right] - \mathbb{E}_\pi \left[\langle v, L^\mu(\rho_g.X, \hat{\rho}_g.Y)\rangle\right]\right) d\lambda_G(g)$$

As this holds for any $v \in \mathbb{R}, |v| \leqslant 1$, then:

$$|R(\mu) - R^G(\mu)| \leqslant \int_G W_1(L^\mu\#g\#\pi, L^\mu\#\pi) d\lambda_G(g)$$

Now, by a known property of the Wasserstein metric with respect to pushforward measures of Lipschitz functions[12], $\forall \mu \in \mathcal{P}(\mathcal{Z})$ and $g \in G$:

$$W_1(L^\mu\#g\#\pi, L^\mu\#\pi) \leqslant C W_1(g\#\pi, \pi)$$

so that (as the bound doesn't depend on μ):

$$\sup_{\mu \in \mathcal{P}(\mathcal{Z})} |R(\mu) - R^G(\mu)| \leqslant C \int_G W_1(g\#\pi, \pi) d\lambda_G(g)$$

[12] Which comes from noting that for γ an optimal coupling between π and $g\#\pi$, $(L^\mu, L^\mu)\#\gamma$ is a coupling between $L^\mu\#\pi$ and $L^\mu\#g\#\pi$; allowing to conclude using the C-Lipschitzness of L^μ uniformly on μ

Assuming both ℓ and σ_* to be Lipschitz (with constants C_1 and C_2 respectively) and defining $C = \max\{C_1, C_1C_2\}$, tells us that, for any $\mu \in \mathcal{P}(\mathcal{Z})$:

$$\begin{aligned}
|L^\mu(x_1, y_1) - L^\mu(x_2, y_2)| &= |\ell(\langle\sigma_*(x_1, \cdot), \mu\rangle, y_1) - \ell(\langle\sigma_*(x_2, \cdot), \mu\rangle, y_2)| \\
&\leqslant C_1 \left(\|\langle\sigma_*(x_1, \cdot), \mu\rangle - \langle\sigma_*(x_2, \cdot), \mu\rangle\| + \|y_1 - y_2\|\right) \\
&\leqslant C_1 \left\|\int (\sigma_*(x_1, \theta) - \sigma_*(x_2, \theta))\, d\mu(\theta)\right\| + C_1\|y_1 - y_2\| \\
&\leqslant C_1 \int \|(\sigma_*(x_1, \theta) - \sigma_*(x_2, \theta))\|\, d\mu(\theta) + C_1\|y_1 - y_2\| \\
&\leqslant C_1 \int C_2 \left(\|x_1 - x_2\| + \|\theta - \theta\|\right) d\mu(\theta) + C_1\|y_1 - y_2\| \\
&\leqslant C \left(\|x_1 - x_2\| + \|y_1 - y_2\|\right)
\end{aligned}$$

Implying the desired (uniform in μ) global C-Lipschitzness. □

C.16 Proofs for Section 5.3

Proof of Proposition 45. We know that a family $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ satisfies the WGF equation in the weak sense if $\forall \varphi \in C_c^\infty(\mathcal{Z} \times (0,T))$:

$$\int_0^T \int_{\mathcal{Z}} \left(\partial_t \varphi(z,t) - \langle \varsigma(t) D_\mu R(\mu_t, z), \nabla_z \varphi(z,t)\rangle\right) d\mu_t(z)\, dt = 0$$

Now, profiting from the **uniqueness** of the solutions of this equation, it will be enough to show that, given a solution $(\mu_t)_{t\geqslant 0} \subseteq \mathcal{P}_2(\mathcal{Z})$ of the **WGF**, then $(\mu_t^G)_{t\geqslant 0} \subseteq \mathcal{P}_2^G(\mathcal{Z})$ **is also a solution**. Indeed, consider, for $g \in G$, $\tilde{\mu}_t = M_g \# \mu_t$, and notice that for $\varphi \in C_c^\infty(\mathcal{Z} \times (0,T))$:

$$\int_0^T \int_{\mathcal{Z}} (\partial_t \varphi(z,t) - \langle \varsigma(t) D_\mu R(\tilde{\mu}_t, z), \nabla_z \varphi(z,t)\rangle) d\tilde{\mu}_t(z)\, dt$$
$$= \int_0^T \int_{\mathcal{Z}} \left(\partial_t \varphi(M_g.z,t) - \langle \varsigma(t) D_\mu R(M_g\#\mu_t, M_g.z), \nabla_z \varphi(M_g.z,t)\rangle\right) d\mu_t(z)\, dt =: \star$$

Now, we can define $\varphi^g \in C_c^\infty(\mathcal{Z} \times (0,T))$ given by $\forall (z,t) \in \mathcal{Z} \times (0,T)$ $\varphi^g(z,t) = \varphi(M_g.z,t)$, which satisfies:

$$\partial_t \varphi^g(z,t) = \partial_t \varphi(M_g.z,t) \ \text{ and } \ \nabla_z \varphi^g(z,t) = M_g^T \nabla_z \varphi(M_g.z,t)$$

So that, by also using proposition 31 and the orthogonality of the group action, we get:

$$\star = \int_0^T \int_{\mathcal{Z}} \left(\partial_t \varphi^g(z,t) - \langle M_g.\varsigma(t) D_\mu R(\mu_t, z), M_g \nabla_z \varphi^g(z,t)\rangle\right) d\mu_t(z)\, dt$$
$$= \int_0^T \int_{\mathcal{Z}} \left(\partial_t \varphi^g(z,t) - \langle \varsigma(t) D_\mu R(\mu_t, z), \nabla_z \varphi^g(z,t)\rangle\right) d\mu_t(z)\, dt = 0$$

Where the last equality comes from the fact that $(\mu_t)_{t\geqslant 0}$ is a solution to the **WGF**.

In particular, as we also have that $\tilde{\mu}_0 = M_g \# \mu_0 = \mu_0$ (because $\mu_0 \in \mathcal{P}^G(\mathcal{Z})$), by uniqueness we can conclude that this means that $\forall g \in G, \forall t \in (0,T)$ λ-a.e., $\mu_t = M_g \# \mu_t$.

This may seem *weaker* that what we want to prove. Nevertheless, as our group is compact and has a unique normalized Haar measure, we can proceed as follows: let $f : [0,T] \times \mathcal{Z} \to \mathbb{R}_+$ be any positive and measurable function. Given $g \in G$, take $\Omega_g \subseteq [0,T]$ a full measure set where it holds that $\mu_t = M_g \# \mu_t$. In particular, $f_t = f(t,\cdot) : \mathcal{Z} \to \mathbb{R}$ is positive and measurable, so that:

$$\forall t \in \Omega_g,\ \langle f_t, \mu_t\rangle = \langle f_t, M_g\#\mu_t\rangle = \langle f_t \circ M_g, \mu_t\rangle$$

and we can integrate this equality to get: $\int_0^T \langle f_t, \mu_t\rangle dt = \int_0^T \langle f_t \circ M_g, \mu_t\rangle dt$ Now, by integrating both sides with respect to the Haar measure, and applying Fubini's theorem (because everything is positive) we get:

$$\int_0^T \langle f_t, \mu_t\rangle dt = \int_G \int_0^T \langle f_t, \mu_t\rangle dt d\lambda_G(g) = \int_G \int_0^T \langle f_t \circ M_g, \mu_t\rangle dt d\lambda_G(g)$$
$$= \int_0^T \int_G \langle f_t \circ M_g, \mu_t\rangle d\lambda_G(g) dt = \int_0^T \langle f_t, \mu_t^G\rangle dt$$

Implying[13] $\forall t \in [0,T]$ a.e. $\mu_t = \mu_t^G$, and therefore: $\forall t \in [0,T]$ a.e. $\mu_t \in \mathcal{P}^G(\mathcal{Z})$.

□

Proof of Theorem 14. From Hu et al. [38] (or Sznitman [83]) we know that in all of the presented cases a unique solution to the Fokker-Planck equation exists.

