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

CONTRIBUTIONS TO THE STUDY OF THE NEURAL TANGENT KERNEL REGIME FROM A MEAN FIELD PERSPECTIVE

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

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

ARIE WORTSMAN ZURICH

PROFESOR GUÍA:
JOAQUÍN FONTBONA TORRES

MIEMBROS DE LA COMISIÓN:
DANIEL REMENIK ZISIS
FELIPE TOBAR HENRÍQUEZ
MIRCEA PETRACHE

Este trabajo ha sido parcialmente financiado por:
Proyecto FONDECYT 1201948
CMM ANID BASAL FB210005

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS
Y MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL MATEMÁTICO
PROF. GUÍA: JOAQUÍN FONTBONA TORRES
POR: ARIE WORTSMAN ZURICH
FECHA: 2023

CONTRIBUCIONES AL ESTUDIO DEL NEURAL TANGENT KERNEL REGIME DESDE UNA PERSPECTIVA DE CAMPO MEDIO

El Aprendizaje de Máquinas, y en particular las redes neuronales, han existido en la comunidad científica desde la década de 1980. Sin embargo, han sido adoptadas como una práctica común sólo en la última década, con la nueva disponibilidad de capacidad computacional. En la última década, el aprendizaje de máquinas y especialmente el aprendizaje profundo han visto muchos avances, alcanzando grandes hitos en tareas particularmente difíciles en visión computacional, generación de audio, clasificación, salud, bioinformática y muchos otros campos.

Pese a que ha habido grandes logros en los últimos años por el uso de redes neuronales, el por qué estas funcionan, y en particular, por qué generalizan bien pese a estar altamente sobre parametrizadas, aún no es comprendido completamente (por ejemplo, Alexa-net de Google's tiene alrededor de 10^8 parámetros). En este contexto, motivado por las aplicaciones de la Teoría de Probabilidad en Mecánica Estadística, una línea de investigación ha propuesto estudiar el objeto matemático que surge cuándo el ancho de la red tiende a infinito.

Dado que las redes neuronales clásicas son claramente inestables en el límite cuando la cantidad de neuronas tiene a infinito, se necesitan otras parametrizaciones para poder estudiar estos objetos matemáticos. Dos parametrizaciones han ganado especial popularidad: La parametrización del NTK,y al parametrización de Campo Medio.

Ambas parametrizaciones han sido ampliamente estudiadas, pero en el caso del NTK, no se han encontrado límites en términos de Ecuaciones en Derivadas Parciales (EDPs), que si es el caso en las parametrizaciones de Campo Medio. Además, un fenómeno llamado *Lazy Training*, que consiste en la distribución de los parámetros siendo muy similar a la distribución inicial, fue reportado por Chizat and Bach en 2018 para la parametrización del NTK.

En este trabajo, estudiamos el límite de la parametrización del NTK para redes poco profundas (con una capa escondida) entrenadas con descenso de gradiente estocástico usando medidas empíricas. Con esto, encontramos EDPs límite que no han sido parte de la literatura.

Por otra parte, también se estudia el límite de la red cuando la cantidad de neuronas tiende a infinito, para lo que se ocupan herramientas de transporte óptimo. También se estudia el límite cuando el tiempo de entrenamiento es largo en este setting.

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS
Y MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL MATEMÁTICO
PROF. GUÍA: JOAQUÍN FONTBONA TORRES
POR: ARIE WORTSMAN ZURICH
FECHA: 2023

CONTRIBUTIONS TO THE STUDY OF THE NEURAL TANGENT KERNEL REGIME FROM A MEAN FIELD PERSPECTIVE

Machine Learning and neural networks have been around in the scientific and engineering disciplines since the 1980's. Nevertheless, they were adopted as a common practice only in the last decade as a consequence of the new availability of computing power. In the last decade, machine learning and specially deep learning have seen lots of advances, achieving great success in incredibly difficult tasks in computer vision, audio generation, classification, healthcare, bio-informatics and a lot of other fields.

Even though there's been great achievements by using neural networks, we still don't fully understand why they work, and particularly, why do they generalize well in spite of the fact of being heavily over-parameterized (e.g Google's Alexa-net had around 10^8 parameters). In this context, motivated by applications of probability theory to statistical mechanics, a line of research has studied what happens if we study the mathematical object that arises when the wide of the networks go to infinity.

Since the classical neural network parametrization is unstable in the limit as the number of neurons go to infinity, other parametrizations are needed. Two parametrizations have gained special popularity in this line of research: The NTK parametrization and the Mean Field parametrization.

Both the NTK and the Mean Field parametrization have been extensively studied, but in the last one,limits for the quantity of neurons going to infinity have been found in the form of Partial Differential Equations, which is not the case in the former. Also a phenomena called *Lazy Training*, consisting on the distribution of parameters moving only slightly form the parameter's initialization, has been found to occur in a recent paper by Chizat and Bach in 2018.

We study the NTK limit for shallow neural networks (one hidden layer) from an empirical measure perspective when the training is done by Stochastic Gradient Descent or by Langevin Dynamics. With this, we find novel PDE limits, which have different solutions depending on the training. By doing this, we gain insights on what makes Lazy Training occur in the NTK regime for shallow neural networks, and how Langevin Dynamics differs from SGD in this parametrization. On the other hand, we also study the limit as the amount of neurons go to infinity for the neural network process itself. We study this limit by using tools from optimal transport theory. We also study the limit as the training time goes to infinity in this setting.

A mis papás, Marcelo y Yael.

Agradecimientos

En primer lugar agradezco a mis papás: Marcelo y Yael. Esta tesis es de ustedes. Gracias a ellos estoy donde estoy en este momento y he tenido todo lo que he tenido. Les agradezco siempre haber incentivado todos mis intereses, tanto en la matemática como en otros ámbitos, y haber siempre priorizado mi educación y salud ante todo. También agradezco profundamente a mi hermana Sigal y mi hermano Eitán: Gracias por soportarme estos seis años y por todas las risas y historias y por todo el apoyo que me han dado desde que tengo memoria.

Agradezco a mis amigos del colegio Alejandro, Alex, Benjamín, Daniel, David, Eitán, Matías y Martín, y a muchos otros que no alcanzo a mencionar. A mis amigos del DIM; Cristian, David, Felipe, Nicolás, Pablo, Sebastián y Tristán: Muchas gracias por todas las tardes de estudio y matraca y las salidas a comer. También al Alvaro y al Vicente infinitas gracias por todas las historias que me dieron desde el 2017. También agradezco a Pablo Ugalde por haberme acompañado y guiado mi trabajo durante 3 meses en Bordeaux en mi pasantía.

Agradezco a todos mis profesores de la educación media y superior. Agradezco especialmente al Profesor Alberto Araniz, quien me enseñó el mundo de las matemáticas y que además fue un importante guía cuando estaba en el colegio.

También agradezco al Profesor Joaquín Fontbona por todo su apoyo, su infinita paciencia, su disponibilidad para ayudarme en todo momento y su preocupación por esta tesis. Este tesis no habría sido lo mismo sin el. También agradezco a los Profesores Daniel Remenik y Felipe Tobar por su constante apoyo en el desarrollo de esta tesis y durante mis estudios universitarios. También agradezco al Profesor Mircea Petrache por su disposición a formar parte de esta comisión. Agradezco también a todos los funcionarios del DIM, especialmente a Natacha, Eterin, Karen, Silvia, Luis Mella y Oscar Mori, que permiten que el DIM sea el mejor lugar del mundo.

Agradezco a mis tíos, tías, abuelos y abuelas. A mi tío Claudio y a mis tías Ximena, Guisela y Tamara: Gracias por su constante apoyo. Agradezco a mi Abuelo Lázaro y a mi Abuelita Tati, que sé que me hubiesen llenado de buenos consejos desde mi entrada a la Universidad de Chile. También agradezco a mi Abuelo Isaias, que sé que habría estado más feliz que nadie por este trabajo y que lo habría celebrado junto a mi. También agradezco a mi Abuelita Gloria, quien tengo la fortuna de que siga apoyándome día a día.

Por último, agradezco a una persona que me tuvo que aguantar estos 6 años más que todos, y que a cambio sólo me entregó risas, cariños, abrazos, una Bimba y hermosos momentos. Parte de esta tesis es para ti Anto.

Table of Content

Chapter 1

Preliminaries

1.1. Introduction

In the last decade, the research areas of machine learning and artificial intelligence have witnessed major advances and breakthroughs, being guided specially by major advances in terms of *hardware*, but also in terms of different techniques of great complexity. Most of these breakthroughs are related to what today is called Deep Learning, which is the sub-area of machine learning concerned with what we call *neural networks*.

Neural networks were introduced in the year 1943 by Warren McCulloch and Walter Pitts, in [1], but it's only in the last decade that they gained special popularity, by achieving great success in areas such as healthcare, computer vision, physics, bio-informatics and numerical methods, among a lot of others, see for example [2].

Despite the major advances in the use and applications of deep learning techniques, the nowadays available theory still can't completely explain the success of neural networks. In particular, the fact that **neural networks can generalize on the test set very well while being heavily over-parametrized** remains a major challenge in the theoretical study of neural networks, specially considering the fact that most of the time they achieve zero training error (see, e.g [3]). As a matter of fact, the most widely used architectures for computer vision and natural language processing have approximately 10^8 parameters, a number that has only been growing in the last years.

In this context, different lines of research have emerged in the quest of explaining the good generalization properties of neural networks. In this work, we study a very recent one, which consist in studying the mathematical object that arises when the quantity of neurons in a neural networks goes to infinity. Even though this line of research has studied both shallow (one hidden layer) and deep neural networks, we'll focus only on the former case.

In this chapter, we'll study the basic ideas behind neural networks and how they are trained. We'll start by defining neural networks, explaining the classical parametrizations and after that we'll study gradient descent, and his stochastic equivalent, stochastic gradient descent. We'll also discuss connections with two algorithms commonly used to train neural networks. We'll end by introducing Reproducing Kernel Hilbert Spaces and give some intuitions about them.

1.2. Neural Networks (NNs)

This section is based on the Chapter of Bengio et al [2].

If we are going to study modern neural networks, we must begin by saying what are they. In the context of machine learning, we'll think of neural networks ad mathematical functions mapping some input to some outputs. This functions are formed by composing many simpler functions, which we call **neurons**. When we stack different neurons not connected between them, but connected to other groups of this kind, we obtain a **layer** of neurons.

Being functions, one might think that there must be a reason behind the success of neural networks. In 1981, Kurt Hornik proved, in [4], that neural networks have very good approximation capabilities. As a matter of fact, Hornik proved that neural networks can approximate arbitrarily good any continuous function. Yet, as almost always in mathematical analysis, we still don't know which is the right way to approximate any function, we can only rely on the fact that this special way exists.

The way we connect different neurons and layers inside a neural network is called the NN's architecture. A lot innovations in this field have ocurred by introducing new architectures. Since different architectures work very differently, both in their training and in the applications they are used for, we must introduce some background on these different architectures.

Over the years, many architectures have been widely used in the machine learning community, so the definition of a neural networks depends highly on which type of architecture are we talking about. To tackle this problem, we'll start by defining the most simple type of neural network: The feedforward neural networks. After that, we'll talk briefly about to other architectures: Convolutional Neural Networks and Recurrent Neural Networks.

1.2.1. Feedforward Neural Networks

Feedforward neural networks, also called multilayer perceptrons, are machine learning models. Their goal is to approximate a function $f^*(x)$, in order to do a classification or regression tasks, among others. They are called feedforward because information flows from the input x, through intermediate computations, and finally to an output y.

They are called networks, because the final computation is a product of the composition and weighting of other function, meaning at the end the neural networks f can usually be written as an **iterated composition**:

$$f = f^{(k)}(f^{(k-1)}(\dots(f^{(1)}(x))\dots)).$$

The function $f^{(i)}$ is called the i-th layer of the neural network. The length of this chain is called the **depth** of the model.

Each layer has a basic element, called a **neuron**, which is essentially the evaluation of an **activation function** on a given input. In feedforward neural networks, this input is the output of the last layer weighted by the neuron's weights assigned to each of them. The quantity of neurons in a given layer is called it's **width**.

Layers between the first and last layers are called **hidden layers**. Graphically, a neural network with one hidden layer can be seen in Figure 1.1.

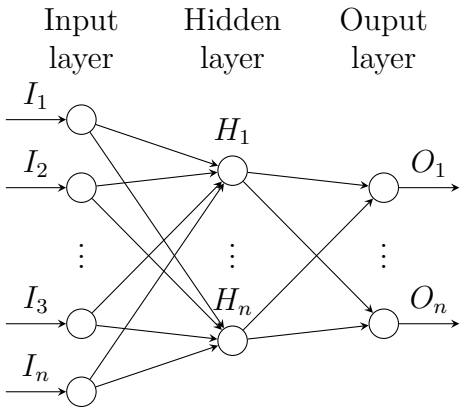

Figure 1.1: Shallow neural networks. The input go through the first layer, and after intermediate computations and output is given in the last layer.

Typically, a shallow neural network f, with m neurons and activation function σ has the form:

$$f(x) = \sum_{i=1}^{m} c_i \sigma(w_i, x).$$

the vectors $(c_i)_{i=1}^m$ and , $(w_i)_{i=1}^m$ are called weights, and for each $w_i, c_i \in \mathbb{R}$.

Remark The parametrization given above is not convergent when m goes to infinity, so it would need a normalization term if we were trying to study this limit.

1.2.2. Other Architectures for Neural Networks

Sometimes, fully-connected NNs have simply too many connections. There are cases were **having less (or better) connections can have an important effect on the accuracy** of the neural network. This is the case of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Although we'll not be studying this type of architectures, we'll define them and explain their uses for completeness.

There are also architectures that we will not discuss, but are also widely used. This is the case, for example, for transformers and other models that have been used only for the last couple of years.

1.2.2.1. Convolutional NNs (CNNs)

Convolutional neural networks (CNNs) have less connections than classical feedforward neural networks. They are specially used for processing data with grid-like topologies, like audio, images and video.

The operation that has special role in this kind of architectures is, as it's name says, the convolution operation. The convolution operation between two functions is essentially a type of weighted average. In a more rigorous way, the convolution of two function $x(t)$ and $w(t)$,

denoted by $x * w(t)$ is given by

$$x * w(t) := \int x(s)w(t-s)ds,$$

where $t \in \mathbb{R}$. If x and w only take integer values as inputs, we can also define a discrete convolution:

$$x * w(t) := \sum_{i=-\infty}^{\infty} x(i)w(t-i)ds.$$

We can also define convolution for multi dimensional input using **kernels**. For example, given an image I and a kernel K, we can define:

$$S(i,j) = (I * K) := \sum_m \sum_n I(m,n)K(i-m, j-n).$$

By using Kernels that only use part of the images, we can give different importance weights to different parts of an image, which allows us to get more information for a neighborhood of the picture, multiple times. This allows the neural networks to be better at learning invariances in data. This is the reason behind the huge success of CNNs in computer vision.

Now, if we a feedforward NN where we add layers that act as convolution kernels (and hence, are not necessarily fully connected we the previous layer), what we obtain is called a convolutional neural network.

1.2.2.2. Recurrent Neural Networks

Just like CNNs are better for data that has grid-like topologies, Recurrent Neural Networks (RNNs) are made for sequentially ordered data, e.g time series. The main improvement of RNNs over classical feedforward neural networks is that there are parameters shared by different parts of the model.

This last fact is essential for generalizing ad level t the learning to variable length inputs. The idea is that the output $h(t)$ of the neural network is allowed to depend on $x(t)$, an input observed just before the current level. This can be seen in mathematical language in the following equation:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \theta).$$

The parameter θ can be shared by multiple computations in the model, and the current computation of a hidden state $h^{(t)}$ is allowed to use the last one, $h^{(t-1)}$.

1.3. Training Neural Networks and Machine Learning Models

The architecture of neural networks is definitely important for the problem one is trying to solve. But once the model is set, it has to be trained with data in order to succeed at a given task.

Learning in this context means optimizing a functional (in particular, minimizing a risk function) using training data (or a **training set**) and being able to generalize this knowledge

to previously unseen data.

In a mathematical way, neural networks $f : \Theta \times \mathcal{X} \to \mathcal{Y}$ are trained to minimize a functional $L(\theta)$. This means, we try to solve the following optimization problem:

$$\min_{\theta \in \Theta} L(\theta).$$

We recall that θ stand for all the parameters of the NN. Classical choices for L are the quadratic loss, $L(\theta) = \mathbb{E}_{X,Y\sim\pi}[(Y - f(\theta, X))^2]$, were π is the distribution of our training set, or categorical loss for classification tasks. Also, sometimes L has incorporated some type of regularization so it minimizes the chances of **overfitting**. Overfitting refers to the case when a machine learning model has been over-adjusted to it's training set, which causes an under-perfomance in it's generalization capabilities.

The main obstacle in this optimization problem is that usually, L is not convex ad a function of the parameters. Therefore, conditions for attaining global minima can be very difficult to grant, so a minimum must be found, usually, using algorithms and numerical methods. Also, finding the global minima of this functional can be of little use if we can't **generalize** well to unseen data, which means that our model is not just trained for it's training set, but also for data that was not part of this process (which is what we aim for when training machine learning models).

In this section we will explain the optimization algorithms that neural networks and most of the machine learning models use for minimizing the functional L.

Most algorithms for training machine learning models are gradient-based algorithms. This means the gradient at the current point has to be computed. In most machine learning models, neural networks included, the gradient or an approximation of it can be computed efficiently. The process in neural networks by which gradient are computed in practice in a very efficient and elegant way is called **automatic differentiation**.

Next, We present the two main optimization algorithms for machine learning models, and variations for one of them. Both algorithms require the user to **initialize** the parameters. **Different initializations can have very different results in the different optimization algorithms**. Normally, this initialization is performed by sampling the parameters out of a distribution, and without necessarily requiring them to be independent at initialization.