Furthermore, from the examples in section 4.4, we know that, whenever R, r and U are G-invariant functions, the functionals R, R^τ and $R_\nu^{\tau,\beta}$ are all G-invariant. As from the examples in section 2.2.5 we know that $\forall \mu \in \mathcal{P}(\mathcal{Z})$ s.t. $\mu \lll \nu$, $\forall \theta \in \mathcal{Z}$:

$$\begin{aligned} D_\mu R_\nu^{\tau,\beta}(\mu,\theta) &= D_\mu R(\mu,\theta) + \tau \nabla_\theta r(\theta) + \beta\left(\nabla_\theta \log(\mu(\theta)) - \nabla_\theta \log(\nu(\theta))\right) \\ &= D_\mu R(\mu,\theta) + \tau \nabla_\theta r(\theta) + \beta\left(\nabla_\theta \log(\mu(\theta)) + \nabla_\theta U(\theta)\right) \\ &= D_\mu R(\mu,\theta) + \tau \nabla_\theta r(\theta) + \beta \nabla_\theta U(\theta) + \beta\left(\frac{1}{\mu(\theta)} \nabla_\theta \mu(\theta)\right) \end{aligned}$$

So that $\forall \mu \in \mathcal{P}(\mathcal{Z})$ s.t. $\mu \lll \nu$, $\forall \theta \in \mathcal{Z}$:

$$\begin{aligned} \operatorname{div}\left(D_\mu R_\nu^{\tau,\beta}(\mu,\theta)\,\mu(\theta)\right) &= \operatorname{div}\left(\left(D_\mu R(\mu,\theta) + \tau \nabla_\theta r(\theta) + \beta \nabla_\theta U(\theta) + \beta\left(\frac{1}{\mu(\theta)} \nabla_\theta \mu(\theta)\right)\right) \mu(\theta)\right) \\ &= \operatorname{div}\left(\left(D_\mu R(\mu,\theta) + \tau \nabla_\theta r(\theta) + \beta \nabla_\theta U(\theta)\right) \mu(\theta) + \beta \nabla_\theta \mu(\theta)\,\right) \\ &= \operatorname{div}\left(\left(D_\mu R(\mu,\theta) + \tau \nabla_\theta r(\theta) + \beta \nabla_\theta U(\theta)\right) \mu(\theta)\,\right) + \beta \Delta_\theta \mu(\theta) \end{aligned}$$

That is, $R_\nu^{\tau,\beta}$ is a G-invariant functional such that the WGF as given by the expression of proposition 45 reads:

$$\partial_t \mu_t = \varsigma(t)\left[\operatorname{div}\left(D_\mu R_\nu^{\tau,\beta}(\mu_t,\cdot)\,\mu_t\right)\right] = \varsigma(t)\left[\operatorname{div}\left(\left(D_\mu R(\mu_t,\cdot) + \tau \nabla_\theta r + \beta \nabla_\theta U\right) \mu_t\right) + \beta \Delta \mu_t\right]$$

and it has indeed a **unique** solution (weak if $\beta = 0$ and strong if $\beta > 0$). So, proposition 45 applies and allows us to conclude. As for the property of the *density*, we apply proposition 26 directly. □

Proof of Corollary 10. Though it is enough to reference theorem 14, we provide an independent proof, as it highlights other interesting elements from the theory developped during chapter 4.

To establish this result, we note the following classical result regarding Brownian motions:

Lemma 15 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(B_t)_{t \geqslant 0}$ be a standard D-dimensional BM. If $M \in O(D)$ is an orthogonal matrix, THEN $(W_t)_{t \geqslant 0} := (MB_t)_{t \geqslant 0}$ is also a standard D-dimensional BM.*

Proof of Lemma 15. It's not complicated to check this, noting that, according to the usual characterization of the standard BM:

[13] Indeed, one can choose any positive measurable $f : \mathcal{Z} \to \mathbb{R}$ and define the sets $A_+ = \{t \in [0,T] \ : \langle f, \mu_t \rangle > \langle f, \mu_t^G \rangle\}$ (And A_- analogously); so that $f_\pm(t,z) = 1_{A_\pm}(t) f(z)$, and therefore, the fact that: $\int_0^T 1_{A_\pm}(t) \langle f, \mu_t \rangle dt = \int_0^T 1_{A_\pm}(t) \langle f, \mu_t^G \rangle dt$ implies that both $A_\pm$ have null measure

- $W_0 = MB_0 = 0$
- For almost all $\omega \in \Omega$, the function $t \mapsto B_t(\omega)$ is continuous, which implies that $t \mapsto MB_t(\omega)$ is also continuous.
- For any $t_0 < \cdots < t_n$, the vector $(B_{t_{i+1}} - B_{t_i})_{i=0}^{n-1}$ is independent, which implies that $(MB_{t_{i+1}} - MB_{t_i})_{i=0}^{n-1}$ is also independent (as it's a measurable function of independent random variables).
- Finally, to see the stationary increments property, let $0 \leqslant s \leqslant t$. By studying the characteristic function, we notice that for any $\xi \in \mathbb{R}^D$:

  $$\begin{aligned}\mathbb{E}\left[e^{i\xi^T(W_t - W_s)}\right] = \mathbb{E}\left[e^{i\xi^T M(B_t - B_s)}\right] = \mathbb{E}\left[e^{i(M^T\xi)^T(B_t - B_s)}\right] = \mathbb{E}\left[e^{i(M^T\xi)^T(B_{t-s})}\right] \\ = \exp\left(-\frac{1}{2}(t-s)|M^T\xi|^2\right) = \exp\left(-\frac{1}{2}(t-s)|\xi|^2\right)\end{aligned}$$

  Here, we've used that $B_{t-s} \sim \mathcal{N}(0, (t-s)\mathrm{Id}_D)$ (and thus its characteristic function is known), as well as the orthogonality of M. In particular, we've shown that the characteristic function of $W_t - W_s$ is exactly that of a normal distribution $\mathcal{N}(0, (t-s)\mathrm{Id}_D)$.

With all of this, we conclude that $(W_t)_{t\geqslant 0}$ is a D-dimensional BM. □

Now, we look at the McKean-Vlasov SDE given by (where we include the *covariance term* from Bortoli et al. [9], assuming the law π is jointly G-invariant):

$$dZ_t = \varsigma(t)\left[-\left(D_\mu R(\mu_t, Z_t) + \tau\nabla_\theta r(Z_t) + \beta\nabla_\theta U(Z_t)\right)dt + \sqrt{\frac{\alpha}{B}}\sqrt{\Sigma}(\mu_t, Z_t)dB_t + \sqrt{2\beta}d\tilde{B}_t\right]$$

Where $\mu_t = \mathrm{Law}(Z_t)$, $(B_t)_{t\geqslant 0}$ and $(\tilde{B}_t)_{t\geqslant 0}$ are (independent) D-dimensional Brownian Motions, and $\Sigma(\mu, \theta) := \mathbb{E}_\pi\left[\left(D_\mu L^\tau_{X,Y}(\mu,\theta) - D_\mu R^\tau(\mu,\theta)\right) \otimes \left(D_\mu L^\tau_{X,Y}(\mu,\theta) - D_\mu R^\tau(\mu,\theta)\right)\right]$.

Let $(Z_t)_{t\geqslant 0}$ be the **unique strong solution** to this SDE (which exists by a result by Bortoli et al. [9]). By the strong solvability, we must have that:

$$Z_t \overset{a.s.}{=} Z_0 - \int_0^t \varsigma(s)\left(D_\mu R^\tau(\mu_s, Z_s) + \beta\nabla_\theta U(Z_s)\right)ds + \int_0^t \sqrt{\frac{\alpha}{B}}\varsigma(s)\sqrt{\Sigma}(\mu_s, Z_s)dB_s + \int_0^t \varsigma(s)\sqrt{2\beta}d\tilde{B}_s$$

and **Law**$(Z_0) = \mu^0$ (initial condition)

Now, in the context of this SDE, the solution is **uniquely trajectory-wise determined**; that is, if $(Y_t)_{t\geqslant 0}$ is another solution over $[0, T]$ with respect to the **same Brownian motion and in the same probability space**, then $Z_t \overset{a.s.}{=} Y_t$ for all $t \geqslant 0$. This is slightly more than what we need, as the Yamada-Watanabe theorem guarantees that **trajectory-wise uniqueness implies uniqueness in law** for an SDE. Consequently, if we can find *another solution to the SDE, potentially with respect to a different probability space and different BM, their laws must be the same.*

In other words, to successfully demonstrate that $\forall g \in G,\ \mu_t = M_g\#\mu_t$, it suffices to show that for a fixed $g \in G$, the process $\tilde{Z}_t := M_g Z_t$ satisfies the same SDE, potentially with a different BM and a different probability space. By uniqueness in law, this implies that $\forall t \geqslant 0$ (a.e.):

$$\mu_t = \mathbf{Law}(Z_t) = \mathbf{Law}(\tilde{Z}_t) = \mathbf{Law}(M_g Z_t) = M_g \# \mu_t$$

This in turn implies (integrating over $\mathbb{R}_+$ and then over G the evaluation against an arbitrary positive measurable function $f : \mathbb{R}_+ \times \mathcal{Z} \to \mathbb{R}$; see the last part of the proof of theorem 14) that $\forall t \geqslant 0$ (a.e.) $\mu_t = \mu_t^G$, and therefore, $\mu_t \in \mathcal{P}^G(\mathcal{Z})$.