1.3.1. Gradient Descent

The most simple and widely used algorithm for optimization is gradient descent. The idea behind it is to repeatedly go in the direction of descent of the gradient of the function. This means, for a function $L(\theta)$, if our parameter at the n-th iteration is denoted by θ_n, then the next iteration is found by calculating:

$$\theta_{n+1} = \theta_n - \gamma_n \nabla L(\theta_n),$$

were $\gamma_n > 0$ is a parameter which is typically constant. The choice of the negative gradient as a descent direction makes sense, since this is the direction of steepest descent of the object

function f. For more details, a good reference is [5].

A big obstacle in training by gradient descent is the fact that since the optimization problem is usually non convex and high-dimensional, gradient descent is very ineffective and it usually gets trapped in saddle points or local minima instead of reaching a global minima.

It's also very important to note that if L is not differentiable, then it's gradient can't be calculated and only approximations or sub-gradients can be used.

1.3.2. Stochastic Gradient Descent and Variations

In stochastic gradient descent (SGD), we don't require the direction updated to be based exactly on the gradient. Instead, we allow the update to be a **random vector whose expected value is the gradient** (See, e.g [6]).

More precisely, for an instance θ_n, we sample $(X_n, Y_n) \sim \pi$, with π the distribution of the training set, and we update according to:

$$\theta_{n+1} = \theta_n - \gamma_n \hat{L}(X_n, Y_n, \theta_n),$$

where $\mathbb{E}[\hat{L}(X, Y, \theta)] = L(\theta)$ for all $\theta \in \Theta$ and (X, Y) random with law π. The fact that stochastic gradient descent incorporates noise to the process makes it easier for the algorithm to get out of local minima. Even more, it's common for SGD to achieve better generalization out of the training set than GD. In this case γ_n must typically go to 0 slowly when $n \to \infty$, see [7].

Sometimes, using only one sample can be too noisy. To achieve a more stable convergence to a global or local minima, practitioners use what is called **mini-batches**. A mini-batch of size k is a sample $\left\{(X_{nk}, Y_{nk}), (X_{nk+1}, Y_{nk+1})..., (X_{(n+1)k-1}, Y_{(n+1)k-1})\right\}$. It is incorporated in the algorithm by redefining the updates as:

$$\theta_{n+1} = \theta_k - \gamma_k \frac{1}{k} \sum_{l=nk}^{(n+1)k-1} \hat{L}(X_i, Y_i, \theta_n).$$

Note that in this case, we still have for all θ:

$$\mathbb{E}_{(X,Y)} \left[\frac{1}{k} \sum_{l=nk}^{(n+1)k-1} \hat{L}(X_i, Y_i, \theta) \right] = L(\theta).$$

Another way to modify these algorithm is by adding noise. This is called **Stochastic Gradient Langevin Dynamics**. The updates are given by:

$$\theta_{n+1} = \theta_n - \gamma_n \hat{L}(X, Y, \theta_n) + \lambda \eta_n,$$

where, for example, $\eta_n \sim \mathcal{N}(\vec{0}, \Sigma)$ are i.i.d.

1.4. Reproducing Kernel Hilbert Spaces

One last thing we should study before our main problem, are Reproducing Kernel Hilbert Spaces (RKHS). The reader should notice the fact that Kernels are well studied mathematical objects, and Kernel Regression in particular is very well understood. This is why searching for connections between Kernels and neural networks is a good starting point for understanding better neural networks. We'll follow the approach presented in [8]. We refer the reader to the book [9] for the proof of the results presented in this section.

We'll begin by defining what is a kernel, to later define what are the so-called Reproducing Kernel Hilbert Spaces or RKHS.

The definition below is the most general definition of a Kernel one could use.

Definition 1.1 (Kernel) *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel over $\mathcal{X}$.*

Under certain conditions, one can grant that K is such that given $x, x' \in \mathcal{X}$

$$K(x, x') = \langle \phi(x), \phi(x') \rangle,$$

for a function $\phi : \mathcal{X} \to \mathcal{H}$, with $(\mathcal{H}, \langle \cdot, \cdot \rangle$ a Hilbert Space called the *features* space. In this regard, the next theorem is very important.

Theorem 1.1 (Mercer's Condition) *Let $\mathcal{X} \subseteq \mathbb{R}^N$ be a compact set and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous symmetric function such that:*

$$\int_{\mathcal{X} \times \mathcal{X}} K(x, x') dx dx' < \infty.$$

Then K admits a uniformly convergent expansion in the following form:

$$K(x, x') = \sum_{n \geq 0} a_n \langle \phi_n(x), \phi_n(x') \rangle,$$

where $(\phi_n)_n$ are a base of $\mathcal{H}$, and with $a_n > 0$ if and only f the kernel is positive semi-definite .

But, what does positive semi-definite means in this context ?

Definition 1.2 (Positive Semi-definite Kernel) *A K is called positive semi-definite (PSD) if it's Gramm matrix, that is, the matrix $(K(x_i, x_j))_{i,j=1}^D$ is always positive semi-definite.*

Lemma 1.1 (Cauchy Schwarz for Kernels) *Let K be a PSD kernel. Then:*

$$\forall x, x' \in \mathcal{X}, K(x, x')^2 \leq K(x, x) K(x' x').$$

Theorem 1.2 (Reproducing Kernel Hilbert Space (RKHS)) *Let $K : \mathcal{X} \times \mathcal{X}$ be a PSD kernel. Then there exists a Hilbert Space $\mathcal{H}$ and a function $\phi : X \to \mathcal{H}$ such that:*

$$\forall x, x' \in \mathcal{X}, K(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Even more, $\mathcal{H}$ satisfies the ***reproducibility property****, mainly:*

$$\forall h \in \mathcal{H}, \forall x \in \mathcal{X}, h(x) = \langle h, K(x, \cdot) \rangle.$$

$\mathcal{H}$ is called the Reproducing Kernel Hilbert Space associated to K.

Remark The space L^2 is not a function space and therefore it is not a $RKHS$.

It is possible to give an alternative (and equivalent) definition of RKHS:

Definition 1.3 *A RKHS $\mathcal{H}$ is a Hilbert space of real functions into $\mathcal{X}$ such that for all $x \in \mathcal{X}$ the operator*

$$\mathcal{E}_x : \mathcal{H} \to \mathbb{R}$$
$$f \to f(x),$$

is bounded.

1.4.1. Building a RKHS

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel, and define for all $x \in \mathcal{X}$

$$\phi(x) = K(x, \cdot).$$

Consider the set $\mathcal{B} := \{\phi(x) : x \in \mathcal{X}\}$, and consider:

$$\mathcal{H}_K := span(\mathcal{B}),$$

where the span of a set is the set of all linear combinations of the elements of the set. Consider the following operation on $\mathcal{H}_K$: Given $f = \sum_i a_i K(x_i, \cdot)$ and $g = \sum_j b_j K(x_j, \cdot)$, we define

$$\langle f, g \rangle_{\mathcal{H}_k} = \sum_{i,j} a_i b_j K(x_i, x_j).$$

Proposition 1.1 *The operation $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ defines an inner product in $\mathcal{H}_K$. With this inner product, $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space.*

1.4.2. Interpolation and Adjustment

Let's suppose we have n samples of a function $f^* : \mathcal{X} \to \mathbb{R}$. Two questions arise naturally:

1. For a fixed space $\mathcal{F}$, does there exist a function $f \in \mathcal{F}$ such that $f(x_i) = y_i$, $\forall i \in [n]$?
2. If so, of all the functions in $\mathcal{F}$ such that $f(x_i) = y_i$ for all i, which one is the best?

Both questions can be answered by the following approach: Given and RKHS, of all functions that adjust to data, we choose the one with the least norm in the RKHS. This can be formulated mathematically by:

$$\textbf{(P)} \quad \min_{f \in \mathcal{H}} \quad \|f\|_{\mathcal{H}}$$
$$\text{s.a} \quad f(x_i) = y_i \quad \forall i \in [n]$$

This method is known ad **minimum - norm interpolation**.

Lemma 1.2 *Let $K \in \mathbb{R}^{n\times n}$ be a matrix with entrances*

$$K_{ij} = \frac{K(x_i, x_j)}{n}.$$

*Then **(P)** is feasible if and only if $y \in rang(K)$. In this case, the solution is given by:*

$$\hat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\alpha}_i K(\cdot, x_i),$$

where $K\hat{\alpha} = \frac{y}{\sqrt{n}}$.

In a statistical framework, is unrealistic to assume that we can interpolate our data in an exact way, specially considering our data might be subject to noise. For this reason, it's more realistic to assume a noisy observation process. In this case, we have:

$$y_i = f(x_i) + w_i,$$

for $i = 1, ..n$. The vector $(w_i)_{i=1}^n$ models the noise in our observations. In this context, a more realistic model might search to solve:

$$\min_{f\in\mathcal{H}} \quad \|f\|_{\mathcal{H}}, s.t \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \delta^2,$$

where δ is a tolerance parameter. Alternatively, we might minimize the following problem:

$$\min_{f\in\mathcal{H}} \quad \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \|f\|_{\mathcal{H}}, s.t \|f\|_{\mathcal{H}} \leq R,$$

for an appropriately chosen radius $R > 0$. Since both minimization problems are convex, their sum is also convex. Hence, it can be obtained (by a Lagrangian duality argument) that this is equivalent to solve the following problem:

$$\min_{f\in\mathcal{H}} \quad \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{H}}^2,$$

where the regularization parameter $\lambda_n \geq 0$ is a function of the tolerance δ and the radius R.

Lemma 1.3 *Let $K \in \mathbb{R}^{n\times n}$ be a matrix with entrances*

$$K_{ij} = \frac{K(x_i, x_j)}{n}.$$

*then **(P)** is feasible if and only if $y \in rang(K)$. In this case, $\forall \quad \lambda_n > 0$ the solution is given by*

$$\hat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\alpha}_i K(\cdot, x_i),$$

where $(K + \lambda_n I_n)\hat{\alpha} = \frac{y}{\sqrt{n}}$.

1.5. Functional Gradient Descent

Let K be a PSD symmetric kernel. Consider the evaluation functional $\mathcal{E}_x : \mathcal{H}_K \to \mathbb{R}$, given by $\mathcal{E}_x[f] = f(x)$. Then by the reproducing property:

$$\mathcal{E}_x[f] = \langle f, K(x, \cdot)\rangle.$$

This way, the differential of $\mathcal{E}_x[f]$ is given by:

$$\nabla\mathcal{E}_x[f] = K(x, \cdot).$$

Now, consider the loss function $L : \mathcal{H}_K \to \mathbb{R}$

$$L[f] = \frac{1}{2}\mathbb{E}_{X,Y}[(Y - f(X))^2],$$

where (X, Y) are random vectors in $\mathcal{X} \times \mathbb{R}$. Then, we'd like to ask who's the differential of L at a function f. We obtain, by the chain rule:

$$\nabla L[f] = \mathbb{E}_{X,Y}[(Y - f(X))\nabla\mathcal{E}_X[f]],$$

and by replacing the differential of $\nabla\mathcal{E}_X[f]$:

$$\nabla L[f] = \mathbb{E}_{X,Y}[(Y - f(X))K(X, \cdot)].$$

Definition 1.4 *Let $f(t) : \mathbb{R} \to \mathcal{H}_K$ be a function such that for all $t \geq 0$*

$$\frac{d}{dt}f = \mathbb{E}_{X,Y}[(Y - f_t(X))K(X, \cdot)].$$

We say that $f_t(\cdot)$ satisfies functional gradient descent, or ***Kernel Gradient Descent****.*

Chapter 2

Different Parametrizations of Shallow Neural Networks

Neural Networks have achieved great success in a number of tasks that a decade ago seemed impossible or very difficult. Different challenges in computer vision, finance, simulation, among others, have been solved by the use of deep neural networks. Also, the practical use of neural networks has witnessed lots of improvements and innovations, most of them guided by empirical tests in data. Despite this last fact, theory has failed to catch up to this pace. It's for this reason that in the last years different theories have tried to explain why and how neural networks work.

One big question in the Theory of Deep Learning has been: Why do Neural Networks have good generalization properties, even though they are heavily over-parametrized? Since the very first models in Statistical Inference, it's been widely known that over-parametrized models only cause overfitting to training data, and it has bad generalization properties. Yet deep neural networks, having millions or hundreds of millions of parameters, seem to be an exception to this statement.

A natural approach to study neural networks would be to study the mathematical object that arises when the number of neurons diverges to infinity. Even though this might sound weird in a first thought, the truth is this idea has been central in the field of Statistical Mechanics for a number of decades. By taking the number of neurons to infinity, the results can be studied as a Law of Large Numbers for a particle system, with the neurons being it's particles and the training being it's interactions.

The main goal of studying this object is to find a structure that can be identified as a regularization for the parameters of the neural network, without it being explicit in the training scheme. This is called implicit over-parametrization.

Two main regimes (and parametrizations) have been studied by the community, achieving different results. In Section 2 we'll study the first one: the Mean Field Regime. In this regime, the variance of the initialization of the parameters are smaller, and the dynamics can be written as a gradient flow. Next, we'll study the second one: The Neural Tangent Kernel (NTK) Regime. The relation between both will be studied in Section 3. We'll end by studying what is now called Lazy Training, and important property that arises in the NTK regime.

2.1. Mean Field Regime

In this section, we'll discuss the results exposed in [10], [11], [12] and [13]. This works consider the **mean-field scaling** of shallow neural networks, and study convergence and approximation results by studying the mean field limit. Mean Field limits consist in studying a high dimensional system by approximating it by infinite-dimensional systems. They have lots of applications, such as in physics or economics (see for example, [14] or [15]. In this context, a major concept is the one of *Propagation of Chaos*, which refers to the property of asymptotic Independence of the different variables when the dimension of the system grows. For the reader interested in Propagation of Chaos, we refer them to [16]. In the context of NNs, there also exists studies about propagation of chaos, such as [17] which contributed to the theory by proving propagation of chaos in the mean field regime, and by studying different scaling for the SGD step-sizes.

Let f_θ^m be a shallow neural network, given by:

$$f_\theta^m(x) = \frac{1}{m}\sum_{i=1}^{m} c^i \sigma(w^i x),$$

where for each $i \in \{1, ..., N\}$, $c^i \in \mathbb{R}$ y $x \in \mathcal{X}$ and $w^i \in \mathbb{R}^p$. The network has parameters $\theta = (c^1, ..., c^m, w^1, ..., w^m) \in \mathbb{R}^{(1+p)N}$, which are estimated from the data by minimizing a certain loss function.

The function $\sigma : \mathbb{R} \to \mathbb{R}$ is non-linear, such as the sigmoid function or the ReLu function, $ReLu(x) = \max\{0, x\}$. The quantity $\sigma(w^i x)$ is called the i-th hidden unit. and $(\sigma(w^1 x), ..., \sigma(w^m x))$ is called the network's hidden layer.

The loss function is given by:

$$L^m(\theta) = \frac{1}{2}\mathbb{E}_{X,Y}[(Y - g_\theta^m(X))^2],$$

where $X, Y \sim \pi(dx, dy)$. The parameters are trained by stochastic gradient descent on each step, that is

$$W_{n+1}^i = W_n^i - \nabla l(X, Y, W_n),$$

where $\mathbb{E}_{X,Y}[\nabla l(X, Y, W_n)] = \nabla L(W_n)$. At the n-th iteration, the empirical measure of the parameters is denoted by

$$\mu_k^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{(w_i^k, c_i^k)}.$$

Even though each group in this area has worked with different hypothesis on the data, on the activation function and on the parameter's initial distribution, all of them are based on the fact that we can define:

$$f(\mu, x) = \int c\sigma(w, x)\mu(dc, dw),$$

and this way $f_\theta^m(x) = f(\mu^m, x)$. Then, by noting that:

$$L^N(\theta) = L_\# + \frac{2}{m}\sum_{i=1}^{m} V(\theta_i) + \frac{1}{m^2}\sum_{i,j=1}^{m} U(\theta_i, \theta_j),$$

where:

- $V(\theta) = -\mathbb{E}\{yc_i\sigma(w_i x)\}$,
- $U(\theta_1, \theta_2) = \mathbb{E}\{c_1\sigma(w_1 x)c_2\sigma(w_2 x)\}$,
- $L_\# = \mathbb{E}\{y^2\}$, is the risk of $f^m \equiv 0$;

we can generalize L to general probability measures, because:

$$L^N(\theta^k) = R_\# + \int V(\theta)\mu_k^N(d\theta) + \int U(\theta_1, \theta_2)\mu_k^N(d\theta_1)\mu_k^N(d\theta_1).$$

This way:

$$L(\mu) = R_\# + \int V(\theta)\mu(d\theta) + \int U(\theta_1, \theta_2)\mu(d\theta_1)\mu(d\theta_1),$$

for μ a probability measure over the parameters. The different works study the limiting dynamic of μ in continuous time, for example in [11] they obtain the convergence in distribution of μ^m to the solution of the PDE

$$\langle f, \bar{\nu}_t\rangle = \langle f, \bar{\nu}_0\rangle + \int_0^t \left(\int_{X\times Y} \alpha(y - \langle c\sigma(wx), \bar{\nu}_s\rangle)\langle\nabla(c\sigma(wx))\nabla f, \bar{\nu}_s\rangle\pi(dx, dy)\right) ds.$$

Note that by assuming that μ has a density, we could derive the classical form of a non-linear partial differential equation called 'McKean-Vlasov' (see [18]), which is what [10], [12] and [13] do.

2.2. Neural Tangent Kernel Regime

In 2018, Arhur Jacot, Franck Gabriel y Clement Hongler published [19], where they studied a different parametrization, who's limit could be studied using Kernels, and in particular the Neural Tangent Kernel (NTK). The parametrization studied in [19], called the NTK parametrization, for shallow neural networks was the following:

$$f^m(W, x)\frac{1}{\sqrt{m}}\sum_{k=1}^{m} c_k\nabla\sigma(W^k, x).$$

The training is done by minimizing, through standard gradient descent, the loss:

$$L(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(w, x_i))^2.$$

If we consider the gradient descent algorithm to have very small stepsizes, then it's dynamics can be written as:

$$\frac{dw(t)}{dt} = -\nabla l(w).$$

But, what about the dynamics of the predictions of the neural network? The work presented in [19] gives the following Lemma:

Lemma 2.1 *Let $u(t) = (f(w(t), x_i))_{i=1}^n$ be the network's outputs at time t for the inputs in the training set. Then $u(t)$ follows the dynamics:*

$$\frac{du(t)}{dt} = H(t) \cdot (u(t) - y),$$

where $H(t)$ is a semi definite positive matrix in $R^{n \times n}$ such that

$$H_{ij} = \left\langle \frac{\partial f(w(t), x_i)}{\partial w}, \frac{\partial f(w(t), x_j)}{\partial w} \right\rangle.$$

Proof. In the first place, we know that the parameters evolve by following the dynamic:

$$\frac{dw(t)}{dt} = -\nabla l(w) = -\sum_{i=1}^{n} (f(w(t), x_i) - y_i) \frac{\partial f(w(t), x_i)}{\partial w}.$$

Then:

$$\frac{df(w(t), x_i)}{dt} = \left\langle \frac{\partial f(w(t), x_i)}{\partial w}, \frac{dw(t}{dt} \right\rangle = -\sum_{j=1}^{n} (f(w(t), x_j) - y_j) \left\langle \frac{\partial f(w(t), x_i)}{\partial w}, \frac{\partial f(w(t), x_j)}{\partial w} \right\rangle$$

Note that the right-hand side corresponds to the product of $(u(t) - y)$ and the i-th roe of $H(t)$. Then, we can write the dynamic $u(t)$ as:

$$\frac{du(t)}{dt} = H(t)(u(t) - y).$$

■ □

The main idea behind this theory is that by allowing the quantity of neurons m go to infinity H will become constant during training, i.e equal to $H(0)$. Even more, $H(0)$ will converge in probability to H^*, which will be the NTK $k(\cdot, \cdot)$ evaluated in the training set. Note that in this case, formally:

$$\frac{du(t)}{dt} = H^*(u(t) - y).$$

This dynamic corresponds to the dynamics that appear when training an RKHS regression using Kernel Gradient Descent. Note that:

- In this study, contrasting to the one for the mean field regime, there is no characterization in function of the empirical measure. The paper [11] is a technical note that studies this by considering centered initializations.
- [19] and [11] do not study what happens for different scaling of step sizes in SGD training, and in particular [19] does not study SGD training.
- [20] studies the NTK regime for SGD training, but they don't do it directly: They use the results for the mean field regime and re-scale the dynamics appropriately.

It's also important to recall that the NTK regime is also valid for other types or NNs architectures, such as CNNs. For this, we refer the reader to [21]. As a matter of fact, research points that provably all types of architectures can enter de NTK regime.

2.2.1. Lazy Training in Neural Networks

In [22], it was proved that models like the NTK regime of NN have a very particular property: When highly overparametrized, as in the case of neural networks, the parameters of this models barely move away from their initialization. Formally, let $F : \mathbb{R}^p \to \mathbb{R}$ be an objective function, and let w_0 be the initial parameters of the model.

Let $h : \mathbb{R}^p \to \mathcal{F}$ be our model. We define the linearized model:

$$\bar{h}(w) = h(w_0) + Df(w_0)(w - w_0).$$

Chizat et al. proved in [22] that as the quantity of parameters grows and the model becomes heavily over-parametrized, then the parameters of the model converge to the parameters of the linearized model. They also make a detailed study of when do models enter to the different regimes. Even tough the study is for models and parameters itself, the study of how does this can be interpreted in terms of the evolution of empirical measures is missing in the literature.

Chapter 3

Main Results

There's been two big approaches on a mathematical theory of Deep Learning in the recent years. On of them , the Neural Tangent Kernel (see [19]), describes the limiting dynamics of a neural network initialized with Xavier's initialization. The other approach, the one of mean field analysis of neural networks (see [10]), have described the limit of neural networks when they are initialized with a much smaller variance. In this work, we study how do both techniques are related.

With this objective in mind, we study the Neural Tangent Kernel setting by using tools from mean field analysis: empirical measures and limiting theorems. We use the technique described in [17] to study everything in a continuous setting. We find a limiting PDE for the dynamics of the limiting empirical measure of the parameters, and we also study the limiting object, which is a function in $L^2(\mathcal{X})$, where $\mathcal{X}$ is the input space.

We'll also study if there's any way by which we can see how Lazy Training (see [22]) works in terms of the empirical measure of the parameters.

3.1. The limiting dynamics of the empirical measure

We consider a shallow neural network with m neurons in it's hidden layer, inputs in $\mathcal{X}$ and outputs in $\mathcal{Y} \subseteq \mathbb{R}$, and activation function $\sigma \in \mathcal{C}_b^2$ as a function $f^m : \mathcal{X} \to \mathcal{Y}$ given by:

$$f^m(x) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} c_k \sigma(W^{k,m}, X),$$

where $x \in \mathcal{X}$, $c_k \in \mathbb{R}$ and $W^{k,m} \in \mathbb{R}^p$, with $p \in \mathbb{N}$. At first, we make the following assumptions:

- The activation function $\sigma(w, x)$ is bounded, with bounded-in-norm gradient $\nabla\sigma(w, x)$ and bounded-in-norm hessian $\mathcal{H}_w\sigma(w, x)$.
- The coefficients c_k are initialized with it's first four moments being bounded and **they are not trained after initialization**. We make more comments on this in the following.

- Let $m \in \mathbb{N}$. The parameters $W^{k,m}$, for $k \in \{1, \dots, m\}$, are initialized with a distribution with density and with it's first eight moments being finite.
- All parameters are initialized independently.

Remark If the parameters are initialized with Gaussian distributions, all the hypothesis above are satisfied. Also, note the assumptions above are more general than what we defined as Xavier initialization in section 3, i.e the distributions are not required to be centered gaussians at first. Nevertheless, we will study this specific setting.

We consider that the hidden layer is trained by stochastic gradient descent (SGD) with the loss:

$$\mathbb{E}_{X,Y}[(Y - f^m(X))].$$

On the other hand, we consider the last layer, c_k to be left untrained. This setting is described in [10] and is named fixed-coefficients setting.

We'd like to use all the tools of stochastic calculus and limit theorems for continuous processes. For this, we apply the approach described in [17]. In continuous time, we consider the process $(W_s^{k,m})_{s\in[0,T]}$ of parameters to be guided by the dynamics in the SDE:

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}, \tag{3.1}$$

where, for $\lambda, \tau \geq 0$,

$$h^{k,m}(W_n^m) := -\lambda W_n^{k,m} + \mathbb{E}_{X,Y}\left[(Y - f^m(W_n^m, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X)\right],$$

$$\xi_{k,m}(w) := (Y - f^m(W_n, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X) - h^{k,m}(w),$$

and

$$\Sigma_{k,m}(w) = \mathbb{E}_{X,Y}[\xi_{k,m}(w)\xi_{k,m}^T(w)] \in \mathbb{R}^{p\times p}.$$

The heuristic by which we obtain the dynamics in equation (4.18) are described in the following chapter, and it's details are fully described in [23]. We also study the different regimes that arise when we vary the parameter α. Our first result is the following

Theorem 3.1 *Let $m \in \mathbb{N}$, $(c_k)_{k=1}^m$ and σ with the assumptions considered above. Then, equation* (4.18)

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m},$$

has strong solutions, subject to the fixed coefficients $(c_k)_{k=1}^m$.

In order to prove the convergence of the empirical measure as the number of neurons go to infinity, we use limit theorems. In particular, we prove the tightness of the empirical measures process in $\mathcal{C}([0,T], \mathbb{P}(\mathbb{R}\times\mathbb{R}^p))$. More specifically, we prove the following

Proposition 3.1 *Let $\sigma : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a bounded, Lipschitz, with bounded-in-norm*

hessian activation function for a one hidden layer neural network $f^m(c,w)$, whose parameters are initialized such that $W^{k,m}$ are i.i.d and have their first four moments finite, and c_k are initialized i.i.d with it's first four moments bounded. Let $(\mu_t^m)_t$ be the empirical measure of the process the process $(c_k, W_t^m)_t$ when trained in continuous time by the SDE:

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}.$$

Then the laws of the process given by the empirical measures of the process $(W_t^m)_t$, with $m \in \mathbb{N}$, are tight.

This proposition is a cornerstone in order to prove convergence to a limit in some sense. In particular, by using this results and techniques described in Snitzman's book [16], we prove our main result, which is stated in the following Theorem.

Theorem 3.2 *Let $\alpha > 0$, $\lambda \in [0,1)$, $\gamma \geq 0$, and μ_t^m denote the empirical measure process that represents the weights of a shallow neural network, who's parameters are trained in continuous time by the dynamics:*

$$dW_t^{k,m} = h^{k,m}(W_t^m)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m},$$

where $W_t^{k,m}$ denotes one neuron in the hidden layer. Let μ_0 denote the initialization distribution for the pair (C,W). Then, in the limit as m goes to infinity, the empirical measure converges in Law to the unique solution of the non-linear Focker Planck Equation:

- *If $\alpha = 0$:*

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \nabla\varphi(\tilde{W})^T \tilde{W} \mu_s(d\tilde{c}, d\tilde{W})ds + \int_0^t \mathbb{E}_{X,Y}\left[(\langle c\sigma, \mu_t\rangle)\left\langle c\nabla\sigma(\cdot, X)\nabla\varphi, \mu_s\right\rangle\right]ds \\ &+ \gamma\int_0^t \langle \mathrm{Tr}\left(S(x,\mu_s)^T \mathcal{H}_w\varphi(\cdot)\right), \mu_s\rangle ds + \sqrt{2\tau}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c}, d\tilde{W})ds.\end{aligned} \tag{3.2}$$

- *If $\alpha > 0$:*

$$\langle \varphi, \mu_t - \mu_0 \rangle = -\lambda \int_0^t \nabla\varphi(\tilde{W})^T \tilde{W} \mu_s(d\tilde{c}, d\tilde{W})ds + \sqrt{2\tau}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c}, d\tilde{W})ds. \tag{3.3}$$

This results is similar to the ones described in Sirignano and Spiliopoulos' technical note [11], yet there's a couple of differences:

- While our work is mainly based in continuous time arguments, their work is based mainly in a discrete time setting.
- The initialization distribution of the parameters is more general in our setting, yet is also allows the study made in Sirignano and Spiliopoulos' work. In particular, the work in [11] only studies centered distributions, which described Xavier initialization.

The results described in Sirignano and Spiliopoulos' technical note [11] can be seen in the following:

Corolary 3.1 *If $\lambda = \tau = 0$ and $\mathbb{E}[c] = 0$, then equation 4.5 becomes:*

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle. \tag{3.4}$$

This result has a very interesting interpretation: When the number of neurons go to infinity, the parameters will tend to stay close to it's initial distribution. This is another way to prove the results found in [22], which state that the Neural Tangent Kernel Regime exhibit Lazy Training. Lazy training is defined as the phenomena where parameters tend to stay close ti it's initialization.

With this result, we can also define the Neural Tangent Kernel in our setting: If we consider $\varphi = \sigma$, then we get that given $x_1, x_2 \in \mathcal{X}$, the following convergence is satisfied in law:

$$\langle c^2 \nabla\sigma(\tilde{W}, x_1)^T \nabla\sigma(\tilde{W}, x_2), \mu_s^m \rangle \overset{m\to\infty}{\longrightarrow} \langle c^2 \nabla\sigma(\tilde{W}, x_1)^T \nabla\sigma(\tilde{W}, x_2), \mu_0 \rangle.$$

Considering this, we define the continuous version of the **Neural Tangent Kernel**, first defined in [19], associated to our neural network as the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that

$$K(x_1, x_2) = \langle c^2 \nabla\sigma(\tilde{W}, x_1)^T \nabla\sigma(\tilde{W}, x_2), \mu_0 \rangle. \tag{3.5}$$

3.2. The Limit of the Neural Network

In the last section, we stated our results on how does the empirical measure process behaves when the number of neurons go to infinity. Nevertheless, this does not end the study in our setting: Since the neural network's corresponding scaling does not integrate directly the empirical measure, we have to study it's dynamic separately, but dependent on the empirical measure process.

We only study the case of centered initializations and $\alpha > \frac{1}{2}$, since the more general setting we studied for the empirical measure becomes quite harder. Nevertheless, we do make conjectures about the different results that are possible in the different cases.

To study the limit of the neural network, we define a white noise in the space $L^2(\mathbb{R})$ with covariance μ_0. The details of this construction can be seen in chapter 5. By using this white noise and optimal transport arguments from [24] and [25], we prove a convergence result for

$$f^m(x) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} c_k \sigma(W^{k,m}, X),$$

which is stated in the following

Theorem 3.3 *Let η_0 be a white noise with covariance μ_0, and let f_t be a solution of the*

equation in $L^2(\mathcal{X})$:

$$f_t(x) - f_0(x) = \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds$$
$$- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w \sigma(w)\right)\right)\mu_0(dc,dw)ds., \tag{3.6}$$

with $f_0 = \langle \sigma, \eta_0 \rangle$. Then, for every $t \geq 0$:

$$\lim_{m\to\infty} \|f_t^m - f_t\| = 0.$$

If we state the results in terms of the value of α, we know that when $\alpha = \frac{1}{2}$ the limiting dynamic of f will be:

$$f_t(x) - f_0(x) = \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds$$
$$- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w \sigma(w)\right)\right)\mu_0(dc,dw)ds,$$

and when $\alpha > \frac{1}{2}$:

$$f_t(x) - f_0(x) = \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds,$$

where the initial condition for both equations is $f_0 = \langle \sigma, \eta_0 \rangle$.

We the result in Theorem 5.1 states the convergence of the neural network process. Knowing this, it becomes natural to ask ourselves if there's something we can say about the limit of f_t when t become large. The answer to this questions require the introduction of a new Loss Function, which is defined in terms of the continuous NTK, which we recall is given by

$$K(x_1, x_2) = \langle c^2 \nabla\sigma(\tilde{W}, x_1)^T \nabla\sigma(\tilde{W}, x_2), \mu_0 \rangle.$$

We consider the case when $\alpha > \frac{1}{2}$. With this, we can re-write f_t's dynamics in the following way:

$$f_t(x) = f_0(x) + \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X))K(X,\cdot)]ds.$$

By considering the RKHS definition we gave in the background section, we obtain f_t follows Kernel Gradient Descent on the Reproducing Kernel Hilbert Space, with the Kernel equal to the Neural Tangent Kernel. That is, f_t follows gradient descent on the RKHS associated to the NTK with respect to the loss:

$$L_K : \mathcal{H}_K \to \mathbb{R}$$

$$L_K[f] = \|Y - f(X))\|^2_{\mathcal{H}_K}.$$

We prove the following theorem, which was already proved in [11] and [19]. Nevertheless, we state the theorem and prove it in a different way.

Theorem 3.4 *Let L be the loss we defined above, and consider f_t such that*

$$f_t(x) = f_0(x) + \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X))K(X, \cdot)]ds.$$

Then, if K is a positive definite kernel,

$$\lim_{t\to\infty} L[f_t] = L^*.$$

It's important for the reader to notice that this theorem does not guarantee convergence to 0, unless the minimum of L_K is actually 0. Another interpretation of the Theorem above is the following: In the limit, the neural network will always be overfitted to the data in which it was trained, which can be seen from the fact that, in the limit, it will always minimize a loss that is directly constructed by the training data.

Chapter 4

The NTK Regime through the lens of mean field models

Let's consider a a shallow neural network f^m (i.e with one hidden layer) with m neurons trained by stochastic gradient descent. We parametrize the network with weights $w_i \in (\mathbb{R}^p)^m$, $c_i \in \mathbb{R}$, $i \in \{1, \dots m\}$. Let $\mathcal{X}$ denote our input space, with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ our output space, which we'll consider one-dimensional, i.e $\mathcal{Y} \subseteq \mathbb{R}$. We can write:

$$f^m(w, x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} c_i \sigma(w^i, x).$$

The function $\sigma(\cdot, \cdot) : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ will be an activation function, which we'll consider bounded, Lipschitz and with bounded second derivative. We'll also assume that our input and outputs X, Y have finite second moment.

Let ν be the distribution of our data (X, Y). Then, our neural network f^m will be trained by minimizing the population risk,

$$L(w) = \frac{1}{2} \mathbb{E}_{(X,Y)\sim\nu} \left[(Y - f^m(c, w, X))^2 \right].$$

and performing stochastic gradient descent. We consider the parameters to be initialized as $W_i \sim p_w$ and $c_i \sim p_c$, where W_i are i.i.d. Just for simplicity, we will not train the last layer of the network. This can be seen in the fact that we do not show the dependence of L on c, since c is considered to be fixed. This model is known as 'fixed coefficients' in [10].