With this in mind, take the solution $(Z_t)_{t\geqslant 0}$ and $g \in G$, and observe that $\tilde{Z} = (M_g.Z_t)_{t\geqslant 0}$ satisfies the SDE with the same initial condition but with respect to **another Brownian motion**. Indeed, we have:

- $\mathbf{Law}(\tilde{Z}_0) = \mathbf{Law}(M_g Z_0) = M_g \# \mu^0 = \mu^0$ (due to the assumed G-invariant initial condition).
- Denoting $\nu_s := M_g \# \mu_s$ as the law of $\tilde{Z}_s$, we have to show that for all $t > 0$:

$$\tilde{Z}_t = \tilde{Z}_0 - \int_0^t \varsigma(s) \left(D_\mu R^\tau(\nu_s, \tilde{Z}_s) + \beta \nabla_\theta U(Z_s) \right) ds + \int_0^t \sqrt{\frac{\alpha}{B}} \varsigma(s) \sqrt{\Sigma}(\nu_s, \tilde{Z}_s) dB_s + \int_0^t \varsigma(s) \sqrt{2\beta} d\tilde{B}_s$$

Specifically, for a given $t > 0$:

$$\begin{aligned}
\tilde{Z}_t &= M_g Z_t \\
&= M_g Z_0 - M_g \left(\int_0^t \varsigma(s) \left(D_\mu R^\tau(\mu_s, Z_s) + \beta \nabla_\theta U(Z_s) \right) ds \right) \\
&\quad + M_g \left(\int_0^t \sqrt{\frac{\alpha}{B}} \varsigma(s) \sqrt{\Sigma}(Z_s, \mu_s) dB_s + \int_0^t \varsigma(s) \sqrt{2\beta} d\tilde{B}_s \right) \\
&= \tilde{Z}_0 - \int_0^t \varsigma(s) \left(M_g D_\mu R^\tau(\mu_s, Z_s) + \beta M_g \nabla_\theta U(Z_s) \right) ds \\
&\quad + \sqrt{\frac{\alpha}{B}} M_g \int_0^t \varsigma(s) \sqrt{\Sigma}(Z_s, \mu_s) dB_s + \sqrt{2\beta} M_g \int_0^t \varsigma(s) d\tilde{B}_s \\
&= \tilde{Z}_0 - \int_0^t \varsigma(s) \left(D_\mu R^\tau(\nu_s, \tilde{Z}_s) + \beta \nabla_\theta U(\tilde{Z}_s) \right) ds \\
&\quad + \sqrt{\frac{\alpha}{B}} \int_0^t \varsigma(s) M_g \sqrt{\Sigma}(Z_s, \mu_s) dB_s + \sqrt{2\beta} \int_0^t \varsigma(s) M_g d\tilde{B}_s
\end{aligned}$$

Here, we used the linearity of the integral (and stochastic integral) and proposition 31 as well as proposition 29, which hold $\forall \theta \in \mathcal{Z}, \forall \mu \in \mathcal{P}(\mathcal{Z})$ (including $\theta = Z_s(\omega)$, $\forall \omega \in \Omega$ and $\mu_s = \mathbf{Law}(Z_s)$). Furthermore, we can observe that by the same propositions (applied to random variables and their laws) we have:

$$\begin{aligned}
A(X, Y, \nu_s, \tilde{Z}_s) &:= D_\mu L^\tau_{X,Y}(\nu_s, \tilde{Z}_s) - D_\mu R^\tau(\nu_s, \tilde{Z}_s \\
&= D_\mu L^\tau_{X,Y}(M_g \# \mu_s, M_g.Z_s) - D_\mu R^\tau(M_g \# \mu_s, M_g.Z_s) \\
&= M_g D_\mu L^\tau_{\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y}(\mu_s, Z_s) - M_g.D_\mu R^\tau(\mu_s, Z_s) \\
&= M_g \left(D_\mu L^\tau_{\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y}(\mu_s, Z_s) - D_\mu R^\tau(\mu_s, Z_s) \right) \\
&= M_g.A(\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y, \mu_s, Z_s)
\end{aligned}$$

$$
\begin{aligned}
\Sigma(\nu_s, \tilde{Z}_s) &= \mathbb{E}_\pi \left[\Big(A(X, Y, \nu_s, \tilde{Z}_s) \Big) \otimes \Big(A(X, Y, \nu_s, \tilde{Z}_s) \Big) \right] \\
&= \mathbb{E}_\pi \left[\big(M_g.A(\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y, \mu_s, Z_s) \big) \otimes \big(M_g.A(\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y, \mu_s, Z_s) \big) \right] \\
&= M_g.\mathbb{E}_\pi \left[\big(A(\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y, \mu_s, Z_s) \big) \otimes \big(A(\rho_g^{-1}.X, \hat{\rho}_g^{-1}.Y, \mu_s, Z_s) \big) \right] .M_g^T \\
&= M_g.\mathbb{E}_\pi \left[(A(X, Y, \mu_s, Z_s)) \otimes (A(X, Y, \mu_s, Z_s)) \right] .M_g^T \\
&= M_g.\Sigma(\mu_s, Z_s).M_g^T
\end{aligned}
$$

Where we've used the G-invariance of π and the definition of the *outer product* $\otimes$. In short, we've proven that:

$$
\Sigma(\nu_s, \tilde{Z}_s) = M_g \Sigma(\mu_s, Z_s) M_g^T
$$

In particular, as the *square root of a symmetric positive definite matrix is unique*, it must hold that:

$$
\begin{gathered}
\sqrt{\Sigma}(\nu_s, \tilde{Z}_s) = M_g \sqrt{\Sigma}(\mu_s, Z_s) M_g^T \\
\Rightarrow M_g \sqrt{\Sigma}(\mu_s, Z_s) = \sqrt{\Sigma}(\nu_s, \tilde{Z}_s) M_g
\end{gathered}
$$

With this, we have[14]:

$$
\begin{aligned}
\tilde{Z}_t = \tilde{Z}_0 &- \int_0^t \varsigma(s) \Big(D_\mu R^\tau(\nu_s, \tilde{Z}_s) + \beta \nabla_\theta U(\tilde{Z}_s) \Big) ds \\
&+ \sqrt{\frac{\alpha}{B}} \int_0^t \varsigma(s) \sqrt{\Sigma}(\nu_s, \tilde{Z}_s) M_g dB_s + \sqrt{2\beta} \int_0^t \varsigma(s) M_g d\tilde{B}_s \\
= \tilde{Z}_0 &- \int_0^t \varsigma(s) \Big(D_\mu R^\tau(\nu_s, \tilde{Z}_s) + \beta \nabla_\theta U(\tilde{Z}_s) \Big) ds \\
&+ \sqrt{\frac{\alpha}{B}} \int_0^t \varsigma(s) \sqrt{\Sigma}(\tilde{Z}_s, \nu_s) d M_g B_s + \sqrt{2\beta} \int_0^t \varsigma(s) d M_g \tilde{B}_s
\end{aligned}
$$

Thus, $(\tilde{Z}_t)_{t \geqslant 0}$ is a solution to the SDE (with the same initial condition) but with the D-dimensional BMs $(W_t)_{t \geqslant 0} := (M_g B_t)_{t \geqslant 0}$ and $(\tilde{W}_t)_{t \geqslant 0} := (M_g \tilde{B}_t)_{t \geqslant 0}$. As previously mentioned, due to uniqueness in law (guaranteed by the Yamada-Watanabe theorem), this implies that $\tilde{Z}_t \overset{(d)}{=} Z_t$ for all $t \geqslant 0$, and consequently, $\mu_t = M_g \# \mu_t$, $\forall t \geqslant 0$, which concludes the proof.

□

Proof of Theorem 15 and Theorem 16. We will prove theorem 16, since the proof of theorem 15 shall be recovered by formally considering $\beta = 0$.

Consider the (**pathwise unique**) solution of the SDE (5.6), $Z = (Z_t)_{t \geqslant 0}$. This means that it satisfies:

$$
Z_t \overset{a.s.}{=} Z_0 - \int_0^t \varsigma(s) D_\mu R^\tau(\mu_s, Z_s) ds + \sqrt{2\beta} \int_0^t \varsigma(s) P_{\mathcal{E}^G} dB_s, \text{ and } Z_0 = \xi_0 \text{ (initial condition)} \tag{C.3}
$$

[14] We have used that $\int_0^t X(s) dW_s \overset{a.s.}{=} \int_0^t X(s) M_g dB_s$, which follows from the definition of the BM as a limit of increments.