Starting from W_0, we train our network by stochastic gradient descent, which can be written in the following way:

$$W_{n+1} = (1 - \frac{\lambda\gamma}{m^\alpha}) W_n - \frac{\gamma}{2m^\alpha} \nabla_w l^m(W_n, X_n, Y_n) + \sqrt{\frac{2\tau\gamma}{m^\alpha}} Z_n, \tag{4.1}$$

where we defined $l^m(w, x, y) = (y - f^m(c, w, x))^2$, $\lambda, \tau \geq 0$ and Z_n is a multivariate standard Gaussian. As the reader may note, this setting is the one of Stochastic Gradient Langevin

dynamics, which we introduced in Chapter 1. Note that:

$$\partial_{w_k} l(w, x, y) = -2\frac{c_k}{\sqrt{m}}(y - f^m(c, w, x))\nabla\sigma(W^{k,m}, x), \tag{4.2}$$

where $W^{k,m}$ are the p weights of the $k-th$ neuron. By re-writing (4.1) focusing on each neuron, and replacing (4.2) into it, we get:

$$W_{n+1}^{k,m} = (1 - \frac{\lambda\gamma}{m^\alpha})W_n^{k,m} + \frac{\gamma c_k}{m^{\alpha+\frac{1}{2}}}(Y - f^m(W_n, X))\nabla_w\sigma(W_n^{k,m}, X) + \sqrt{\frac{2\tau\gamma}{m^\alpha}}Z_n. \tag{4.3}$$

With the aim of using the tools that stochastic calculus can offer to us, we'll study a dynamic that approximates (4.3). This approach was first in [17] and then applied in [23]. Following [17], let $\tilde{\gamma}(m) = \frac{\gamma}{m^\alpha}$ and $\tilde{W^{k,m}}_t$ denote the interpolation of $W_n^{k,m}$, i.e for t in $[n\tilde{\gamma}, (n+1)\tilde{\gamma}]$, we have:

$$\tilde{W_t^{k,m}} = \frac{(t - n\tilde{\gamma})W_{n+1}^{k,m} + ((n+1)\tilde{\gamma} - t)W_n^{k,m}}{\tilde{\gamma}}.$$

the heuristics for finding an approximation for our discrete dynamics in continuous time is the following: In the first place, we can simply write:

$$\begin{aligned}\tilde{W}_{(n+1)\tilde{\gamma}}^{k,m} - \tilde{W}_{n\tilde{\gamma}}^{k,m} &\approx W_{n+1}^{k,m} - W_n^{k,m} \\ &= \lambda\tilde{\gamma}W_n^{k,m} + \tilde{\gamma}(Y - f^m(W_n, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X) + \sqrt{2\tau}\tilde{\gamma}^{\frac{1}{2}}Z_n.\end{aligned}$$

By considering that this random variable depending on X and Y is approximately Gaussian, we can rewrite the last expression in the following form:

$$\approx \tilde{\gamma}\left(-\lambda W_m^{k,m} + \mathbb{E}_{X,Y}\left[(Y - f^m(W_n, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X)\right]\right) + \tilde{\gamma}\Sigma^{\frac{1}{2}}(\tilde{W}_{n\tilde{\gamma}}^{k,m})G_{n+1} + \sqrt{2\tau}\tilde{\gamma}^{\frac{1}{2}}Z_n,$$

where Σ we'll be a covariance matrix that we will specify in a moment, $\Sigma^{\frac{1}{2}}$ is it's unique squared root, and G_{n+1} is a standard Gaussian random variable. Finally, by rewriting as integrals, we get:

$$\begin{aligned}&\approx \int_{n\tilde{\gamma}}^{(n+1)\tilde{\gamma}}\left(-\lambda W_n^{k,m} + \mathbb{E}_{X,Y}\left[(Y - f^m(W_n, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X)\right]\right)ds \\ &+\sqrt{\tilde{\gamma}}\int_{n\tilde{\gamma}}^{(n+1)\tilde{\gamma}}\Sigma_{k,m}^{\frac{1}{2}}(\tilde{W}_s^{k,m})dB_s + \int_{n\tilde{\gamma}}^{(n+1)\tilde{\gamma}}\sqrt{2\tau}d\tilde{B}_s.\end{aligned}$$

where B_s and $\tilde{B}_s$ denote p-dimensional Brownian Motions. Let

$$h^{k,m}(W_n^m) := -\lambda W_n^{k,m} + \mathbb{E}_{X,Y}\left[(Y - f^m(W_n^m, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X)\right],$$

and

$$\xi_{k,m}(w) := (Y - f^m(W_n, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X) - h^{k,m}(w).$$

Remember that in all this setting, we are considering the different c_k's to be fixed. For the

$k-th$ neuron, we define the covariance matrix $\Sigma_{k,m}$

$$\Sigma_{k,m}(w) = \mathbb{E}_{X,Y}[\xi_{k,m}(w)\xi_{k,m}^T(w)] \in \mathbb{R}^{p\times p}$$

In summary, we'll approximate the SGD dynamics in continuous time for the k-th neuron by

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}, \tag{4.4}$$

where $(B_t^{k,m})$ and $(\tilde{B}_t^{k,m})$ are p-dimensional Brownian Motions. We remind the reader to check [17] for a rigorous derivation of this SDE.

4.1. Training Dynamics

Let $\mu_t^m \in \mathcal{M}(\mathbb{R}\times\mathbb{R}^p)$ be the empirical measure associated with the vectors $(\vec{c^m}, (W_t^m)_t)$, i.e

$$\mu_t^m = \frac{1}{m}\sum_{k=1}^m \delta_{(c_k, W_t^{k,m})}. \tag{4.5}$$

. We'd like to study the dynamics of the empirical measure μ_t^m when m diverges to infinity. Why would it be interesting to study the dynamics of μ_t^m in the limit? Because, given the results presented by Jacot et al. in [19], we hope that if the initialization of the parameters are independent centered Gaussians, then the limit of the empirical measure will be the initial measure, i.e we hope that in this case $\mu_t^m \xrightarrow{m\to\infty} \mu_0$ in some way since the NTK regime tells us that in the limit, the NTK stays frozen through training. We hope to generalize this results and check the limits in other settings and different training dynamics than the ones studied in [19].

To follow this study rigorously, we have to solve three previous steps:

1. Prove that the process $(\mu_t^m)_m$ is tight in some *good space* (Which will turn out to be the space of continuous path in the space of probability measures).
2. Prove the existence, and identify the limiting point as m diverges to infinity.
3. Prove the uniqueness of the limit.

We will extensively use

Lemma 4.1 (Itô's Lemma) *Let $X_t \in \mathbb{R}^p$ be a stochastic process and let $B_t \in \mathbb{R}^p$ a Brownian motion, with X_t adapted w.r.t the Brownian filtration. Assume X_t follows the Stochastic Differential Equation:*

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t,$$

with initial condition X_0 independent of B_s for all $s \geq 0$. For a function $f \in \mathcal{C}^2(\mathbb{R}^p)$, we have:

$$f(X_t) = f(X_0) + \int_0^t (\nabla_X f(X_s))^T b(s, X_s)ds + \frac{1}{2}\int_0^t \mathrm{Tr}\Big(\sigma(s, X_s)^T H_X f(X_s)\sigma(s, X_s)\Big)ds + \int_0^t (\nabla_X f(X_s))^T \sigma(s, X_s)dB_s.$$

Considering the dynamics for $W_t^{k,m}$, $k \in [m]$, given a function $\varphi(\cdot) \in \mathcal{C}^2(\mathbb{R}^p)$ we can apply Itô's lemma, which gives us:

$$\begin{aligned}\varphi(W_t^{k,m}) = \varphi(W_0^{k,m}) &+ \int_0^t \nabla\varphi(W_s^{k,m})^T h^{k,m}(W_s^m)ds + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\int_0^t \nabla\varphi(W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k \\ &+ \int_0^t \sqrt{2\tau}\nabla\varphi(W_s^{k,m})^T d\tilde{B}_s^{k,m} + \frac{\gamma}{2m^\alpha}\int_0^t \operatorname{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds \\ &+ \int_0^t 2\tau \operatorname{Tr}\left(H_w\varphi(W_s^{k,m})\right)ds. \end{aligned} \tag{4.6}$$

Hence, given $\varphi \in \mathcal{C}^{0,2}(\mathbb{R}\times\mathbb{R}^p)$, we can *test* $\mu_t^m = \frac{1}{m}\sum_{k=1}^m \delta_{(c_k, W_t^{k,m})}$ with φ, obtaining:

$$\langle \varphi, \mu_t^m\rangle = \frac{1}{m}\sum_{k=1}^m \varphi(c^k, W_t^{k,m})$$

and by replacing equation (5.1) in this formula, and considering that we are not training c and hence they are constant in time, we get (4.23).

$$\begin{aligned}\langle \varphi, \mu_t^m - \mu_0^m\rangle = &\frac{1}{m}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m)ds + \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^m\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k \\ &+ \frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^m\int_0^t \operatorname{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds \\ &+ \frac{1}{m}\sum_{k=1}^m\int_0^t \sqrt{2\tau}\nabla\varphi(c_k, W_s^{k,m})^T d\tilde{B}_s^{k,m} + \frac{1}{m}\sum_{k=1}^m\int_0^t 2\tau \operatorname{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds. \end{aligned} \tag{4.7}$$

4.1.1. Technical Lemmas

In order to prove tightness, we need some bounds and controls on $\Sigma_{k,m}$, for all $k \in \{1, \dots, \}$, and on $L(W^m)$, with $W^m \in (\mathbb{R}^p)^m$. Let's start by the latter.

Let $L(w) = \mathbb{E}_{X,Y}[(Y - f_m(W^m, X))]$. Then we have the following:

Lemma 4.2 *Let $m \in \mathbb{N}$ and $t \geq 0$. Then, for any $W \in (\mathbb{R}^p)^m$,*

$$\left|\frac{L(W)}{m}\right| \leq C\left(\frac{1}{m} + a_m\right),$$

with $a_m = \sum_{i=1}^m \frac{c_i^2}{m}$.

Proof. We'll use the fact that $\sigma(\cdot,\cdot)$ is bounded. In the first place, using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we get:

$$\left|\frac{L(W)}{m}\right| \leq \frac{1}{m}\left(2\mathbb{E}_{X,Y}[Y^2] + 2\mathbb{E}_{X,Y}[f_t^m(W,X)^2]\right),$$

where the expectation on the second term Recalling that we assume that Y has a finite second moment and σ is bounded, by using the definition of f_t^m we get:

$$\begin{aligned}\left|\frac{L(W)}{m}\right| &\leq \frac{1}{m}\left(C + \frac{1}{m}\sum_{i,j=1}^{m}|c_i||c_j||\mathbb{E}_{X,Y}[\sigma(W,X)\sigma(W,X)]|\right)\\ &\leq \frac{C}{m} + \frac{C}{m^2}\left(\sum_{i=1}^{m}|c_i|\right)^2, \qquad \text{because } \sigma \text{ is bounded.}\end{aligned}$$

Bounding $\left(\sum_{i=1}^m |c_i|\right)^2 \leq m\sum_{i=1}^m c_i^2$ by Cauchy-Schwarz inequality, and defining the quantity $a_m = \frac{1}{m}\sum_{i=1}^m c_i^2$, we conclude:

$$\left|\frac{L(W)}{m}\right| \leq C\left(\frac{1}{m} + a_m\right).$$

□

For Σ, we'll need a control over it's norm.

Lemma 4.3 *Let $m \in \mathbb{N}, k \in \{1,\dots,m\}$ and $s \geq 0$. Then; with $\|\cdot\|_{Frob}$ denoting the Frobenius norm of matrices, we have:*

$$\left\|\Sigma(W_s^{k,m})\right\|_{Frob} \leq C\frac{c_k^2}{m}L(W_s^m),$$

with C being a positive constant .

Recall that $a_m := \sum_{i=1}^m \frac{c_i^2}{m}$.

Proof. Before we start, it's important to notice that given $m \in \mathbb{N}$, $k \in \{1,\dots,m\}$, $w \in \mathbb{R}^p$ and $\Sigma_{k,m}$ has the following structure:

$$\begin{aligned}\Sigma_{k,m}(W^{m,k}) = \mathbb{E}_{X,Y}&\left[(Y - f^m(W^m,X))^2\frac{c_k^2}{m}\nabla\tilde{\sigma}(W_s^{k,m},X)\nabla\tilde{\sigma}(W_s^{k,m},X)^T\right]\\ &- \mathbb{E}_{X,Y}\left[(Y - f^m(W^m,X))\frac{c_k}{\sqrt{m}}\nabla\sigma(W_s^{k,m},X)\right]\mathbb{E}_{X,Y}\left[(Y - f^m(W^m,X))\frac{c_k}{\sqrt{m}}\nabla\sigma(W_s^{k,m},X)\right]^T\end{aligned}$$

and for $i,j \in \{1,\dots,p\}$:

$$\begin{aligned}\Sigma_{k,m}(w)_{i,j} = &\frac{c_k^2}{m}\mathbb{E}_{X,Y}\left[(Y - f^m(W^m,X))^2\partial_i\tilde{\sigma}(W_s^{k,m},X)\partial_j\tilde{\sigma}(W_s^{k,m},X)\right]\\ &- \frac{c_k^2}{m}\mathbb{E}_{X,Y}\left[(Y - f^m(W^m,X))\partial_i\tilde{\sigma}(W_s^{k,m},X)\right]\mathbb{E}_{X,Y}\left[(Y - f^m(W^m,X))\partial_j\tilde{\sigma}(W_s^{k,m},X)\right],\end{aligned}$$

Now, for the proof, let $u_t^m(x,y) = y - f^m(W_t^m,x)$. By definition

$$\left\|\Sigma(W_s^{k,m})\right\|_{\text{Frob}}^2 = \sum_{i,j}|\Sigma_{k,m}(W_s^{k,m})_{i,j}|^2.$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, and the fact that σ is a Lipschitz function and therefore it's derivative is bounded:

$$|\Sigma_{k,m}(W_s^{k,m})_{i,j}|^2 \leq C\frac{c_k^4}{m^2}(\mathbb{E}_{X,Y}[u_s^m(X,Y)^2] + \mathbb{E}_{X,Y}[u_s^m(X,Y)]^2)^2. \tag{4.8}$$

By applying Jensen's inequality in the second term of the right-hand side, we get:

$$|\Sigma_{k,m}(W_s^{k,m})_{i,j}|^2 \leq C\frac{c_k^4}{m^2}L(W_s^m)^2,$$

and this way:

$$\left\|\Sigma_{k,m}(W_s^{k,m})\right\|_{\text{Frob}}^2 = \sum_{i,j}|\Sigma_{k,m}(W_s^{k,m})_{i,j}|^2 \leq Cp^2\frac{c_k^4}{m^2}L(W_s^m)^2,$$

where C remains the same constant as the last equation. By applying the square root on both sides, we conclude:

$$\left\|\Sigma_{k,m}(W_s^{k,m})\right\|_{\text{Frob}} = \sqrt{\sum_{i,j}|\Sigma_{k,m}(W_s^{k,m})_{i,j}|^2} \leq C\frac{c_k^2}{m}L(W_s^m).$$

□

Lemma 4.3 tells us that if we have a control over the expectation of $L(W_s^m)$, then we can also control $\Sigma(W_s^m)$'s norm. That'll be useful in the future to prove the tightness of the process $(\mu_t^m)_m$.

Even though we already said that c's are fixed, we'll specify how are the initialized: From now on we consider

Assumption: The distribution that initializes all these coefficients is such that $\mathbb{E}[c]$ is finite and $\mathbb{E}[c^4]$ is finite.

A direct consequence of the uniform bound from Lemma 4.2 is the following:

Lemma 4.4 *For $m \in \mathbb{N}$, $t \geq 0$,*

$$\frac{\mathbb{E}[L(W_t^m)]}{m} \leq \mathbb{E}\left[C(\frac{1}{m} + a_m)\right] \leq C(1 + \underbrace{\mathbb{E}\left[a_m\right]}_{\leq C}) \leq C.$$

Since we are controlling the expectation of $L(W_S^m)$, we might as well think about doing the same with $f(W_s^m)$. We'll do this in the following:

In order to prove the next Lemma, we present a special type of Gronwall's inequality in the following:

Lemma 4.5 (A version of Gronwall's Lemma) *Let $x : [0, +\infty) \to \mathbb{R}$ be a locally absolutely*

continuous function, let $a, b \in L^1_{loc}([0,+\infty))$ *be given functions satisfying, for* $\lambda \in \mathbb{R}$,

$$\frac{d}{dt}x^2(t) + 2\lambda x^2(t) \leq a(t) + 2b(t)x(t), \text{for } \mathcal{L}^1 a.e. t > 0.$$

Then for every $T > 0$ *we have*

$$e^{\lambda T}|x(T)| \leq \left(x^2(0) + \sup_{t\in[0,T]} \int_0^t e^{2\lambda s} a(s)ds\right)^{\frac{1}{2}} + 2\int_0^T e^{\lambda t} b(t)dt.$$

Lemma 4.6 *Given* m, *and* $k \in \{1, .., m\}$*:*

$$\mathbb{E}[\|W_t^{k,m}\|^2]^{\frac{1}{2}} \leq \left(p + \frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}} + Ct.$$

Proof. Let $m \in \mathbb{N}$, and let $g : \mathbb{R}^p \to \mathbb{R}$, given by $g(w) = \|w\|^2$, for all $w \in \mathbb{R}^p$. Note that

$$\nabla g(w) = 2w \text{ and } \mathcal{H}_w g(w) = 2\mathcal{I}. \tag{4.9}$$

Recall that for each $k \in \{1, .., m\}$, we know that when $\tau = 0$, $(W_s^{k,m})_{s\geq 0}$ follows the SDE:

$$dW_s^{k,m} = h^{k,m}(W_s^{k,m})ds + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^{k,m}.$$

Then, by applying Itô's Lemma to g, for $W_s^{k,m}$, for $t > 0$ we get:

$$\begin{aligned} g(W_t^{k,m}) = g(W_0^{k,m}) &+ \int_0^t \nabla g(W_s^{k,m)})h^{k,m}(W_s^{k,m})ds + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\int_0^t \nabla g(W_s^{k,m)})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^{k,m} \\ &+ \frac{\gamma}{2m^\alpha}\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\mathcal{H}_w g(W_s^{k,m})\Big)ds. \end{aligned}$$

By replacing (4.9) in this equation, we get:

$$\begin{aligned} \|W_t^{k,m}\|^2 = \|W_0^{k,m}\|^2 &+ 2\int_0^t (W_s^{k,m)})^T h^{k,m}(W_s^{k,m})ds + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\int_0^t \nabla\varphi f(W_s^{k,m)})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^{k,m} \\ &+ \frac{\gamma}{m^\alpha}\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds. \end{aligned}$$

Now, if we apply expectation on both sides the local martingale term will be transformed to 0. This way :

$$\mathbb{E}[\|W_t^{k,m}\|^2] = \mathbb{E}[\|W_0^{k,m}\|^2] + 2\mathbb{E}\left[\int_0^t (W_s^{k,m)})^T h^{k,m}(W_s^{k,m})ds\right] + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right]. \tag{4.10}$$

Remember that, in the common SGD setting, $h^{k,m}(W_s k, m) = -\mathbb{E}_{X,Y}[(Y - f_s^m(W_s^m))\frac{c_k}{\sqrt{m}}\nabla\sigma(W_s^{k,m})]$.