With ξ_0 such that $\text{Law}(\xi_0) = \mu^0$.

We first let $g \in G$ be an arbitrary group element, and we study how the process $\tilde{Z} = (\tilde{Z}_t)_{t \geqslant 0} := (M_g Z_t)_{t \geqslant 0}$ satisfies this same equation (C.3).

Denote $\nu_s := M_g \# \mu_s$ as the law of $\tilde{Z}_s$ (see Lemma 7), we want to show that for all $t \geqslant 0$:

$$\tilde{Z}_t \overset{a.s.}{=} \tilde{Z}_0 - \int_0^t \varsigma(s) D_\mu R^\tau(\nu_s, \tilde{Z}_s) ds + \sqrt{2\beta} \int_0^t \varsigma(s) P_{\mathcal{E}^G} dB_s \tag{C.4}$$

Indeed, first notice that:

1. Let Ω be the full measure set where $\xi_0 \in \mathcal{E}^G$ (which we can do since $\mu_0 \in \mathcal{P}(\mathcal{E}^G)$, or, equivalently: $\mathbb{P}(\xi_0 \in \mathcal{E}^G) = 1$). Then, $\forall \omega \in \Omega,\ Z_0(\omega) = \xi_0(\omega) \in \mathcal{E}^G$. In particular, $\forall \omega \in \Omega,\ \forall g \in G, \tilde{Z}_0(\omega) = M_g Z_0(\omega) = M_g \xi_0(\omega) = \xi_0(\omega) = Z_0(\omega)$. That is, $\tilde{Z}_0 \overset{a.s.}{=} Z_0$.
2. Now, the equation is satisfied by $(Z_t)_{t \geqslant 0}$ and therefore, for $t \geqslant 0$, we have:

$$\begin{aligned}
\tilde{Z}_t = M_g Z_t = M_g Z_0 - M_g \left(\int_0^t \varsigma(s) D_\mu R^\tau(\mu_s, Z_s) ds \right) + \sqrt{2\beta} M_g \int_0^t \varsigma(s) P_{\mathcal{E}^G} dB_s \\
= \tilde{Z}_0 - \int_0^t \varsigma(s) M_g D_\mu R^\tau(\mu_s, Z_s)) ds + \sqrt{2\beta} \int_0^t \varsigma(s) M_g . P_{\mathcal{E}^G} dB_s \\
= \tilde{Z}_0 - \int_0^t \varsigma(s) D_\mu R^\tau(M_g \# \mu_s, M_g . Z_s) ds + \sqrt{2\beta} \int_0^t \varsigma(s) P_{\mathcal{E}^G} dB_s \\
= \tilde{Z}_0 - \int_0^t \varsigma(s) D_\mu R^\tau(\nu_s, \tilde{Z}_s) ds + \sqrt{2\beta} \int_0^t \varsigma(s) P_{\mathcal{E}^G} dB_s
\end{aligned}$$

   Here, we used the linearity of the integral (and the stochastic integral), the fact that $\forall g \in G,\ M_g P_{\mathcal{E}^G} = P_{\mathcal{E}^G}$, and Proposition 31 (as R is G-invariant, as well as $r : \mathcal{X} \to \mathbb{R}$), which holds for $\forall \theta \in \mathcal{Z}, \forall \mu \in \mathcal{P}(\mathcal{Z})$ (in particular for $\theta = Z_s(\omega),\ \forall \omega \in \Omega$ and $\mu_s = \text{Law}(Z_s)$). Thus, $\forall g \in G$, (C.4) holds.

By the **pathwise uniqueness** of the solution $(Z_t)_{t \geqslant 0}$, we (5.6) we have[15]:

$$\mathbb{P}\left(\sup_{t \geqslant 0} \| Z_t - \tilde{Z}_t \| = 0 \right) = 1$$

In particular, as $g \in G$ was arbitrary, we have that:

$$\forall g \in G,\ \sup_{t \geqslant 0} \| Z_t - M_g Z_t \| \overset{a.s.}{=} 0 \tag{C.5}$$

We now want to be able to **interchange** the $\forall g \in G$ with the probability measure. Fortunately, we are dealing with a compact group with a normalized Haar measure λ_G.

Indeed, from equation (C.5) we deduce that $\forall g \in G,\ \forall t \geqslant 0,\ \mathbb{P}(\| Z_t - M_g Z_t \| = 0) = 1$.

[15] More information shall be found in Ethier and Kurtz [29]

Now, notice that, for any $t \geqslant 0$ and $\omega \in \Omega$:

$$\begin{aligned}
\|Z_t(\omega) - P_{\mathcal{E}^G} Z_t(\omega)\| &= \left\| Z_t(\omega) - \int_G M_g.Z_t(\omega) d\lambda_G(g) \right\| \\
&= \left\| \int_G (Z_t(\omega) - M_g.Z_t(\omega)) d\lambda_G(g) \right\| \\
\textit{by property of the integral } &\leqslant \int_G \|Z_t(\omega) - M_g.Z_t(\omega)\| \, d\lambda_G(g)
\end{aligned}$$

Now, we can integrate both sides by $\mathbb{P}$ to get (using Fubini as functions are positive and measurable):

$$\begin{aligned}
0 \leqslant \int_\Omega \|Z_t(\omega) - P_{\mathcal{E}^G} Z_t(\omega)\| d\mathbb{P}(\omega) &\leqslant \int_\Omega \int_G \|Z_t(\omega) - M_g.Z_t(\omega)\| \, d\lambda_G(g) d\mathbb{P}(\omega) \\
&\leqslant \int_G \int_\Omega \|Z_t(\omega) - M_g.Z_t(\omega)\| \, d\mathbb{P}(\omega) d\lambda_G(g) = 0
\end{aligned}$$

where in the last step we have used the fact that $\forall g \in G,\ \forall t \geqslant 0,\ \mathbb{P}(\|Z_t - M_g Z_t\| = 0) = 1$, so that $\forall t \geqslant 0,\ \forall g \in G,\ \int_\Omega \|Z_t(\omega) - M_g.Z_t(\omega)\| \, d\mathbb{P}(\omega) = 0$.

This implies that $\forall t \geqslant 0$ $\mathbb{P}$-a.s. $Z_t = P_{\mathcal{E}^G} Z_t$, i.e. $\mathbb{P}(Z_t \in \mathcal{E}^G) = \mu_t(\mathcal{E}^G) = 1$, or, in fancier words, $\forall t \geqslant 0,\ \mu_t \in \mathcal{P}(\mathcal{E}^G)$. □

Proof of the equivalent condition resulting from Theorem 15. As for all $t \geqslant 0$ we get: $1 = \mathbb{P}\left(\forall \tilde{t} \geqslant 0, Z_{\tilde{t}} \in \mathcal{E}^G\right) \leqslant \mathbb{P}\left(Z_t \in \mathcal{E}^G\right) =: \mu_t(\mathcal{E}^G) \leqslant 1$, we get the first implication.

For the reverse implication, assume that for all $t \geqslant 0$ we have $\mu_t(\mathcal{E}^G) = 1$. i.e. if we define $\Omega_t := \{Z_t \in \mathcal{E}^G\}$, we have $\mathbb{P}(\Omega_t) = 1,\ \forall t \geqslant 0$. It will suffice to show that

$$\bigcap_{t \in \mathbb{Q}_+} \Omega_t = \bigcap_{t \geqslant 0} \Omega_t$$

as we will then have: $\mathbb{P}\left(\bigcap_{t \in \mathbb{Q}_+} \Omega_t\right) = 1$ as a countable intersection of almost sure events. This will, in turn, give us that:

$$\mathbb{P}\left(\forall t \in [0, T], Z_t \in \mathcal{E}^G\right) = 1$$

To show our claim, the $\supseteq$ inclusion is direct; the other inclusion follows from the **continuity of the solution of the SDE**[16]. Let $\omega \in \bigcap_{t \in \mathbb{Q}_+} \Omega_t$, then $\forall \tilde{t} \in \mathbb{Q}_+,\ Z_{\tilde{t}}(\omega) \in \mathcal{E}^G$. Let $t \geqslant 0$. we know that $\exists (t_n)_{n \in \mathbb{N}} \subseteq \mathbb{Q}_+$ such that $t_n \xrightarrow[n \to \infty]{} t$; by the continuity of $(Z_t(\omega))_{t \geqslant 0}$, we get that $\lim_{n \to \infty} \|Z_{t_n}(\omega) - Z_t(\omega)\| = 0$. As $\forall n \in \mathbb{N},\ Z_{t_n}(\omega) \in \mathcal{E}^G$, and $\mathcal{E}^G$ is closed[17], we get that $Z_t(\omega) \in \mathcal{E}^G$. i.e. we have proven that $\forall t \geqslant 0,\ Z_t(\omega) \in \mathcal{E}^G$, which means that $\omega \in \bigcap_{t \geqslant 0} \Omega_t$. □

[16] This shouldn't be hard to verify, as $\forall \omega \in \Omega,\ (Z_t(\omega))_{t \geqslant 0}$ just follows an equation driven by $(B_t(\omega))_{t \geqslant 0}$, which has continuous trajectories.

[17] $\mathcal{E}^G = \bigcap_{g \in G} \{\theta \in \mathbb{R}^D \ : \ (M_g - I)\theta = 0\} = \bigcap_{g \in G} (M_g - \mathrm{Id}_D)^{-1}(\{0\})$ which is an intersection of closed sets (thanks to the continuity of $M_g,\ \forall g \in G$)

Proof of Corollary 13 and Corollary 14. Let $(\mu_t^{FA})_{t\geqslant 0}$ and $(\mu_t)_{t\geqslant 0}$ be the WGF solutions starting from μ_0 for R^{FA} and R respectively. Then, as R^{FA} is G-invariant (see proposition 40), by theorem 14, $(a.e.)\forall t \geqslant 0,\ \mu_t^{FA} \in \mathcal{P}^G(\mathcal{Z})$. Now, let's see that this process actually **also satisfies** the WGF for R, forcing both processes to coincide by the uniqueness of the WGF.

Indeed, we know that $(\mu_t^{FA})_{t\geqslant 0}$ satisfies: $\forall \varphi \in C_c^\infty(\mathcal{Z} \times (0,T))$:

$$\int_0^T \int_{\mathcal{Z}} \big(\partial_t \varphi(z,t) - \langle \varsigma(t) D_\mu R^{FA}(\mu_t^{FA}, z), \nabla_z \varphi(z,t)\rangle\big)\, d\mu_t^{FA}(z)\, dt = 0$$

Now, as: $(a.e.)\forall t \geqslant 0,\ \mu_t^{FA} \in \mathcal{P}^G(\mathcal{Z})$ (and thus, $\mu_t^{FA} = (\mu_t^{FA})^G$), we have that, $\forall z \in \mathcal{Z}$:

$$D_\mu R^{FA}(\mu_t^{FA}, z) = \int_G M_g^T . D_\mu R((\mu_t^{FA})^G, M_g.z) d\lambda_G(g) = \int_G M_g^T . D_\mu R(\mu_t^{FA}, M_g.z) d\lambda_G(g)$$

Where we've expanded the expression of $D_\mu R^{FA}$ using proposition 41. Now, as R is supposed to be G-invariant, we can use proposition 31 to see that (once again using that μ_t^{FA} is G-invariant):

$$D_\mu R(\mu_t^{FA}, M_g.z) = M_g . D_\mu R(M_g^{-1}.\mu_t^{FA}, z) = M_g . D_\mu R(\mu_t^{FA}, z)$$

Putting it all together, we have:

$$D_\mu R^{FA}(\mu_t^{FA}, z) = \int_G M_g^T . M_g . D_\mu R(\mu_t^{FA}, z) d\lambda_G(g) = D_\mu R(\mu_t^{FA}, z)$$

So, in particular, $(\mu_t^{FA})_{t\geqslant 0}$ satisfies $\forall \varphi \in C_c^\infty(\mathcal{Z} \times (0,T))$:

$$\int_0^T \int_{\mathcal{Z}} \big(\partial_t \varphi(z,t) - \langle \varsigma(t) D_\mu R(\mu_t^{FA}, z), \nabla_z \varphi(z,t)\rangle\big)\, d\mu_t^{FA}(z)\, dt = 0$$

Implying that $(\mu_t^{FA})_{t\geqslant 0}$ is a solution of the WGF of R starting from μ_0. By uniqueness, we get: $(\mu_t^{FA})_{t\geqslant 0} = (\mu_t)_{t\geqslant 0}$.

The extension to the case of corollary 14 comes from the fact that, even when R isn't G-invariant, we have from proposition 41 (and also using the G-invariance of μ_t^{FA}), $\forall z \in \mathcal{Z}$:

$$D_\mu R^{FA}(\mu_t^{FA}, z) = \int_G M_g^T . D_\mu R((\mu_t^{FA})^G, M_g.z) d\lambda_G(g) = \int_G M_g^T . D_\mu R(\mu_t^{FA}, M_g.z) d\lambda_G(g)$$

$$D_\mu R^G(\mu_t^{FA}, z) = \int_G M_g^T . D_\mu R(M_g \# \mu_t^{FA}, M_g.z) d\lambda_G(g) = \int_G M_g^T . D_\mu R(\mu_t^{FA}, M_g.z) d\lambda_G(g)$$

i.e. $D_\mu R^{FA}(\mu_t^{FA}, z) = D_\mu R^G(\mu_t^{FA}, z),\ \forall z \in \mathcal{Z}$, allowing us to conclude as before. □

Annex D

Reference Results from the Literature

D.1 Mean Field Theory of Shallow NNs

D.1.1 Key Assumptions for the results of chapter 2

Here we present some of the *standard* assumptions from the literature for the different results presented throughout our work.

Assumption 7 (Assumptions for Proposition 7 (taken from Chizat and Bach [16])) *Consider a setting as described in proposition 6, with* $R(\mu) := L(\langle \Phi, \mu \rangle) + \tau \int_{\mathcal{Z}} r d\mu$.

1. *Let* $\mathcal{Z}$ *to be the closure of a convex open set within some finite-dimensional euclidean space.*
2. *Let* $L : \mathcal{H} \to \mathbb{R}^+$ *be differentiable, with a differential* dL *that is* ***Lipschitz on bounded sets and bounded on sublevel sets****.*
3. *Let* $\Phi : \mathcal{Z} \to \mathcal{H}$ *be differentiable and* $V : \mathcal{Z} \to \mathbb{R}^+$ *be* ***semiconvex*** *(i.e.* $\exists \lambda \in \mathbb{R} : V + \lambda |\cdot|^2$ *is convex).*
4. *There exists a family* $(Q_r)_{r>0}$ *of* ***nested nonempty closed convex subsets*** *of* $\mathcal{Z}$ *such that:*
   (a) $\{u \in \Omega; dist(u, Q_{r'}) \leqslant r\} \subset Q_{r+r'}$ *for all* $r, r' > 0$,
   (b) Φ *and* V *are bounded, and* $d\Phi$ *is Lipschitz on each* Q_r
   (c) $\exists C_1, C_2 > 0$ *such that* $\sup_{u \in Q_r} (\|d\Phi(u)\| + \|\partial V(u)\|) \leqslant C_1 + C_2 r$ *for all* $r > 0$*, where* $\|\partial V(u)\|$ *stands for the maximal norm of an element in* $\partial V(u)$.

Assumption 8 (Assumptions for the POC result, Theorem 4) *We state some of the common assumptions for which results are known:*

1. *Sirignano and Spiliopoulos [78]. Let* $\mathcal{Z} = \mathbb{R} \times \mathbb{R}^d$ *with* $\sigma_*(x, (w, a)) = w\sigma(a^T x)$.

- *$\sigma \in C_b^2(\mathbb{R})$ i.e. σ is twice continuously differentiable and bounded.*
- *The data is i.i.d. distributed according to $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that $\mathbb{E}_\pi[\|X\|^4 + |Y|^4]$ is bounded.*
- *The parameters are initialized i.i.d. drawn from a distribution $\mu_0 \in \mathcal{P}(\mathcal{Z})$ is such that: $\mathbb{E}_{(W,A)\sim\mu_0}[\exp(q|W|)] < C$ for some $0 < q < \infty$ and $\mathbb{E}_{(W,A)\sim\mu_0}[\|A\|^4)] < C$*