We use this to bound (4.10):

$$\begin{aligned}
\mathbb{E}[\|W_t^{k,m}\|^2] &= \mathbb{E}[\|W_0^{k,m}\|^2] + 2\mathbb{E}\left[\int_0^t (W_s^{k,m)})^T h^{k,m}(W_s^{k,m})ds\right] + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right] \\
&= \mathbb{E}[\|W_0^{k,m}\|^2] - 2\mathbb{E}\left[\int_0^t (W_s^{k,m)})^T \mathbb{E}_{X,Y}[(Y - f_s^m(W_s^m))\frac{c_k}{\sqrt{m}}\nabla\sigma(W_s^{k,m})]ds\right] \\
&+ \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right] \\
&\overset{C-S}{\leq} \mathbb{E}[\|W_0^{k,m}\|^2] + 2\int_0^t \mathbb{E}\left[\|W_s^{k,m)}\|\mathbb{E}_{X,Y}[|Y - f_s^m(W_s^m)|\frac{|c_k|}{\sqrt{m}}\|\nabla\sigma(W_s^{k,m})\|]\right]ds \\
&+ \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right].
\end{aligned}$$

Recall that $\|\nabla\sigma\| \leq C$. By applying this:

$$\begin{aligned}
\mathbb{E}[\|W_t^{k,m}\|^2] &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}\left[\|W_s^{k,m)}\|L(W_s^{k,m})^{\frac{1}{2}}\frac{|c_k|}{\sqrt{m}}\right]ds + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right] \\
&\overset{C-S}{\leq} \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}}\mathbb{E}[L(W_s^{k,m})\frac{c_k^2}{m}]^{\frac{1}{2}}ds + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right] \\
&\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}}\mathbb{E}[L(W_s^{k,m})^2]^{\frac{1}{4}}\mathbb{E}[\frac{c_k^4}{m^2}]^{\frac{1}{4}}ds + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right]
\end{aligned}$$

Recall we assume that all c_k have finite 4th moment. Then, we can bound the expectation in our equation and get:

$$\begin{aligned}
\mathbb{E}[\|W_t^{k,m}\|^2] &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}}\frac{\mathbb{E}[L(W_s^{k,m})^2]^{\frac{1}{4}}}{\sqrt{m}}ds + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right] \\
&\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}}\frac{\mathbb{E}[L(W_s^{k,m})^2]^{\frac{1}{4}}}{\sqrt{m}}ds + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big)ds\right].
\end{aligned} \tag{4.11}$$

By Lemma 4.2,

$$\frac{L(W_s^m)}{m} \leq C(\frac{1}{m} + a_m),$$

where $a_m = \frac{1}{m}\sum_{k=1}^m c_k$. Then:

$$\frac{L(W_s^m)^2}{m^2} \leq C(\frac{1}{m^2} + a_m^2),$$

and by applying expectation:

$$\mathbb{E}[\frac{L(W_s^m)^2}{m^2}] \leq C(\frac{1}{m^2} + \mathbb{E}[a_m^2]).$$

Recall we assumed that all c_k's where independent and centered. This way $\mathbb{E}[a_m^2] < \infty$. By replacing this:

$$\mathbb{E}[\frac{L(W_s^m)^2}{m^2}] \leq C. \tag{4.12}$$

By replacing (4.12) on (4.11), we get:

$$\mathbb{E}[\|W_t^{k,m}\|^2] \leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}} ds + \frac{\gamma}{m^\alpha}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big) ds\right]. \quad (4.13)$$

Now let's bound the other term. By Lemma 4.3, we know that:

$$|\Sigma_{k,m}(W_s^{k,m})_{i,i}| \leq C\frac{c_k^2}{m} L(W_s^m).$$

Then:

$$\mathrm{Tr}\Big(\Sigma_{k,m}(W_s^{k,m})\Big) \leq C\frac{c_k^2}{m} L(W_s^m). \quad (4.14)$$

By replacing 4.14 in 4.13 and applying Fubini, we get:

$$\begin{aligned}\mathbb{E}[\|W_t^{k,m}\|^2] &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}} ds + \frac{C\gamma}{m^\alpha}\int_0^t \mathbb{E}[\frac{c_k^2}{m} L(W_s^m)] ds \\ &\overset{C-S}{\leq} \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}} ds + \frac{C\gamma}{m^\alpha}\int_0^t \mathbb{E}[c_k^4]^{\frac{1}{2}}\mathbb{E}[\frac{L(W_s^m)^2}{m^2}]^{\frac{1}{2}} ds.\end{aligned}$$

By replacing (4.12), and using that the c_k's have finite 4th moment:

$$\begin{aligned}\mathbb{E}[\|W_t^{k,m}\|^2] &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}} ds + \frac{C\gamma}{m^\alpha}\int_0^t \mathbb{E}[\frac{c_k^2}{m} L(W_s^m)] ds \\ &\overset{C-S}{\leq} \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}} ds + \frac{C\gamma}{m^\alpha}\int_0^t ds \\ &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}[\|W_s^{k,m)}\|^2]^{\frac{1}{2}} ds + \frac{C\gamma t}{m^\alpha}.\end{aligned}$$

Since both sides are equal at $t = 0$, we apply the Radon-Nikodym derivative on both sides and obtain:

$$\frac{d}{dt}\mathbb{E}[\|W_t^{k,m}\|^2] \leq C\mathbb{E}[\|W_t^{k,m)}\|^2]^{\frac{1}{2}} + \frac{C\gamma}{m^\alpha}. \quad (4.15)$$

Next, we apply Lemma 4.5 with $\lambda = 0$, $a(t) = \dfrac{C\gamma}{m^\alpha}$ and $b(t) = C$. Then, for every $t \geq 0$

$$\mathbb{E}[\|W_t^{k,m}\|^2]^{\frac{1}{2}} \leq \left(\mathbb{E}[\|W_0^{k,m)}\|^2] + \frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}} + Ct. \quad (4.16)$$

Recall that each element of $W_0^{k,m}$ has a finite expectation by our assumptions, and that $W^{k,m} \in \mathbb{R}^p$. Therefore:

$$\mathbb{E}[\|W_0^{k,m}\|] = p.$$

With this, we conclude:

$$\mathbb{E}[\|W_t^{k,m}\|^2]^{\frac{1}{2}} \leq \left(p + \frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}} + Ct. \quad (4.17)$$

□

Having all of these technical results we continue to the next section, where we'll prove the ex of solutions for the Stochastic Differential Equations we have. Next, we'll prove the

tension of the laws of the empirical measure process.

4.2. Existence of Solutions for the SDE

Let $m \in \mathbb{N}$ be the quantity of neurons we have in our neural network, whose parameters are initialized such that $W_0^{k,m}$ has finite expectation for $k \in \{1, .., m\}$ i.i.d, and c_k i.i.d such that it's first four moments are finite. For each $k \in \{1, .., m\}$ and $t \geq 0$ we have:

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}, \tag{4.18}$$

where $(B_t^{k,m})$ and $(\tilde{B}_t^{k,m})$ are p-dimensional Brownian Motions,

$$h^{k,m}(W_n^m) := -\lambda W_n^{k,m} + \mathbb{E}_{X,Y}\left[(Y - f^m(W_n^m, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X)\right],$$

and for

$$\xi_{k,m}(w) := (Y - f^m(W_n, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X) - h^{k,m}(w),$$

we defined:

$$\Sigma_{k,m}(w) = \mathbb{E}_{X,Y}[\xi_{k,m}(w)\xi_{k,m}^T(w)] \in \mathbb{R}^{p\times p}.$$

Theorem 4.1 *Let $m \in \mathbb{N}$ and $(c_k)_{k=1}^m$ with the properties considered above. Let σ a bounded function, whose gradient and hessian have a bounded norm. Then, equation* (4.18) *has strong solutions, subject to the fixed coefficients $(c_k)_{k=1}^m$.*

This Theorem will allow us to make our study without being uncertain on the existence of solutions to our Stochastic Differential Equation. It's proof will go back to proving that both coefficients are Lipschitz. For the existence and uniqueness Theorem of solutions for Stochastic Differential Equations, we refer the reader to Karatzas and Shreve's book [26].

The reader may also note the fact that we are proving existence of solutions **given the parameters** $(c_k)_{k=1}^m$. This means that, even though they are initialized randomly, by being independent on the rest of the parameters, we can assure the existence of solutions. Note, however, that this doesn't mean that the coefficients are not random variables.

Proof. let $m \in \mathbb{N}$. We must prove that for all $k \in \{1, ..., m\}$, the equation

$$dW_t^{k,m} = h^{k,m}(W_t^{k,m})dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m},$$

has strong solutions. The reader may note that each coefficient depends both on the particle itself, but also on the other particles in the system (i.e the rest of the parameters. To surpass this difficulty, we'll prove that for each $k \in \{1, ..., m\}$, given $w_1, w_2 \in \mathbb{R}^p$, we have

$$\|h^{k,m}(w_1) - h^{k,m}(w_2)\| + \|\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(w_1) - \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(w_2)\| \leq C\|W_1 - W_2\|,$$

i.e that both coefficients are Lipschitz in W, which is the matrix of weights that contains w in it's k-th row. Having this, we can make sure that the same can be concluded for the SDE

that rule $W^m \in (\mathbb{R}^p)^m$ because it will be a vectored version of the dynamics for each k. Let $w_1, w_2 \in \mathbb{R}^p$ and $k \in \{1, ..., m\}$. We begin by proving that $h^{k,m}$ is Lipschitz for W. We have:

$$\begin{aligned}\|h^{k,m}(w_1) - h^{k,m}(w_2)\| = \Bigg\|&\lambda(w_1 - w_2) + \mathbb{E}_{X,Y}\left[(Y - f^m(W_1, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(w_1, X)\right] \\ &-\mathbb{E}_{X,Y}\left[(Y - f^m(W_2, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(w_2, X)\right]\Bigg\| \\ &\|\lambda(w_1 - w_2)\| + \Bigg\|\mathbb{E}_{X,Y}\left[(Y - f^m(W_1, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(w_1, X)\right] \\ &-\mathbb{E}_{X,Y}\left[(Y - f^m(W_2, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(w_2, X)\right]\Bigg\|.\end{aligned}$$

The first term is already Lipschitz. Hence, we focus on the case $\lambda = 0$. By adding and subtracting $\mathbb{E}_{X,Y}\left[(Y - f^m(W_1, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(w_2, X)\right]$:

$$\begin{aligned}\|h^{k,m}(w_1) - h^{k,m}(w_2)\| &\leq \left\|\mathbb{E}_{X,Y}\left[(f^m(W_1, X) - f^m(W_2, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(w_2, X)\right]\right\| \\ &+ \left\|\mathbb{E}_{X,Y}\left[(Y - f^m(W_1, X))\frac{c_k}{\sqrt{m}}(\nabla_w\sigma(w_2, X) - \nabla_w\sigma(w_1, X))\right]\right\| \\ &\leq \mathbb{E}_{X,Y}\left[|f^m(W_1, X) - f^m(W_2, X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(w_2, X)\|\right] \\ &+ \mathbb{E}_{X,Y}\left[|Y - f^m(W_1, X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(w_2, X) - \nabla_w\sigma(w_1, X)\|\right] \\ &\leq \mathbb{E}_{X,Y}\left[|\frac{1}{m}\sum_{i=1}^{m} c_i(\sigma(W_1^i, X) - \sigma(W_i^k, X)||c_k|\|\nabla_w\sigma(w_2, X)\|\right] \\ &+ \mathbb{E}_{X,Y}\left[|Y - f^m(W_1, X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(w_2, X) - \nabla_w\sigma(w_1, X)\|\right].\end{aligned}$$

Recall that in our hypothesis, σ is Lipschitz and bounded, with bounded gradient. Therefore

$$\begin{aligned}
\|h^{k,m}(w_1) - h^{k,m}(w_2)\| &\leq C\mathbb{E}_{X,Y}\left[\frac{1}{m}\sum_{i=1}^{m}|c_i||\sigma(W_1^i, X) - \sigma(W_2^i, X)||c_k|\right] \\
&+ \mathbb{E}_{X,Y}\left[|Y - f^m(W_1, X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(w_2, X) - \nabla_w\sigma(w_1, X)\|\right] \\
&\leq C\mathbb{E}_{X,Y}\left[\frac{1}{m}\sum_{i=1}^{m}|c_i|\|W_1^i - W_2^i\||c_k|\right] \\
&+ C\mathbb{E}_{X,Y}\left[|\frac{Y}{\sqrt{m}} - \frac{1}{m}\sum_{i=1}^{m}c_i\sigma(W_1^i, X)||c_k|\|w_2 - w_1\|\right] \\
&\leq C\frac{1}{m}\sum_{i=1}^{m}|c_i|\|W_1^i - W_2^i\||c_k| \\
&+ C\mathbb{E}_{X,Y}\left[\left(\frac{|Y|}{\sqrt{m}} + \frac{1}{m}\sum_{i=1}^{m}|c_i||\sigma(W_1^i, X)|\right)|c_k|\|w_2 - w_1\|\right].
\end{aligned}$$

Note that $\|W_1^i - W_2^i\| \leq \max_{1\leq i\leq m}\|W_1^i - W_2^i\| \leq C\|W_1 - W_2\|$, because all norms in a finite space are equivalent. On the other hand, by using that the activation function is bounded, we obtain:

$$\|h^{k,m}(w_1) - h^{k,m}(w_2)\| \leq C\left(\frac{1}{m}\sum_{i=1}^{m}|c_i||c_k|\right)\|W_1 - W_2\| + C\mathbb{E}_{X,Y}[\left(\frac{|Y|}{\sqrt{m}} + \frac{1}{m}\sum_{i=1}^{m}|c_i|\right)|c_k|\|W_2 - W_1\|].$$

Since in our setting, Y has finite second moment, we obtain that:

$$\|h^{k,m}(w_1) - h^{k,m}(w_2)\| \leq C\left(\frac{1}{m}\sum_{i=1}^{m}|c_i||c_k| + C\left(C + \frac{1}{m}\sum_{i=1}^{m}|c_i|\right)|c_k|\right)\|W_2 - W_1\|. \quad (4.19)$$

With this, we conclude that $h^{k,m}$ is Lipschitz for fixed coefficients $(c_k)_{k=1}^m$. The full proof that $\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})$ can be derived from the arguments used in the existence and uniqueness results for the Mean Field SDE, and it uses results from Stroock and Varadhan's book, [27]. With this, we conclude the existence of solutions for the SDE for fixed coefficients $(c_k)_{k=1}^m$. □

Having proved this results, we can now be relaxed about the existence of the objects we are studying, which is not trivial. Being sure of this, we can proceed to prove the tightness of our process.

4.3. Tightness of the laws of the empirical measure process

To prove tightness, there are two big steps. We'll separate them in two different sections, each concerning one of the steps.

4.3.1. First Part of the Proof

To prove tightness of the laws of the process μ_t^m in some *good space*, we'll start by proving tightness of a different -but related- object.

Lemma 4.7 *Given $\varphi \in \mathcal{C}_0(\mathbb{R} \times \mathbb{R}^p)$, the process $(\langle \varphi, \mu_t^m \rangle)_t$ is tight.*

The demonstration of Lemma 4.7, relies on Aldous Criterion, which can be found in [28].

Lemma 4.8 (Aldous' Criterion) *For every $n \in \mathbb{N}$, let $(X_t^n)_t$ be a cadlag process on a filtered probability space $(\Sigma, \mathcal{F}, (\mathcal{F}_t)_t, \mathbb{P})$. We suppose that the process satisfies:*

1. *For every $N \in \mathbb{N}$ and for all $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ and $K > 0$ such that*

$$n \geq n_0 \implies \mathbb{P}\left(\sup_{t \leq N} |X_t^n| > K\right) \leq \varepsilon.$$

2. *For every $N \in \mathbb{N}$ and for all $\varepsilon > 0$*

$$\lim_{\theta \downarrow 0} \limsup_n \sup_{S,T \in \mathcal{T}_N : S \leq T \leq S+\theta} \mathbb{P}(|X_T^n - X_S^n| \geq \varepsilon) = 0, \tag{4.20}$$

where $\mathcal{T}_N$ are all the stopping times in $(\mathcal{F}_t)_t$ that are bounded by N.

Then the process' probability laws are tight.

Remark For the first condition, by Markov's Inequality it is sufficient to bound the expectation of the random variable at time N.

We'll prove part of Lemma 4.10 using this remark.

Another important criterion will be the Aldous - Rebolledo Criterion, which works better for semi-martingales. For a full proof and statement we refer the reader to [29], and for a discussion about it, to [30]. We present it's statement in the following:

Lemma 4.9 *(Aldous - Rebolledo Criterion). Let $(X_t^n)_{t \geq 0, n \geq 0}$ be a sequence of càdlàg square integrable semi-martingales. Let us write the decomposition $X_t^n = A_t^n + M_t^n$, where $(M_t^n)_{t \geq 0}$ is a local square integrable martingale and $(A_t^n)_{t \geq 0}$ is an adapted finite variation paths process. If the two following conditions are fulfilled, then the sequences of processes $(M_t^n)_{t \geq 0}$, $(\langle M^n \rangle_t)_{t \geq 0}$ and $(X_t^n)_t$ are tight.*

1. *For every t within a dense subset of $\mathbb{R}_+$, $(M_t^n)_{t \geq 0}$ and $(A_t^n)_{t \geq 0}$ are tight sequences.*
2. *Both processes $(\langle M^n \rangle_t)_{t \geq 0}$ and $(A_t^n)_{t \geq 0}$ satisfy Aldous' Criterion.*

Remark When X_t^n is a continuous semi-martingale, (2) implies (1).

With both tightness criterions in our minds, we are ready to state the following:

Lemma 4.10 *Given $\varphi \in \mathcal{C}_0^\infty(\mathbb{R})$, the finite variation and the quadratic variation of the martingale parts of the process $\langle \varphi, \mu_t^m \rangle$ satisfy the first condition of Aldous Criterion.*

Let's recall the dynamics we found for the empirical measure in Equation in equation

(4.23):

$$\langle \varphi, \mu_t^m - \mu_0^m \rangle = \underbrace{\frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m)ds}_{(1)} - \underbrace{\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k}_{(2)}$$
$$+ \underbrace{\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds}_{(3)}.$$
$$+ \underbrace{\frac{1}{m}\sum_{k=1}^{m}\int_0^t \sqrt{2\tau}\nabla\varphi(c_k, W_s^{k,m})^T d\tilde{B}_s^{k,m}}_{(4)} + \underbrace{\frac{1}{m}\sum_{k=1}^{m}\int_0^t 2\tau\,\mathrm{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds}_{(5)}$$

The finite variation part of $\langle \varphi, \mu_t^m \rangle$ corresponds to:

$$A_t^n = (1) + (3) + (5).$$

On the other hand, the local martingale part corresponds to:

$$M_t^n = (2) + (4).$$

With these definitions, we go for the proof of tightness.

Proof.

Let's begin with A_t^n. Recall we want to prove the first condition of Aldous' criterion, that is, we want to prove that $\mathbb{E}[\sup_{t\leq N} A_t^n]$ is finite. For this, it's sufficient to show that the expectation of the modules of each term, (1), (3) and (5) are finite.

Let's start with (1). We begin by bounding the norm of φ, $\nabla\varphi$ and $\nabla\sigma$, since we are assuming they are all bounded. Also, we use some classical inequalities for the module and the expectation. We have:

$$\begin{aligned}
\mathbb{E}[\sup_{t\leq N}|(1)|] &= \mathbb{E}\left[\sup_{t\leq N}\left|\frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(W_s^{k,m})^T h^{k,m}(W_s^m)ds\right|\right] \\
&\leq \mathbb{E}\left[\sup_{t\leq N}\frac{1}{m}\sum_{k=1}^{m}\int_0^t \|\nabla\varphi(W_s^{k,m})\|\|h^{k,m}(W_s^m)\|ds\right] \\
&\leq C\mathbb{E}\left[\sup_{t\leq N}\frac{1}{m}\sum_{k=1}^{m}\int_0^t \left\|\mathbb{E}_{X,Y}\left[(Y - f^m(W_s^m, X))\frac{c_k}{\sqrt{m}}\nabla_w\sigma(W_n^{k,m}, X)\right]\right\| ds\right. \\
&\left. + \frac{1}{m}\sum_{k=1}^{m}\int_0^t \|W_s^{k,m}\|ds\right] \\
&\leq C\mathbb{E}\left[\sup_{t\leq N}\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}_{X,Y}\left[|Y - f^m(W_s^m, X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(W_n^{k,m}, X)\|\right] ds\right. \\
&\left. + \frac{\lambda}{m}\sum_{k=1}^{m}\int_0^t \|W_s^{k,m}\|ds\right].
\end{aligned}$$

Now, we separate the expectation and obtain:

$$\begin{aligned}\mathbb{E}[\sup_{t\leq N}|(1)|] &\leq C\mathbb{E}\left[\sup_{t\leq N}\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}_{X,Y}\left[|Y-f^m(W_s^m,X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(W_n^{k,m},X)\|\right]ds\right]\\ &+\mathbb{E}\left[\sup_{t\leq N}\frac{\lambda}{m}\sum_{k=1}^{m}\int_0^t\|W_s^{k,m}\|ds\right]\\ &\leq C\mathbb{E}\left[\sup_{t\leq N}\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}_{X,Y}\left[|Y-f^m(W_s^m,X)|\frac{|c_k|}{\sqrt{m}}\|\nabla_w\sigma(W_n^{k,m},X)\|\right]ds\right]\\ &+\frac{\lambda}{m}\sum_{k=1}^{m}\sup_{t\leq N}\int_0^t\mathbb{E}\left[\|W_s^{k,m}\|\right]ds\end{aligned}$$

By using Cauchy-Schwarz inequality plus the fact that the gradient of σ has bounded norm (since σ is a Lipschitz function) we obtain:

$$\begin{aligned}\mathbb{E}_{X,Y}\left[|Y-f^m(W_s^m,X)|\frac{|c_k|}{\sqrt{m}}\nabla_w\|\sigma(W_n^{k,m},X)\|\right] &\leq \frac{C|c_k|}{\sqrt{m}}\mathbb{E}_{X,Y}\left[|Y-f^m(W_s^m,X)|\right]\\ &\leq \frac{C|c_k|}{\sqrt{m}}\mathbb{E}_{X,Y}\left[(Y-f^m(W_s^m,X))^2\right]^{\frac{1}{2}} \quad \text{By Jensen's}\\ &\leq \frac{C|c_k|}{\sqrt{m}}L(W_s^m)^{\frac{1}{2}}.\end{aligned}$$

On the other hand, by Lemma 4.6, we know that:

$$\begin{aligned}\frac{\lambda}{m}\sum_{k=1}^{m}\sup_{t\leq N}\int_0^t\mathbb{E}\left[\|W_s^{k,m}\|\right]ds &\leq \frac{\lambda}{m}\sum_{k=1}^{m}\sup_{t\leq N}\int_0^t\mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}}ds\\ &\leq \frac{\lambda}{m}\sum_{k=1}^{m}\sup_{t\leq N}\int_0^t\left(\left(p+\frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}}+Ct\right)dt\\ &\leq \frac{\lambda}{m}\sum_{k=1}^{m}\int_0^N\left(\left(p+\frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}}+Ct\right)^{\frac{1}{2}}dt\\ &\leq \frac{\lambda}{m}\sum_{k=1}^{m}N\left(\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}}+CN\right)\\ &\leq \lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}}+C\lambda N^2\end{aligned}$$

Replacing this in the last inequality:

$$\mathbb{E}[\sup_{t\leq N}|(1)|]\leq C\mathbb{E}\left[\sup_{t\leq N}\frac{1}{m}\sum_{k=1}^{m}\int_0^t\frac{C|c_k|}{\sqrt{m}}L(W_s^m)^{\frac{1}{2}}ds\right]+\lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}}+C\lambda N^2,$$

and since the inside of the integral is positive, we can erase the dependence on the supreme

and obtain:

$$\mathbb{E}[\sup_{t\leq N}|(1)|] \leq C\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^N \frac{C|c_k|}{\sqrt{m}}L(W_s^m)^{\frac{1}{2}}ds\right] + \lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + C\lambda N^2,$$
$$\leq C\mathbb{E}\left[\int_0^N \frac{C|c_k|}{\sqrt{m}}L(W_s^m)^{\frac{1}{2}}ds\right] + \lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + C\lambda N^2,.$$

By using Fubini and Cauchy-Schwarz:

$$\mathbb{E}[\sup_{t\leq N}|(1)|] \leq C\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^N \frac{C|c_k|}{\sqrt{m}}L(W_s^m)^{\frac{1}{2}}ds\right] + \lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + C\lambda N^2,$$
$$\leq C\mathbb{E}[|c_k|^2]^{\frac{1}{2}}\int_0^N \frac{\mathbb{E}\left[L(W_s^m)^{\frac{1}{2}}\right]^2}{m^{\frac{1}{2}}}ds + \lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + C\lambda N^2.$$

We are assuming that $\mathbb{E}[c_k^2] < \infty$, so we get:

$$\mathbb{E}[\sup_{t\leq N}|(1)|] \leq C\int_0^N \frac{\mathbb{E}\left[L(W_s^m)^{\frac{1}{2}}\right]^2}{m^{\frac{1}{2}}}ds + \lambda N\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + C\lambda N^2.$$

By using Lemma 4.4, which gives us a bound over $\frac{\mathbb{E}[L(W_s^m)]}{m}$ we can take the square root and obtain the following inequality for all $\alpha \geq 0$:

$$\mathbb{E}\left[\sup_{t\leq N}\left|\frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(W_s^{k,m})^T h^{k,m}(W_s^m)ds\right|\right] \leq C.$$

Thus, the expectation of the term (1) is finite. Next, let's see that (3) is finite. We need to bound:

$$\mathbb{E}\left[\sup_{t\leq N}|(3)|\right] = \mathbb{E}\left[\sup_{t\leq N}\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds\right|\right]. \tag{4.21}$$

Given the fact that both $H_w\varphi$ and $\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m}$ are symmetric matrices, the trace of (4.21) satisfies:

$$\mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right) = \mathrm{Tr}\left(\Sigma_{k,m}(W_s^{k,m})H_w\varphi(W_s^{k,m})\right),$$

and by applying the Cauchy-Schwarz type inequality for the trace:

$$\left|\mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)\right| \leq \|\Sigma_{k,m}(W_s^{k,m})\|_{Frob}\|H_w\varphi(W_s^{k,m})\|_{Frob},$$

and since the second derivative of the test function φ is bounded, we get:

$$\left|\mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)\right| \leq C\|\Sigma_{k,m}(W_s^{k,m})\|_{Frob}.$$

Replacing this in the right hand side of equation (4.21):

$$\begin{aligned}\mathbb{E}\left[\sup_{t\leq N}|(3)|\right] &\leq C\mathbb{E}\left[\sup_{t\leq N}\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \|\Sigma_{k,m}(W_s^{k,m})\|_{\text{Frob}}\, ds\right|\right] \\ &\leq C\mathbb{E}\left[\sup_{t\leq N}\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \frac{c_k^2}{m}L(W_s^m)ds\right|\right] \qquad \text{by Lemma 4.3}\end{aligned}$$

Finally, by applying Fubini, we get:

$$\mathbb{E}\left[\sup_{t\leq N}|(3)|\right] \leq C\frac{\gamma}{2m^{1+\alpha}}\int_0^N \sum_{k=1}^{m}\mathbb{E}\left[\frac{c_k^2}{m}L(W_s^m)\right]ds$$

By Cauchy-Schwarz's inequality, and because we assume that c_i has a bounded second moment, we have that:

$$\mathbb{E}\left[\frac{c_k^2}{m}L(W_s^m)\right] \leq \frac{C}{m}\mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}.$$

Then:

$$\begin{aligned}\mathbb{E}\left[\sup_{t\leq N}|(3)|\right] &\leq C\frac{\gamma}{2m^{1+\alpha}}\int_0^N \mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}} ds \\ &\leq C\frac{\gamma}{2m^{\alpha}}\int_0^N \frac{\mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}}{m}ds \\ &\leq C\frac{\gamma N}{2m^{\alpha}} \qquad \text{by Lemma 4.4,}\end{aligned}$$

which is bounded for every $\alpha \geq 0$. For (5), we do just what we did with (3), that is, we bound the norm of the Hessian matrix that appears in the matrix. With this, we conclude that the finite variation part of the semi-martingale A_t^n satisfies the first condition of Aldous' Criterion.

For the local martingale term, Rebolledo's Criterion (Lemma 4.8) tells us we must prove the first part of Aldous Criterion for the quadratic variation for M_t^n.

We begin by calculating the quadratic variation of (2). We get:

$$\langle(2)\rangle = \left\langle \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^{m}\int_0^N \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle.$$

By using the independence of each Brownian motion and the symmetry of the matrices $\Sigma_{k,m}^{\frac{1}{2}}$ for each k, we get:

$$\left\langle\sum_{k=1}^{m}\int_0^N \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle = \int_0^N\sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p}\partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds. \tag{4.22}$$

Hence, We must prove:

$$\mathbb{E}\left[\sup_{t\leq N}\int_0^N \sum_{k=1}^m \sum_{i_1,i_2=1}^p \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds\right] < \infty.$$

Now, by replacing and bounding the derivatives of the function φ:

$$\begin{aligned}
\mathbb{E}[\sup_{t\leq N}\langle(2)\rangle] &= \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\mathbb{E}\left[\int_0^N \sum_{k=1}^m \sum_{i_1,i_2=1}^p \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds\right] \\
&\leq C\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\mathbb{E}\left[\int_0^N \sum_{k=1}^m \sum_{i_1,i_2=1}^p |\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}|ds\right] \\
&\leq Cp^{\frac{1}{2}}C\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\mathbb{E}\left[\int_0^N \sum_{k=1}^m \|\Sigma_{k,m}(W_s^{k,m})\|_{\text{Frob}}ds\right]
\end{aligned}$$

where we used Cauchy- Schwarz in the last inequality. By applying Lemma 4.3, which tells us that

$$\|\Sigma_{k,m}(W_s^{k,m})\|_{\text{Frob}} \leq \frac{c_k^2}{m}L(W_s^m),$$

we get:

$$\begin{aligned}
\mathbb{E}[\sup_{t\leq N}\langle(2)\rangle] &\leq Cp^{\frac{1}{2}}\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\mathbb{E}\left[\int_0^N \sum_{k=1}^m \frac{c_k^2}{m}L(W_s^m)ds\right] \\
&\leq Cp^{\frac{1}{2}}\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\mathbb{E}\left[\int_0^N \sum_{k=1}^m \frac{c_k^2}{m}L(W_s^m)ds\right] && \text{By Jensen's} \\
&\leq Cp^{\frac{1}{2}}\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\int_0^N \mathbb{E}\left[\sum_{k=1}^m \frac{c_k^2}{m}L(W_s^m)\right]ds.
\end{aligned}$$

Now, by considering our assumptions on c_k, we have

$$\begin{aligned}
\sum_{k=1}^m \mathbb{E}\left[\frac{c_k^2}{m}L(W_s^m)\right] &\leq \sum_{k=1}^m \frac{1}{m}\mathbb{E}[c_k^4]^{\frac{1}{2}}\mathbb{E}[L(W_s^m)^2]^{\frac{1}{2}} && \text{by C-S} \\
&\leq \frac{C}{m}\sum_{k=1}^m \mathbb{E}[L(W_s^m)^2]^{\frac{1}{2}} && \text{because } \mathbb{E}[c_k^2] \text{ is bounded} \\
&= C\mathbb{E}[L(W_s^m)^2]^{\frac{1}{2}}.
\end{aligned}$$

Replacing:

$$\begin{aligned}
\mathbb{E}[\sup_{t\leq N}\langle(2)\rangle] &\leq Cp^{\frac{1}{2}}\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\int_0^N C\mathbb{E}[L(W_s^m)^2]^{\frac{1}{2}}ds \\
&\leq Cp^{\frac{1}{2}}\frac{\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}}\int_0^N \left(\frac{\mathbb{E}[L(W_s^m)^2]}{m}\right)^{\frac{1}{2}}ds \\
&\leq CNp^{\frac{1}{2}}\frac{\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}} && \text{by Lemma 4.4.}
\end{aligned}$$

Therefore, we conclude that (2) is finite. to prove the same for (4), it's enough to use a direct extension of this argument. □

To finish this section, we'll prove the following Lemma, which will get us one step closer to prove tightness of the laws of the process.

Lemma 4.11 *Given $\varphi \in \mathcal{C}_0^\infty(\mathbb{R}\times\mathbb{R}^p)$, the random variables A_t^n and $\langle M_t^n\rangle$, i.e the finite variation and the quadratic variation of the martingale term of the process $\langle \varphi, \mu_t^m\rangle$ satisfy the second condition in Aldous Criterion for all $m \in \mathbb{N}$.*

Before the proof, let's recall the dynamics we found for the empirical measure in Equation in equation (4.23):

$$\begin{aligned}\langle \varphi, \mu_t^m - \mu_0^m\rangle = &\underbrace{\frac{1}{m}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m)ds}_{(1)} - \underbrace{\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k}_{(2)}\\ &+ \underbrace{\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^m \int_0^t \operatorname{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds}_{(3)}\\ &+ \underbrace{\frac{1}{m}\sum_{k=1}^m \int_0^t \sqrt{2\tau}\nabla\varphi(c_k, W_s^{k,m})^T d\tilde{B}_s^{k,m}}_{(4)} + \underbrace{\frac{1}{m}\sum_{k=1}^m \int_0^t 2\tau \operatorname{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds}_{(5)}\end{aligned} \tag{4.23}$$

The finite variation part of $\langle \varphi, \mu_t^m\rangle$ corresponds to:

$$\begin{aligned}A_t^n = &\frac{1}{m}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m)ds + \frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^m \int_0^t \operatorname{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right.\\ &+ \frac{1}{m}\sum_{k=1}^m \int_0^t 2\tau \operatorname{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds\end{aligned}$$

On the other hand, the martingale part corresponds to:

$$M_t^n = -\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k + \frac{1}{m}\sum_{k=1}^m \int_0^t \sqrt{2\tau}\nabla\varphi(c_k, W_s^{k,m})^T d\tilde{B}_s^{k,m}.$$

Proof. Let $\theta > 0, n \in \mathbb{N}$, and let $S, T \in \mathcal{T}_N$ such that $S \le T \le S+\theta$. Then by Markov's inequality, for any process X_t^m:

$$\mathbb{P}(|\mathcal{X}_t^m| \ge \varepsilon) \le \frac{1}{\varepsilon}\mathbb{E}\left[|X_t^m|\right]. \tag{4.24}$$

We must prove this for A_t^m and for $\langle M^m \rangle_t$. For A_t^n:

$$\sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[A_t^n\right] \leq \underbrace{\sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\left|\frac{1}{m}\sum_{k=1}^{m}\int_S^T \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m)ds\right|\right]}_{(1)}$$
$$+ \underbrace{\mathbb{E}\left[\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_S^T \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds\right|\right]}_{(2)}$$
$$+ \underbrace{\mathbb{E}\left[\left|\frac{1}{m}\sum_{k=1}^{m}\int_0^t 2\tau\,\mathrm{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds\right|\right]}_{(3)}.$$

Hence, we must prove that (1), (2) and (3) are finite. On the other hand, for $\langle M_t^n \rangle$, we must prove:

$$\sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\langle M_t^n\rangle\right] \leq \underbrace{\sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\int_0^N \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds\right]}_{(4)}$$
$$+ \underbrace{\mathbb{E}\left[\int_0^N \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})ds\right]}_{(5)}.$$

Therefore, we must also prove that (4) and (5) are finite.

The idea behind proving all these terms are finite, will be to find bounds which can allow us to remove the supremum, but keeping θ on each term, allowing us to take $\theta \to 0$ afterwards.