2. *Mei et al. [57]. Let $\mathcal{Z} = \mathbb{R} \times \mathbb{R}^d$.*

   - *The activation function $(x, \theta) \mapsto \sigma_*(x; \theta)$ is **bounded**, with **sub-Gaussian gradient**: $\|\sigma_*\|_\infty \leqslant C_2$, $\|\nabla_\theta \sigma_*(X; \theta)\|_{\psi_2} \leqslant C_2$. **Labels are bounded** $|y_k| \leqslant C_2$.*
   - *The gradients $\theta \mapsto \nabla F(\theta)$, $(\theta_1, \theta_2) \mapsto \nabla_{\theta_1} K(\theta_1, \theta_2)$ are **bounded**, Lipschitz continuous; namely $\|\nabla_\theta F(\theta)\|_2, \|\nabla_{\theta_1} K(\theta_1, \theta_2)\|_2 \leqslant C_3$, $\|\nabla_\theta F(\theta) - \nabla_\theta F(\theta')\|_2 \leqslant C_3\|\theta - \theta'\|_2$, $\|\nabla_{\theta_1} K(\theta_1, \theta_2) - \nabla_{\theta_1} K(\theta_1', \theta_2')\|_2 \leqslant C_3\|(\theta_1, \theta_2) - (\theta_1', \theta_2')\|_2$.*
   - *The initialization is i.i.d. drawn from $\mu_0 \in \mathcal{P}(\mathcal{Z})$.*
   - *In the general case, assume the function $t \mapsto \varsigma(t)$ to be bounded Lipschitz: $\|\varsigma\|_\infty, \|\varsigma\|_{Lip} \leqslant C_1$, with $\int_0^\infty \varsigma(t)\,dt = \infty$. Also, let $\varepsilon = \varepsilon_N$ such that $\lim_{N\to\infty} \varepsilon_N = \lim_{N\to\infty} \varepsilon_N \log(N/\varepsilon_N) = 0$ and $\lim_{N\to\infty} N/\log(N/\varepsilon_N) = \infty$*

   Only for the Noisy case, assume:

   - *$F \in C^4(\mathcal{Z})$, $K \in C^4(\mathcal{Z} \times \mathcal{Z})$ with $\nabla_{\theta_i}^k K(\theta_1, \theta_2)$ uniformly bounded for $0 \leqslant k \leqslant 4$.*

3. *Descours et al. [24]. Consider, for $N \geqslant 1$, the σ-algebras, $\mathcal{F}_N^0 = \sigma\{Z_i^0, i = 1, \ldots, N\}$ and, for $k \geqslant 1$, $\mathcal{F}_N^k = \sigma\left(Z_i^0, \{B_j^{k-1}\}_{j=0}, \{\varepsilon_{i,j}^{k-1}\}_{j=0}, i \in \{1, \ldots, N\}\right)$. Assume:*

   - *For all $k, q \in \mathbb{N}$, $|B_q| \perp\!\!\!\perp (X_n^k, Y_n^k)_{n\geqslant 1}$. Also, $\forall k \in \mathbb{N}$, $|B_k|, (X_n^k, Y_n^k)_{n\geqslant 1} \perp\!\!\!\perp \mathcal{F}_N^k$*
   - *The activation function $\sigma^* : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ belongs to $C_b^\infty(\mathcal{X} \times \mathcal{Z})$.*
   - *For all $\ell \neq k \in \mathbb{N}$, $(X_n^\ell, Y_n^\ell)_{n\geqslant 1} \perp\!\!\!\perp (X_n^k, Y_n^k)_{n\geqslant 1}$. Also, $\forall k \in \mathbb{N}$, the batch of data is drawn i.i.d. from $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ s.t. $\mathbb{E}_\pi[|Y|^{16\gamma^*}]$ is finite.*
   - *Parameters are initialized i.i.d. from $\mu_0 \in \mathcal{P}(\mathcal{Z})$ such that $\mathbb{E}_{\mu_0}[|Z|^{8\gamma^*}] < +\infty$.*
   - *$\forall k \in \mathbb{N}$, $\forall i \in \{1, \ldots, N\}$, $\xi_i^k \sim \mathcal{N}(0, \mathrm{Id}_{\mathcal{Z}})$ and $\xi_i^k \perp\!\!\!\perp \mathcal{F}_N^k$. Also, $\forall k, l \in \mathbb{N}$, $\forall i, j \in \{1, \ldots, N\}$ s.t. $(i, k) \neq (j, l)$, $\xi_i^k \perp\!\!\!\perp \xi_j^l$.*

4. *Bortoli et al. [9]. Let there exist measurable functions $\Phi : \mathcal{X} \to [1, +\infty)$ and $\Psi : \mathcal{Y} \to [1, +\infty)$ such that:*

   - *$\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$ is s.t. $\forall y \in \mathcal{Y}$, $(\tilde{y} \mapsto \ell(\tilde{y}, y)) \in C^3(\mathbb{R})$, and $\forall \hat{y} \in \mathbb{R}$, $\forall y \in \mathcal{Y}$ $|\partial_1 \ell(0, y)| \leqslant \Psi(y)$, $|\partial_1^2 \ell(\hat{y}, y)| + |\partial_1^3 \ell(\hat{y}, y)| \leqslant \Psi(y)$.*
   - *$\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is such that $\forall x \in \mathcal{X}$, $(\theta \mapsto \sigma_*(x, \theta)) \in C^3(\mathcal{Z})$ and*

     $$\|\sigma_*(x, \theta)\| + \|D_\theta^1 \sigma_*(x, \theta)\| + \|D_\theta^2 \sigma(x, \theta) + \|D_\theta^3 \sigma_*(x, \theta)\| \leqslant \Phi(x),$$

   - *$r \in C^3(\mathcal{Z})$ satisfies $\sup_{\theta\in\mathcal{Z}}\{\|D_\theta^2 r(\theta)\| + \|D_\theta^3 r(\theta)\|\} < +\infty$*
   - *The data distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is s.t. $\mathbb{E}_\pi[\Phi^{10}(X) + \Psi^4(Y)] < \infty$*

Assumption 9 (Assumptions for theorem 5 (taken from Chizat and Bach [16])) *Consider the following assumptions:*

1. *Chizat and Bach [16]. Assume their* 1*-partially homogeneous case, with* $\mathcal{Z} = \mathbb{R} \times \tilde{\mathcal{Z}}$ *with* $\tilde{\mathcal{Z}} \subseteq \mathbb{R}^{D-1}$ *and* $\forall (w, \theta) \in \mathcal{Z}$, $\sigma_*(w, \theta) = w\tilde{\sigma_*}(\theta) := [x \in \mathcal{X} \mapsto w\tilde{\sigma_*}(x; \theta)] \in \mathcal{H} := L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ *and* $V(w, \theta) = |w|\tilde{V}(\theta)$. *Let* $L(f) = \mathbb{E}_\pi [\ell(f(X), Y)]$ *defined over* $L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ *and see* R *as* $R(\mu) = L(\langle \sigma_*, \mu \rangle)$.

   - *Let* $\tilde{\sigma}$ *and* $\tilde{V}$ *be bounded and* ***differentiable****, with Lipschitz differential.*
   - L *is convex, differentiable with differential* dL *that is* ***Lipschitz on bounded sets and bounded on sublevel sets****.*
   - $\forall h \in \mathcal{H}$ *(it is enough for it to hold for* $h = L'(\int \sigma_* d\mu)$ *with* $\mu \in \mathcal{M}(\mathcal{Z})$*), the set of* ***regular values*** *of* $g_h : \theta \in \tilde{\mathcal{Z}} \mapsto \langle h, \tilde{\sigma_*}(\theta) \rangle + \tilde{V}(\theta)$ *is* ***dense in its range****.*
   - $\tilde{\sigma_*}$ *behaves nicely at the boundary of the domain. i.e. Either:*
     - $\tilde{\mathcal{Z}} = \mathbb{R}^{D-1}$ *and* $\forall h \in \mathcal{H}$, $\theta \in \mathbb{S}^{D-2} \mapsto g_h(r\theta)$***converges uniformly*** *in* $\mathcal{C}^1(\mathbb{S}^{D-2})$ *as* $r \to \infty$*, to a function satisfying the Sard-type regularity.*
     - $\tilde{\mathcal{Z}}$*is the* ***closure of an bounded open convex set*** *and for all* $h \in \mathcal{H}$, g_h *satisfies* ***Neumann boundary conditions*** *(i.e., for all* $\theta \in \partial\tilde{\mathcal{Z}}$, $d(g_h)_\theta(\vec{n}_\theta) = 0$*, where* $\vec{n}_\theta \in \mathbb{R}^{D-1}$ *is the* ***normal*** *to* $\partial\tilde{\mathcal{Z}}$*at* θ*).*
   - *The initial condition* μ_0 *satisfies that: for some* $r_0 > 0$*, the support of* μ_0 *is contained in* $[-r_0, r_0] \times \tilde{\mathcal{Z}}$ *and separates*[1] $(-r_0) \times \tilde{\mathcal{Z}}$ *from* $(r_0) \times \tilde{\mathcal{Z}}$ *(this is referred to as a separation property of the initial condition.).*

Assumption 10 (Assumptions for the well definedness of Hu et al. [38] and Chen et al. [13]) *Assume that the intrinsic derivative* $D_\mu R : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathcal{Z}$ *of the functional* $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ *exists and satisfies the following conditions:*