We'll begin with (1). We start by taking the module of the integral and bounding by the norm of $\nabla\varphi$:

$$\begin{aligned}(1) &= \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\left|\frac{1}{m}\sum_{k=1}^{m}\int_S^T \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m)ds\right|\right] \\ &\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\frac{1}{m^{1+\frac{1}{2}}}\sum_{k=1}^{m}\int_S^T \mathbb{E}_{X,Y}\left[|Y - f^m(W^m, X)||c_k|\|\nabla\sigma(W^{k,m}, X)\|\right]ds\right] \\ &+ C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_S^T \|W_s^{k,m}\|ds\right]. \end{aligned} \tag{4.25}$$

On the other hand, by Lemma 4.6, we know that:

$$\begin{aligned}
\frac{1}{m}\sum_{k=1}^{m}\int_S^T \mathbb{E}\left[\|W_s^{k,m}\|\right] ds &\leq \frac{1}{m}\sum_{k=1}^{m}\int_S^T \mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}} ds \\
&\leq \frac{1}{m}\sum_{k=1}^{m}\int_S^T \left(\left(p+\frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}} + Ct\right) \\
&\leq \frac{1}{m}\sum_{k=1}^{m}\int_S^T \left(\left(p+\frac{C\gamma t}{m^\alpha}\right)^{\frac{1}{2}} + Ct\right) dt \\
&\leq \frac{}{m}\sum_{k=1}^{m}(T-S)\left(\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + CN\right) \\
&\leq (T-S)\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + CN^2(T-S),
\end{aligned}$$

therefore,

$$\frac{1}{m}\sum_{k=1}^{m}\int_S^T \mathbb{E}\left[\|W_s^{k,m}\|\right] ds \leq (T-S)\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + CN^2(T-S). \tag{4.26}$$

By replacing (4.26) in (4.25):

$$\begin{aligned}
(1) &\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\frac{1}{m^{1+\frac{1}{2}}}\sum_{k=1}^{m}\int_S^T \mathbb{E}_{X,Y}\left[|Y-f^m(W^m,X)||c_k|\|\nabla\sigma(W^{k,m},X)\|\right] ds\right] \\
&\quad + C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} (T-S)\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + CN^2(T-S) \\
&\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\frac{1}{m^{1+\frac{1}{2}}}\sum_{k=1}^{m}\int_S^T \mathbb{E}_{X,Y}\left[|Y-f^m(W^m,X)||c_k|\|\nabla\sigma(W^{k,m},X)\|\right] ds\right] \\
&\quad + \theta\left(p+\frac{C\gamma N}{m^\alpha}\right)^{\frac{1}{2}} + CN^2\theta
\end{aligned}$$

Now, with the other term, by using that $\|\nabla\sigma\|$ is bounded, we have:

$$
\begin{aligned}
(1) &\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\frac{1}{m^{1+\frac{1}{2}}}\sum_{k=1}^{m}\int_S^T \mathbb{E}_{X,Y}\left[|c_k||Y-f^m(W^m,X)|\right]ds\right] + C\theta \\
&\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[\frac{1}{m^{1+\frac{1}{2}}}\sum_{k=1}^{m}\int_S^T |c_k| L(W_s^m)^{\frac{1}{2}}ds\right] + C\theta && \text{by C-S} \\
&\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \int_S^T \sum_{k=1}^{m}\frac{1}{m^{1+\frac{1}{2}}}\mathbb{E}\left[|c_k| L(W_s^m)^{\frac{1}{2}}\right]ds + C\theta \\
&\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \int_S^T \sum_{k=1}^{m}\frac{\mathbb{E}\left[c_k^2\right]^{\frac{1}{2}}}{m}\frac{\mathbb{E}[L(W_s^m)]^{\frac{1}{2}}}{m^{\frac{1}{2}}}ds + C\theta && \text{By C-S} \\
&\leq C \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \int_S^T \sum_{k=1}^{m}\frac{\mathbb{E}\left[c_k^2\right]^{\frac{1}{2}}}{m}ds + C\theta && \text{by Lemma 4.2} \\
&\leq C\theta && \text{Because } c_k's \text{ moments are}
\end{aligned}
$$

With this, we can deduce:

$$\limsup_{m\to\infty}(1) \leq C\theta,$$

and therefore, by remembering (1)'s definition, we obtain:

$$\lim_{\theta\downarrow 0}\ \limsup_{m\to\infty}\ \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta}\mathbb{E}\left[\left|\frac{1}{m}\sum_{k=1}^{m}\int_S^T \nabla\varphi(W_s^{k,m})^T h^{k,m}(W_s^m)ds\right|\right] = 0. \tag{4.27}$$

Next, we have (2). Let's remember it's definition.

$$(2) = \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta}\mathbb{E}\left[\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_S^T \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds\right|\right].$$

By repeating a previous calculation:

$$\mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right) \leq C\|\Sigma_{k,m}(W_s^{k,m})\|_{Frob}.$$

With this, we get:

$$(2) \leq \tilde{C}\sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta}\mathbb{E}\left[\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_S^T \|\Sigma_{k,m}(W_s^{k,m})\|_{\mathrm{Frob}}ds\right|\right].$$

Now we use Lemma 4.3 and conclude:

$$\begin{aligned}
(2) &\leq \tilde{C} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \mathbb{E}\left[\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_S^T \frac{c_k^2}{m}L(W_s^m)ds\right] \\
&\leq \frac{\tilde{C}\gamma}{2m^{1+\alpha}} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \sum_{k=1}^{m}\int_S^T \frac{1}{m}\mathbb{E}\left[c_k^2 L(W_s^m)\right] ds \\
&\leq \frac{\tilde{C}\gamma}{2m^{\alpha}} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \int_S^T \frac{\mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}}{m^2} ds && \text{by C-S} \\
&\leq \frac{\tilde{C}\gamma}{2m^{\alpha}} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \int_S^T ds && \text{by Lemma 4.4} \\
&\leq \frac{\tilde{C}\gamma}{2m^{\alpha}}\theta && \text{by Lemma 4.4}
\end{aligned}$$

With this last inequality, we conclude:

$$\lim_{\theta\downarrow 0} \limsup_{m\to\infty} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \mathbb{E}\left[\left|\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_S^T \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right) ds\right|\right] = 0.$$

By following what we did in terms (1) and (2), the prove for term (3) is direct. Having all of these inequalities, we can conclude that:

$$\lim_{\theta\downarrow 0} \limsup_{m\to\infty} \sup_{S,T\in\mathcal{T}:S\leq T\leq S+\theta} \mathbb{E}\left[A_t^n\right] = 0. \tag{4.28}$$

Let's continue with the martingale term M_t^m. For (4), we have

$$(4) = \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \mathbb{E}\left[\int_S^T \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds\right].$$

By repeating a previous calculation, we get:

$$\int_S^T \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds \leq C\int_0^N \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} |\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}|\mathbb{1}_{[S,T]}(s)ds,$$

and now, by using Cauchy-Schwarz's inequality, we can deduce:

$$\int_S^T \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds \leq Cp\sum_{k=1}^{m}\int_0^N \|\Sigma_{k,m}(W_s^{k,m})\|\mathbb{1}_{[S,T]}(s)ds.$$

In order to bound the right hand side, we apply Lemma 4.3, which gives us bound over the Frobenius norm of $\Sigma_{k,m}(W_s^{k,m})$:

$$\int_S^T \sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p} \partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds \leq Cp\sum_{k=1}^{m}\int_0^N \frac{c_k^2}{m}L(W_s^m)\mathbb{1}_{[S,T]}(s)ds.$$

Replacing, we obtain:

$$(4) \leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}} \mathbb{E}\left[\left(\sum_{k=1}^{m}\int_0^N \frac{c_k^2}{m} L(W_s^m)\mathbb{1}_{[S,T]}(s)ds\right)^{\frac{1}{2}}\right].$$

Now, we apply Jensen's inequality,and continue bounding the expression by using classical inequalities:

$$\begin{aligned}
(2) &\leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}} \mathbb{E}\left[\sum_{k=1}^{m}\int_0^N \frac{c_k^2}{m} L(W_s^m)\mathbb{1}_{[S,T]}(s)ds\right]^{\frac{1}{2}} \\
&\leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}} \left(\int_0^N \sum_{k=1}^{m} \mathbb{E}\left[\frac{c_k^2}{m} L(W_s^m)\right]\mathbb{1}_{[S,T]}(s)ds\right)^{\frac{1}{2}} \qquad \text{by Fubini} \\
&\leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}} \left(\int_0^N \mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}\mathbb{1}_{[S,T]}(s)ds\right)^{\frac{1}{2}} \qquad \text{by C-S.}
\end{aligned}$$

Now, by reordering the terms inside the integral:

$$\begin{aligned}
(2) &\leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}} \left(\int_0^N \frac{\mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}}{m}\mathbb{1}_{[S,T]}(s)ds\right)^{\frac{1}{2}} \\
&\leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}} \left(\int_0^N \mathbb{1}_{[S,T]}(s)ds\right)^{\frac{1}{2}} \qquad \text{by Lemma 4.4} \\
&\leq C\sqrt{p} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \frac{\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}} \sqrt{T-S} \\
&\leq \frac{C\sqrt{p}\sqrt{\theta}\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}}
\end{aligned}$$

At last, following the same steps we applied for (1), we obtain:

$$\lim_{\theta\downarrow 0} \limsup_{m\to\infty} \sup_{S,T\in\mathcal{T}_N:S\leq\mathcal{T}\leq S+\theta} \mathbb{E}\left[\left|\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{i=1}^{m}\int_S^T \varphi'(W_s^i)\sum_{j=1}^{m}\Sigma_{i,j}^{\frac{1}{2}}(W_s^m)dB_t^j\right|\right] = 0 \tag{4.29}$$

Having this, the prove of the fact that the desired expectation of (5) is bounded is straightforward. Hence, we conclude the proof of Lemma 4.11. □

Having the two previous Lemmas, we can finally prove the lemma at the beginning of this section:

Lemma 4.12 *Let $\sigma : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a bounded, Lipschitz, with bounded-in-norm hessian activation function for a one hidden layer neural network $f^m(c,w)$, whose parameters are initialized such that $W^{k,m} \sim \mathcal{N}(0,1)$ i.i.d and c_k are initialized i.i.d with it's first four moments bounded. Let $(\mu_t^m)_t$ be the empirical measure of the process the process $(c_k, W_t^m)_t$ when*

trained in continuous time by the SDE:

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}.$$

. Then, given $\varphi \in \mathcal{C}_0(\mathbb{R}\times\mathbb{R}^p)$, the process $\langle\varphi, \mu_t^m\rangle$ is tight.

Proof. Lemmas 4.10 and 4.11 combined correspond to the second condition of Aldous-Rebolledo Criterion, which is enough to conclude that for $\varphi \in \mathcal{C}_0(\mathbb{R})$ the process $\langle\varphi, \mu_t^m\rangle$ is tight. □

4.3.2. Second Part of the Proof

In order to finish proving tightness, we must prove the following Lemma.

Lemma 4.13 *Let $\sigma : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a bounded, Lipschitz, with bounded-in-norm hessian activation function for a one hidden layer neural network $f^m(c, w)$, whose parameters are initialized such that $W^{k,m}$ are i.i.d and have their first four moments finite and c_k are initialized i.i.d with it's first four moments bounded. Let $(\mu_t^m)_t$ be the empirical measure of the process the process $(c_k, W_t^m)_t$ when trained in continuous time by the SDE:*

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}.$$

Then, for every $\varepsilon > 0$, there exists a compact set K_ε such that

$$\forall m, t \in [0,T], \sup_{m\in\mathbb{N}} \sup_{t\in[0,T]} \mathbb{P}(\mu_t^m \notin K_\varepsilon) \leq \varepsilon.$$

.

Proof.

Let $\varphi(x) = 1 + \|x\|^2$ for $x \in \mathbb{R}\times\mathbb{R}^p$. Note that $\varphi(x) \to \infty$ when $\|x\| \to \infty$. Also, both $\nabla\varphi(x)$ and $H_X\varphi(x)$ are continuous and $H_X\varphi(x)$ is bounded in norm.

By the dynamics of the empirical measure (4.23) we have:

$$\begin{aligned}\langle\varphi, \mu_t^m - \mu_0^m\rangle = &\frac{1}{m}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \nabla h^{k,m}(W_s^m)ds + \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^m \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k \\ &+ \frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^m \int_0^t \operatorname{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds. \\ &+ \frac{1}{m}\sum_{k=1}^m \int_0^t \sqrt{2\tau}\nabla\varphi(c_k, W_s^{k,m})^T d\tilde{B}_s^{k,m} + \frac{1}{m}\sum_{k=1}^m \int_0^t 2\tau \operatorname{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds.\end{aligned}$$

By taking expectation we get:

$$\begin{aligned}\mathbb{E}\left[\langle\varphi,\mu_t^m\rangle\right] = &\underbrace{\mathbb{E}\left[\langle\varphi,\mu_0^m\rangle\right]}_{(1)} + \underbrace{\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(W_s^{k,m})^T h^{k,m}(W_s^m)ds\right]}_{(2)}\\ &+\underbrace{\mathbb{E}\left[\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^{m}\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right]}_{(3)}\\ &+\underbrace{\mathbb{E}\left[\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds\right]}_{(4)}\\ &+\underbrace{\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^t \sqrt{2\tau}\nabla\varphi(W_s^{k,m})^T d\tilde{B}_s^{k,m}\right]}_{(5)} + \underbrace{\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^t 2\tau\,\mathrm{Tr}\left(H_w\varphi(W_s^{k,m})\right)ds\right]}_{(6)}.\end{aligned}$$

We'd like to prove that the modules of (1), (2), (3), (4), (5) and (6) are finite. Let's begin with term (1), which by definition corresponds to

$$|(1)| = \mathbb{E}\left[\langle\varphi,\mu_0^m\rangle\right] = \langle\varphi,\mu_0^m\rangle = \frac{1}{m}\sum_{i=1}^{m}(1+c_k^2) + \frac{1}{m}\sum_{i=1}^{m}(1+\|W_0^{k,m}\|^2).$$

Since both terms are convergent in m by the Law of Large Numbers, they are finite. Having that $|(1)|$ is finite, we can now analyze term (2). For this term, we recall that $\nabla\varphi(c,w) = 2(c,w)$. Then:

$$\begin{aligned}|(2)| &= \left|\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k,W_s^{k,m})^T h^{k,m}(W_s^m)ds\right]\right|\\ &\leq \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^t \|\nabla\varphi(W_s^{k,m})\|\|h^{k,m}(W_s^m)\|ds\right]\\ &\leq C\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\int_0^t |c_k|\|h^{k,m}(W_s^m)\|ds + \int_0^t \|W_s^{k,m}\|\|h^{k,m}(W_s^m)\|ds\right]\\ &\leq C\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[|c_k|\|h^{k,m}(W_s^m)\|\right]ds + C\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[\|W_s^{k,m}\|\|h^{k,m}(W_s^m)\|\right]ds\\ &\overset{C-S}{\leq} C\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[c_k^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|h^{k,m}(W_s^m)\|^2\right]^{\frac{1}{2}}ds + C\frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|h^{k,m}(W_s^m)\|^2\right]^{\frac{1}{2}}ds.\end{aligned} \tag{4.30}$$

By Cauchy Schwartz and by Lemma 4.6:

$$\|h^{k,m}(W_s^m)\| \leq \lambda\|W_s^{k,m}\| + \frac{|c_k|}{m^{\frac{1}{2}}}L(W_s^m)^{\frac{1}{2}}.$$

Then:

$$\begin{aligned}
\mathbb{E}[\|h^{k,m}(W_s^m)\|^2]^{\frac{1}{2}} &\leq 2\mathbb{E}[\lambda\|W_s^{k,m}\|^2]^{\frac{1}{2}} + \frac{2}{m^{\frac{1}{2}}}\mathbb{E}[c_k^2 L(W_s^m)]^{\frac{1}{2}} \\
&\leq 2\left(\left(p + \frac{C\gamma s}{m^\alpha}\right)^{\frac{1}{2}} + Cs\right) + \frac{2}{m^{\frac{1}{2}}}\mathbb{E}[c_k^2 L(W_s^m)]^{\frac{1}{2}} && \text{By Lemma 4.6} \\
&\overset{C-S}{\leq} 2\left(\left(p + \frac{C\gamma s}{m^\alpha}\right)^{\frac{1}{2}} + Cs\right) + \frac{2}{m^{\frac{1}{2}}}\mathbb{E}[c_k^4]^{\frac{1}{4}}\mathbb{E}[L(W_s^m)^2]^{\frac{1}{4}} \\
&\leq 2\left(\left(p + \frac{C\gamma s}{m^\alpha}\right)^{\frac{1}{2}} + Cs\right) + \frac{2}{m^{\frac{1}{2}}}\mathbb{E}[L(W_s^m)^2]^{\frac{1}{4}} && \text{Because } \mathbb{E}[c_k^4] \leq C.
\end{aligned} \tag{4.31}$$

Recall that by Lemma 4.2:

$$\frac{L(W_s)}{m} \leq C\left(\frac{1}{m} + a_m\right).$$

Then, by using that $(a+b)^2 \leq 2a^2 + 2b^2$:

$$\frac{L(W_s)^2}{m^2} \leq 2C\left(\frac{1}{m^2} + a_m^2\right),$$

and since c_k's are centered and i.i.d:

$$\mathbb{E}\left[\frac{L(W_s)^2}{m^2}\right] \leq C.$$

Replacing in (4.31):

$$\mathbb{E}[\|h^{k,m}(W_s^m)\|^2]^{\frac{1}{2}} \leq 2\left(\left(p + \frac{C\gamma s}{m^\alpha}\right)^{\frac{1}{2}} + Cs\right) + C \leq C + Cs^{\frac{1}{2}} + Cs, \tag{4.32}$$