1. *Hu et al. [38]. Assume:*

   - $D_\mu R$ *is bounded and Lipschitz continuous, i.e.* $\exists C_R > 0$ *s.t.* $\forall z, z' \in \mathcal{Z}$, $\forall \mu, \mu' \in \mathcal{P}_2(\mathcal{Z})$,
   $$|D_\mu R(\mu, z) - D_\mu R(\mu', z')| \leqslant C_R[|z - z'| + W_2(\mu, \mu')]$$
   - $\forall \mu \in \mathcal{P}(\mathcal{Z})$, $D_\mu R(\mu, \cdot) \in C^\infty(\mathcal{Z})$.
   - $\nabla D_\mu R : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathcal{Z} \times \mathcal{Z}$ *is jointly continuous.*

2. *Chen et al. [13]. They relax some differentiability conditions at the cost of boundedness assumptions:*

   - $\forall x \in \mathcal{Z}, \forall m, m' \in \mathcal{P}_2(\mathcal{Z}), |D_\mu R(m, x) - D_\mu R(m', x)| \leqslant M^R_{mm} W_1(m, m')$ *for some constant* $M^R_{mm} \geqslant 0$ *(i.e. it is lipschitz on the measure argument).*
   - *Suppose that*
   $$\sup_{\mu \in \mathcal{P}_2(\mathcal{Z})} \sup_{x \in \mathcal{Z}} |\nabla D_\mu R(\mu, x)| \leqslant M^R_{mx}$$
   for some constant $M^R_{mx} \geqslant 0$ *i.e.* $\nabla D_\mu R(\mu, x)$ *is uniformly bounded.* [2]

[1] In an ambient space $\mathcal{Z}$, the set C *separates* the sets A and B if any continuous path in $\mathcal{Z}$ with endpoints in A and B intersects C.

[2] They state that they are NOT requiring a coercivity condition, as is the case in Hu et al. [38] with the assumption 3

Some *less relevant* assumptions, specific to the result by Suzuki et al. [82] are:

Assumption 11 (from Suzuki et al. [82]) *The loss function R and the regularization term r are convex. Specifically,*

1. *$R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ is a convex functional on $\mathcal{P}(\mathcal{Z})$, that is, $R(\theta\mu + (1-\theta)\nu) \leqslant \theta R(\mu) + (1-\theta)R(\nu)$ for any $\theta \in [0,1]$ and $\mu, \nu \in \mathcal{P}(\mathcal{Z})$.*
2. *Moreover, R admits a first-variation at any $\mu \in \mathcal{P}_2(\mathcal{Z})$.*
3. *$r(\cdot)$ is **twice differentiable and convex**.*
4. *There exist constants $\lambda_1, \lambda_2 > 0$ and $c_r > 0$ such that $\lambda_1 \mathrm{Id}_{\mathcal{Z}} \leqslant \nabla\nabla^\top r(x) \leqslant \lambda_2 \mathrm{Id}_{\mathcal{Z}}$ (in the matrix order), $x^\top \nabla r(x) \geqslant \lambda_1 \|x\|^2$, and $0 \leqslant r(x) \leqslant \lambda_2(c_r + \|x\|^2)$ for any $x \in \mathcal{Z}$, and $\nabla r(0) = 0$.*

Assumption 12 (from Suzuki et al. [82]) *Assume that:*

1. *There exists $L > 0$ such that*
$$\|D_\mu R(\mu, x) - D_\mu R(\mu', x')\| \leqslant L(W_2(\mu, \mu') + \|x - x'\|)$$
(i.e. the intrinsic derivative is Lipschitz), and
$$\left\| \frac{\partial^2 R}{\partial \mu^2}(\mu, x, x') \right\| \leqslant L(1 + c_L(\|x\|^2 + \|x'\|^2))$$
for any $\mu, \mu' \in \mathcal{P}_2(\mathcal{Z})$ and $x, x' \in \mathcal{Z}$.
2. *There exists $C > 0$ such that $\|D_\mu R(\mu, x)\| \leqslant C$ for any $\mu \in \mathcal{P}(\mathcal{Z})$ and $x \in \mathcal{Z}$. i.e. the intrinsic derivative is **bounded.***

Assumption 13 (from Suzuki et al. [82]) *Assume that there exists $\vartheta > 0$ such that μ^* and $\hat{\nu}_\theta^N$ satisfy LSI(ϑ) $\forall \theta = (\theta_i)_{i=1}^N \in \mathcal{Z}^N$.*

D.1.2 A note about the Central Limit Theorem of shallow NNs

Similar in spirit to the usual *Central Limit Theorem* (CLT) from probability theory, we will be interested in the *fluctuations* of the *empirical process* around its *mean*. Let $\mu^{N,\varepsilon} = (\nu^N_{[t/\varepsilon]})_{t \geqslant 0} \in D_E([0,T])$ be the empirical process associated to the general SGD Dynamics (equation (2.2)) and $\mu = (\mu_t)_{t \geqslant 0}$ the unique solution of the associated DD (equation (2.5)), both starting from μ_0.

We will be interested in understanding the *fluctuation process* given by:

$$\eta_t^{N,\varepsilon} := \sqrt{N}(\mu_t^{N,\varepsilon} - \mu_t)$$

The main idea will be to try to determine whether this fluctuation process has a fixed asymptotic *law* (possibly another process in $D_E([0,T])$), in a similar way to how the usual CLT works.

The *natural candidate* would correspond to a *Gaussian distribution* in the space $D_E([0,T])$. Let's define it in the following manner (where the exact definition of the relevant spaces shall be found in Descours et al. [24]):

Definition D.1 ((G-process) as in Descours et al. [24]) *We say that $\mathcal{G} \in C(\mathbb{R}^+, H^{-J_0,j_0}(\mathcal{Z}))$ is a G-process if for all $k \geqslant 1$ and $f_1, \ldots, f_k \in H^{J_0,j_0}(\mathcal{Z})$,*

$$\{t \mapsto (\langle f_1, \mathcal{G}_t\rangle, \ldots, \langle f_k, \mathcal{G}_t\rangle)^\top, t \in \mathbb{R}^+\} \in C(\mathbb{R}^+, \mathbb{R}^k)$$

is a process with zero-mean, independent Gaussian increments (and thus a martingale), and with covariance structure given by: for all $1 \leqslant i, j \leqslant k$ and all $0 \leqslant s \leqslant t$,

$$Cov(\langle f_i, \mathcal{G}_t\rangle, \langle f_j, \mathcal{G}_s\rangle) = \alpha^2 \mathbb{E}\left[\frac{1}{|B_\infty|}\int_0^s Cov(Q_v[f_i](x,y), Q_v[f_j](x,y))\, dv\right]$$

where $Q_v[f](x,y) = (y - \langle \sigma_*(\cdot, x), \bar{\mu}_v\rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_v\rangle$ for $f \in H^{J_0,j_0}(\mathcal{Z})$ and $\bar{\mu}$ is the solution to the DD equation (2.7).

Definition D.2 *Let ν be a $H^{-J_0+1,j_0}(\mathcal{Z})$-valued random variable. We say that a $C(\mathbb{R}^+, H^{-J_0+1,j_0}(\mathcal{Z}))$-valued process η on a probability space is a weak solution of equation* (D.1) *with initial distribution ν if there exists a G-process $\mathcal{G} \in C(\mathbb{R}^+, H^{-J_0,j_0}(\mathcal{Z}))$ such that the following equation holds:*

a.s. $\forall f \in H^{J_0,j_0}(\mathcal{Z}), \forall t \in \mathbb{R}^+$,

$$\begin{aligned}\langle f, \eta_t\rangle - \langle f, \eta_0\rangle = &\int_0^t \int_{X\times Y} \alpha(y - \langle \sigma_*(\cdot, x), \bar{\mu}_s\rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \eta_s\rangle \pi(dx, dy) \\ &- \int_0^t \int_{X\times Y} \alpha\langle \sigma_*(\cdot, x), \eta_s\rangle\langle \nabla f \cdot \nabla \sigma^*(\cdot, x), \bar{\mu}_s\rangle \pi(dx, dy) + \langle f, \mathcal{G}_t\rangle\end{aligned} \tag{D.1}$$

and $\eta_0 = \nu$ in distribution.

In addition, we say that weak uniqueness holds if for any weak two solutions η_1 and η_2 of equation (D.1) *(possibly defined on two different probability spaces) with the same initial distributions, it holds $\eta_1 = \eta_2$ in distribution.*

The main theorem of this branch of the literature states as follows:

Assumption 14 (From Descours et al. [24]) *Further assume:*

1. *The distribution $\mu_0 \in \mathcal{P}(\mathcal{Z})$ is compactly supported.*
2. *$|B_k| \to |B_\infty|$ a.s. as $k \to \infty$*

Theorem 18 ((**Central Limit Theorem**) as in Descours et al. [24]) *Assume assumption 8 (3.) and assumption 14. Consider the quadratic loss $\ell(y, \hat{y}) = |y - \hat{y}|^2$; let $\varepsilon_N = \frac{1}{N}$ and $\varsigma \equiv \alpha > 0$ (i.e. $s_k^N = \frac{\alpha}{N}$ is the simple learning rate).*

Let $(\theta^k)_{k\in\mathbb{N}}$ be obtained from following the ***minibatch*** *SGD Dynamics (equation* (2.2) *with $\tau = 0$ and the noise term divided by N^δ with $\delta > \frac{1}{4}$) starting from μ_0; and let*

$\mu := (\mu_t)_{t \geqslant 0} \in D_E([0,T])$ be the unique solution of the simple distributional dynamics (equation (2.7)) starting from μ_0. Define the rescaled empirical training process as $\mu^N := (\nu^N_{[Nt]})_{t \geqslant 0} \in D_E([0,T])$ and the fluctuation process as $\eta^N = (\eta^N_t)_{t \geqslant 0} := (\sqrt{N}(\mu^N_t - \mu_t))_{t \geqslant 0} \in D_E([0,T])$. Then:

1. *(Convergence) The sequence $(\eta^N)_{N \geqslant 1} \subset D(\mathbb{R}^+, H^{-J_0+1,j_0}(\mathcal{Z}))$ converges in distribution to a process $\eta^* \in C(\mathbb{R}^+, H^{-J_0+1,j_0}(\mathcal{Z}))$.*
2. *(Limit equation) Let ν_0 be the unique $H^{-J_0+1,j_0}(\mathcal{Z})$-valued random variable such that for all $k \geqslant 1$ and $f_1, \ldots, f_k \in H^{J_0-1,j_0}(\mathcal{Z})$,*
$$(\langle f_1, \eta^*_0 \rangle, \ldots, \langle f_k, \eta^*_0 \rangle)^\top \sim \mathcal{N}\left(0, \Gamma(f_1, \ldots, f_k)\right),$$
where $\Gamma(f_1, \ldots, f_k)$ is the covariance matrix of the vector $(f_1(W^0_1), \ldots, f_k(W^0_1))^\top$. The process η^* has the same distribution as the unique weak solution $\eta^\star$ of equation (D.1) with initial distribution ν_0.

*This also holds in the setting of Sirignano and Spiliopoulos [81], where $(\theta^k)_{k \in \mathbb{N}}$ just follows the **simple** SGD Dynamics (equation (2.1)); it is a particular case, in which $B_k \equiv 1$ $\forall k \in \mathbb{N}$ and thus $B_\infty = 1$ in the definition of the G-process.*

Remark 1. Notice that equation (D.1) is a **stochastic** PDE that is **linear** (in η), whereas the DD from equation (2.7) is a **deterministic** PDE but NON-Linear. Also, equation (D.1) is clearly coupled to the DD. By its linearity, as the equation is driven by a Gaussian process, the limit η_t itself must also be a Gaussian process.

2. As in the usual probabilistic setting, theorem 18 allows us to *approximate* our rv's behaviour, for large N as:
$$\nu^N_{[Nt]} \approx \mu_t + \frac{1}{\sqrt{N}} \eta_t$$
where η_t has a Gaussian distribution with a specific variance-covariance structure.

3. As shown by the combination of *simple* Propagation of Chaos (theorem 3) and CLT (theorem 18) results, we get some insight about how training of the NN should take place: the number of hidden units and stochastic gradient steps should be of the **same order to have convergence and statistically good behavior** (e.g. $k = NT$ SGD steps for N hidden units).

Having understood how fluctuations of our *training process* evolve overtime; we are now interested in discovering whether this process will converge to a meaningful distribution for the problem.

D.1.3 Existence and Uniqueness of Solutions for the McKean-Vlasov Dynamic

In their paper, Bortoli et al. [9] prove the existence of strong solutions with pathwise uniqueness for the non-homogeneous McKean-Vlasov equation with non-constant covariance matrix (based on Theorem 1.1. of Sznitman [83]); i.e.

$$dZ_t = b(t, Z_t, \mu_t)dt + \sigma(t, Z_t, \mu_t)dB_t$$

where b and σ satisfy the conditions of **B2** (presented right after) and for all $t \geqslant 0$, $\mu_t = \mathbf{Law}(Z_t) \in \mathcal{P}_2(\mathbb{R}^D)$, $(B_t)_{t \geqslant 0}$ is an r-dimensional Brownian motion (in the case where r could be different from D), and Z_0 has law (fixed) $\mu^0 \in \mathcal{P}_2(\mathbb{R}^D)$. For this, consider the following technical assumptions (**B1** and **B2**) which have been taken directly from [9]

Assumption 15 (Assumptions for the existence/uniqueness of solutions in Bortoli et al. [9]) *Consider:*

B1. *There exist a measurable function $g : \mathbb{R}^D \times \mathcal{W} \to \mathbb{R}$, $M_1 \geqslant 0$ and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^D)$ such that for any $N \in \mathbb{N}$, the following hold.*

(a) For any $w_1, w_2 \in \mathbb{R}^D$ and $z \in \mathcal{W}$ we have

$$\|g(w_1, z) - g(w_2, z)\| \leqslant \zeta(z)\|w_1 - w_2\|, \quad \textit{and} \quad \|g(w_1, z)\| \leqslant \zeta(z)$$

with $\int_{\mathcal{W}} \zeta^2(z)\, d\pi_{\mathcal{W}}(z) < +\infty$

(b) $b_N \in C(\mathbb{R}_+ \times \mathbb{R}^D \times \mathcal{P}_2(\mathbb{R}^D), \mathbb{R}^D)$ and $\sigma_N \in C(\mathbb{R}_+ \times \mathbb{R}^D \times \mathcal{P}_2(\mathbb{R}^D), \mathbb{R}^{D \times r})$.

(c) For any $w_1, w_2 \in \mathbb{R}^D$ and $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^D)$

$$\sup_{t \geqslant 0} \Big\{ \|b_N(t, w1, \mu1) - b_N(t, w2, \mu2)\| + \|\sigma_N(t, w_1, \mu_1) - \sigma_N(t, w_2, \mu_2)\| \Big\}$$
$$\leqslant M_1 \Bigg(\|w_1 - w_2\| + \left(\int_{\mathcal{W}} \int_{\mathbb{R}^D} |\langle g(\cdot, z), \mu_1 \rangle - \langle g(\cdot, z), \mu_2 \rangle|^2 \, d\pi_{\mathcal{W}}(z) \right)^{1/2} \Big\} \Bigg),$$
$$\sup_{t \geqslant 0} \{ \|b_N(t, 0, \mu_0)\| + \|\sigma_N(t, 0, \mu_0)\| \} \leqslant M_1.$$

B2. *There exist $M_2 \geqslant 0$, $\kappa > 0$, $b \in C(\mathbb{R}_+ \times \mathbb{R}^D \times \mathcal{P}_2(\mathbb{R}^D), \mathbb{R}^D)$ and $\sigma \in C(\mathbb{R}_+ \times \mathbb{R}^D \times \mathcal{P}_2(\mathbb{R}^D), \mathbb{R}^{D \times r})$ such that*

$$\sup_{t \geqslant 0, w \in \mathbb{R}^D, \mu \in \mathcal{P}_2(\mathbb{R}^D)} \{ \|b_N(t, w, \mu) - b(t, w, \mu)\| + \|\sigma_N(t, w, \mu) - \sigma(t, w, \mu)\| \} \leqslant M_2 N^{-\kappa}.$$

They use this to prove the following key result:

Proposition 47 (11 of Bortoli et al. [9]) *Assuming **B1** and **B2**. Given $\mu^0 \in \mathcal{P}_2(\mathbb{R}^D)$ as a fixed initial condition; THEN, there exists an $(\mathcal{F}_t)_{t \geqslant 0}$-adapted process $(Z_t)_{t \geqslant 0}$ that is the **unique (pathwise) strong solution** of the McKean-Vlasov SDE:*

$$dZ_t = b(t, Z_t, \mu_t)dt + \sigma(t, Z_t, \mu_t)dB_t$$

Additionally, it satisfies for each $T \geqslant 0$: $\sup_{t \in [0,T]} \mathbb{E}[\|Z_t\|^2] < \infty$