From 4.30, we can bound the second moment of c_k, and then replace 4.32:

$$\begin{aligned}
|(2)| &\leq C\frac{1}{m}\sum_{k=1}^m \int_0^t \mathbb{E}\left[c_k^2\right]^{\frac{1}{2}} \mathbb{E}\left[\|h^{k,m}(W_s^m)\|^2\right]^{\frac{1}{2}} ds + C\frac{1}{m}\sum_{k=1}^m \int_0^t \mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}} \mathbb{E}\left[\|h^{k,m}(W_s^m)\|^2\right]^{\frac{1}{2}} ds \\
&\leq \frac{C}{m}\sum_{k=1}^m \int_0^t \mathbb{E}\left[\|h^{k,m}(W_s^m)\|^2\right]^{\frac{1}{2}} ds + C\frac{1}{m}\sum_{k=1}^m \int_0^t \mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}} \mathbb{E}\left[\|h^{k,m}(W_s^m)\|^2\right]^{\frac{1}{2}} ds \\
&\leq \frac{C}{m}\sum_{k=1}^m \int_0^t \left(C + Cs^{\frac{1}{4}} + Cs^{\frac{1}{2}}\right) ds + \frac{C}{m}\sum_{k=1}^m \int_0^t \mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}} \left(C + Cs^{\frac{1}{2}} + Cs\right) ds \\
&\leq \left(C + Ct^{\frac{3}{2}} + Ct^2\right) + \frac{C}{m}\sum_{k=1}^m \int_0^t \mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}} \left(C + Cs^{\frac{1}{2}} + Cs\right) ds.
\end{aligned}$$

On the other hand, by Lemma 4.6, we know that:

$$\mathbb{E}\left[\|W_s^{k,m}\|^2\right]^{\frac{1}{2}} \leq \left(\left(p + \frac{C\gamma s}{m^\alpha}\right)^{\frac{1}{2}} + Cs\right),$$

and replacing:

$$\begin{aligned}
|(2)| &\leq \left(C + Ct^{\frac{3}{2}} + Ct^2\right) + \frac{C}{m}\sum_{k=1}^{m}\int_0^t \left(\left(p + \frac{C\gamma s}{m^\alpha}\right)^{\frac{1}{2}} + Cs\right)\left(C + Cs^{\frac{1}{2}} + Cs\right) ds \\
&\leq \left(C + Ct^{\frac{3}{2}} + Ct^2\right) + \frac{C}{m}\sum_{k=1}^{m}\left(C + Ct^{\frac{1}{2}} + Ct\right)^2 t \\
&\leq \left(C + Ct^{\frac{3}{2}} + Ct^2\right) + \left(C + Ct^{\frac{1}{2}} + Ct\right)^2 t \\
&\leq \left(C + CT^{\frac{3}{2}} + CT^2\right) + \left(C + CT^{\frac{1}{2}} + CT\right)^2 t.
\end{aligned}$$

With this, we conclude that $|(2)|$ is bounded. Let's continue with (3). We use Burkholder-Davis-Gundy's inequality to bound $|(3)|$:

$$\begin{aligned}
|(3)| &= \mathbb{E}\left[\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^{m}\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right] \\
&= \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^{m}\mathbb{E}\left[\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right]. \qquad (4.33)
\end{aligned}$$

By applying Burkholder-David-Gundy, we have:

$$\begin{aligned}
\mathbb{E}\left[\left|\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right|\right] &\leq \mathbb{E}\left[\sup_{u\leq t}\left|\int_0^u \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right|\right] \\
&\overset{C-S}{\leq} C\mathbb{E}\left[\left\langle\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle^{\frac{1}{2}}\right]. \qquad (4.34)
\end{aligned}$$

As we already calculated in Lemma 4.10:

$$\left\langle\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle = \int_0^t\sum_{k=1}^{m}\sum_{i_1,i_2=1}^{p}\partial_{i_1}\varphi(W_s^{k,m})\partial_{i_2}\varphi(W_s^{k,m})\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}ds,$$

which in this context corresponds to:

$$\left\langle\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle = \int_0^t\sum_{k=1}^{m}\nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}(W_s^{k,m})\nabla\varphi(W_s^{k,m})ds.$$

By bounding:

$$\begin{aligned}
\left\langle\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle &\leq \int_0^t |\nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}(W_s^{k,m})\nabla\varphi(W_s^{k,m})|ds \\
&\leq \int_0^t \|\nabla\varphi(W_s^{k,m})\|^2\|\Sigma_{k,m}(W_s^{k,m})\|ds \\
&\leq \int_0^t \|W_s^{k,m}\|^2\|\Sigma_{k,m}(W_s^{k,m})\|ds, \qquad (4.35)
\end{aligned}$$

where we used that $\nabla\varphi(W_s^{k,m}) = 2W_s^{k,m}$. By taking expectation in (4.35) and replacing at

(4.34):

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right|\right] &\leq C\mathbb{E}\left[\left\langle\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right\rangle^{\frac{1}{2}}\right]\\ &\leq C\mathbb{E}\left[\left(2\int_0^t \|W_s^{k,m}\|^2\|\Sigma_{k,m}(W_s^{k,m})\|ds\right)^{\frac{1}{2}}\right]\\ &\leq C\mathbb{E}\left[\int_0^t \|W_s^{k,m}\|^2\|\Sigma_{k,m}(W_s^{k,m})\|ds\right]^{\frac{1}{2}}\\ &\leq C\left(\int_0^t \mathbb{E}\left[\|W_s^{k,m}\|^2\|\Sigma_{k,m}(W_s^{k,m})\|\right]ds\right)^{\frac{1}{2}}. \qquad (4.36)\end{aligned}$$

We used Jensen's inequality and the linearity of the expectation in the last lines. By Lemma 4.3:

$$\|\Sigma_{k,m}(W_s^{k,m})\| \leq C\frac{c_k^2}{m}L(W_s^m).$$

By replacing this in (4.36):

$$\mathbb{E}\left[\left|\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right|\right] \leq C\left(\int_0^t \mathbb{E}\left[\|W_s^{k,m})\|^2\frac{c_k^2}{m}L(W_s^m)\right]ds\right)^{\frac{1}{2}}.$$

By Cauchy-Schwarz, we obtain:

$$\mathbb{E}\left[\left|\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right|\right] \leq C\left(\int_0^t \mathbb{E}[\|W_s^{k,m})\|^4]^{\frac{1}{4}}\mathbb{E}[c_k^4]^{\frac{1}{4}}\mathbb{E}[\frac{L(W_s^m)^2}{m^2}]ds\right)^{\frac{1}{2}}.$$

Now, by the Lemma in the appendix, the fact that c_k's 4th moment is bounded, and the fact that, as we previously calculated, $\mathbb{E}[\frac{L(W_s^m)^2}{m^2}] \leq C$, we get:

$$\mathbb{E}\left[\left|\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right|\right] \leq C.$$

Replacing in (4.33):

$$\begin{aligned}|(3)| &= \mathbb{E}\left[\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^m\int_0^t \nabla\varphi(W_s^{k,m})^T\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k\right]\\ &\leq \frac{C\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\\ &\leq C.\end{aligned}$$

With this, the only thing left is that $|(4)|$ is finite. As a matter of fact, we have

$$\begin{aligned}|(4)| &= \left|\mathbb{E}\left[\frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds\right]\right| \\ &\leq \frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\mathbb{E}\left[\int_0^t |\,\mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)|ds\right].\end{aligned}$$

Now, by repeating previous calculations, we have that:

$$\left|\mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)\right| \leq C\|\Sigma_{k,m}(W_s^{k,m})\|$$

Then, by replacing this in our last inequality:

$$|(4)| \leq \frac{C\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\mathbb{E}\left[\int_0^t \|\Sigma_{k,m}(W_s^{k,m})\|_{Frob}ds\right]$$

Now, just as before, we bound using classical inequalities:

$$\begin{aligned}|(4)| &\leq \frac{C\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[\|\Sigma_{k,m}(W_s^{k,m})\|_{Frob}\right]ds && \text{by Fubini} \\ &= \frac{C\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[\frac{c_k^2}{m}L(W_s^m)\right]ds && \text{by Lemma 4.3} \\ &\leq \frac{C\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathbb{E}\left[c_k^4\right]^{\frac{1}{2}}\frac{\mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}}{m^{\frac{1}{2}}}ds && \text{By C-S.}\end{aligned}$$

By using our assumption that c has a bounded fourth moment, and Lemma 4.4's uniform bound, we have:

$$\begin{aligned}|(4)| &\leq \frac{C\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \frac{\mathbb{E}\left[L(W_s^m)^2\right]^{\frac{1}{2}}}{m^{\frac{1}{2}}}ds \\ &\leq \frac{C\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t ds \\ &\leq \frac{C\gamma t}{2m^{\alpha}}.\end{aligned}$$

By noting again that the right hand side has a limit when m goes to infinity for all $\alpha \geq 0$, we conclude that (4) is bounded and therefore finite.
For $|(5)|$, the proof is analogous to (3), and at last, for (6) we do the same as in the last section, i.e we treat them as an analogue (4).

Then (1),(2),(3), (4), (5) and (6) are finite and hence, there exists a constant $C^\star$ such that

$$\mathbb{E}[\langle\varphi, \mu_t^m\rangle] \leq C^\star,$$

with $\varphi(x) = (1 + \|x\|^2)$. Now, let K_R be the set defined by:

$$K_R := \{\mu \in \mathcal{P}(\mathbb{R}) \mid \langle \varphi, \mu \rangle \leq R\}.$$

It's possible to prove that K_R is a compact for all $R > 0$. Given $\varepsilon > 0$, we consider $R = \dfrac{C^\star}{\varepsilon}$, and obtain that for all $t \in [0, T], m \in \mathbb{N}$:

$$\mathbb{P}(\mu_t^m \notin K_R) \overset{Markov}{\leq} \frac{1}{R}\mathbb{E}[\langle \varphi, \mu_t^m \rangle] \leq \frac{C^\star}{R} \leq \varepsilon.$$

In particular, by taking the supremum over $m \in \mathbb{N}$:

$$\sup_{m \in \mathbb{N}} \sup_{t \in [0,T]} \mathbb{P}(\mu_t^m \notin K_R) \overset{Markov}{\leq} \frac{1}{R}\mathbb{E}[\langle \varphi, \mu_t^m \rangle] \leq \frac{C^\star}{R} \leq \varepsilon,$$

with which we can finish our proof. □

In the following, we state the Theorem that will allow us to finish this section. For a proof, we refer the reader to [31].

Theorem 4.2 *Given a collection of random measures$(\mu_t^m)_t$, with $m \in \mathbb{N}$, the laws of this process are thigh if they satisfy the following conditions:*

1. *Given $\varphi \in \mathcal{C}_0(\mathbb{R})$, the process$\langle \varphi, \mu_t^m \rangle$ is tense.*
2. *For all $\varepsilon > 0$, there exists a compact set K_ε such that*

$$\forall m, t \in [0, T], \sup_{m \in \mathbb{N}} \sup_{t \in [0,T]} \mathbb{P}(\mu_t^m \notin K_\varepsilon) \leq \varepsilon.$$

A straightforward application of this Theorem is the following Proposition, which will be our starting point to prove the convergence of the process of empirical measures on a suitable space.

Proposition 4.1 *Let $\sigma : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a bounded, Lipschitz, with bounded-in-norm hessian activation function for a one hidden layer neural network $f^m(c, w)$, whose parameters are initialized such that $W^{k,m}$ are i.i.d and have their first four moments finite, and c_k are initialized i.i.d with it's first four moments bounded. Let $(\mu_t^m)_t$ be the empirical measure of the process the process $(c_k, W_t^m)_t$ when trained in continuous time by the SDE:*

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}.$$

Then the laws of the process given by the empirical measures of the process $(W_t^m)_t$, with $m \in \mathbb{N}$, are tight.

Proof. It's a consequence of Theorem 4.2, and of Lemmas 4.10 and 4.11. □

4.4. The PDE limit

Let $\sigma : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a bounded, Lipschitz, with bounded-in-norm hessian activation function for a one hidden layer neural network $f^m(c, w)$, whose parameters are initialized such that $W^{k,m} \sim \mathcal{N}(0, 1)$ i.i.d and c_k are initialized i.i.d with it's first four moments bounded. Let $(\mu_t^m)_t$ be the empirical measure of the process the process $(c_k, W_t^m)_t$ when trained in continuous time by the SDE:

$$dW_t^{k,m} = h^{k,m}(w)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}.$$

In the last section we proved that under this hypothesis the laws of the process of empirical measures are tight. This is only the first part of our study, since it gives us a hint on how to prove convergence (if the convergence exists) to some kind of limit. We devote this section to the study of this convergence, which will be represented as a Partial Differential Equation (PDE) in the distributional sense. We begin by identifying the limiting PDE, and next, we prove the convergence to this equation.

4.4.1. Identification of the Limit

Let's remember the dynamics of the empirical measure (4.23):

$$\begin{aligned}\langle \varphi, \mu_t^m - \mu_0^m \rangle = &\int_0^t \left\langle \nabla\varphi(w)^T h^{k,m}(w)ds, \mu_s^m \right\rangle ds - \lambda \int_0^t \left\langle \nabla\varphi(c,w)^T w, \mu_s^m \right\rangle ds \\ &\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}} \int_0^t \left\langle \nabla\varphi(w)^T \Sigma_{k,m}^{\frac{1}{2}}(w), \mu_s^m \right\rangle dB_s^k \\ &+ \frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(w)^T H_w\varphi(w)\Sigma_{k,m}^{\frac{1}{2}}(w)\right), \mu_s^m \right\rangle ds \\ &+ \int_0^t \left\langle \sqrt{2\tau}\nabla\varphi(w)^T, \mu_s^m \right\rangle d\tilde{B}_s^{k,m} + \int_0^t \left\langle 2\tau\,\mathrm{Tr}\left(H_w\varphi(W_s^{k,m})\right), \mu_s^m \right\rangle ds.\end{aligned}$$

We already know about the tightness of the laws of the empirical measure process, but we'd also like to prove convergence to a given equation. The central question in this part we'll be: Who's that limit equation? In the first place, it'll be help full to count with the following Lemma.

Lemma 4.14 *The limit in law of the the empirical measures μ_0^m when $m \to \infty$ is μ_0, the initialization distribution.*

Proof. Since the parameters are initialized independently, it's a straightforward consequence of the Law of Large numbers. □

Recall that μ_0 is the law of the neural network's parameters at initialization. On the first place, it's natural to expect that:

$$-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(c,\tilde{W})^T\tilde{W}\mu_s^m(d\tilde{c}, d\tilde{W})ds \underset{m\to\infty}{\longrightarrow} -\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(c,\tilde{W})^T\tilde{W}\mu_s^m(d\tilde{c}, d\tilde{W})ds.$$

Now, let's study the other terms. For simplicity in our notation, we define the terms

$$A_m(t) = \frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \nabla h^{k,m}(W_s^m)ds; B_m(t) = \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})dB_s^k;$$

$$C_m(t) = \frac{\gamma}{2m^{1+\alpha}}\sum_{k=1}^{m}\int_0^t \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m})\right)ds;$$

$$D_m(t) = \frac{1}{m}\sum_{k=1}^{m}\int_0^t \sqrt{2\tau}\nabla\varphi(c_k, W_s^{k,m})^T d\tilde{B}_s^{k,m} \text{ and } E_m(t) = \frac{1}{m}\sum_{k=1}^{m}\int_0^t 2\tau\,\mathrm{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right)ds.$$

In order to identify the limit in an easier way, we'll re-write the terms. For $A_m(t)$ we have:

$$\begin{aligned}
A_m(t) &= \frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \nabla h^{k,m}(W_s^m)ds \\
&= \frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \mathbb{E}_{X,Y}\left[(Y - f^m(W^m, X))\frac{c_k}{\sqrt{m}}\nabla\sigma(W_s^{k,m}, X)\right]ds \\
&+ \frac{\lambda}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T W_s^{k,m}ds \\
&= \frac{1}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \mathbb{E}_{X,Y}\left[(\frac{Y}{\sqrt{m}} - \langle c\sigma, \mu_s^m\rangle)c_k\nabla\sigma(W_s^{k,m}, X)\right]ds \\
&+ \frac{\lambda}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T W_s^{k,m}ds \\
&= \frac{1}{m}\sum_{k=1}^{m}\int_0^t \mathbb{E}_{X,Y}\left[(\frac{Y}{\sqrt{m}} - \langle c\sigma, \mu_s^m\rangle)c_k\nabla\varphi(c_k, W_s^{k,m})^T\nabla\sigma(W_s^{k,m}, X)\right]ds \\
&+ \frac{\lambda}{m}\sum_{k=1}^{m}\int_0^t \nabla\varphi(c_k, W_s^{k,m})^T W_s^{k,m}ds.
\end{aligned}$$

We can re-write this last expression so its dependence on the empirical measure is clearer. We get:

$$\begin{aligned}
A_m(t) = \int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[(\frac{Y}{\sqrt{m}} - \langle c\sigma(w,x), \mu_s^m\rangle)\tilde{c}\nabla\varphi(\tilde{W})^T\nabla\sigma(\tilde{W}, X)\right]\mu_s^m(d\tilde{c}, d\tilde{w})ds \\
+\lambda\int_0^t \langle\nabla\varphi(c, \tilde{W})^T\tilde{W}, \mu_s^m\rangle ds
\end{aligned} \tag{4.37}$$

Hence, we expect that if a limiting measure process $\mu \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}\times\mathbb{R}^p))$ exists (in law), then:

$$A_m(t) \xrightarrow[n\to\infty]{\mathcal{L}} -\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[\langle c\sigma, \mu_s\rangle)\tilde{c}\nabla\varphi(\tilde{W})^T\nabla\sigma(\tilde{W}, X)\right]\mu_s(d\tilde{c}, d\tilde{W})ds + \lambda\int_0^t\langle\nabla\varphi(\tilde{W})^T\tilde{W}, \mu_s\rangle ds$$

Before we continue with $B_m(t)$ and $C_m(t)$, we'll define the following function, which will act as a limit to $\Sigma_{k,m}(W_s^{k,m})$. Let $S(\mu, w, c)$ be the function

$$S(\mu, w, c) : \mathcal{P}(\mathbb{R}^p)\times\mathbb{R}^p\times\mathbb{R} \to \mathcal{M}_{p,p}(\mathbb{R}),$$

given by

$$S(\mu, w, c) = c^2 \left(\mathbb{E}_X \left[\langle c\sigma, \mu \rangle^2 \partial_i \sigma(w, X) \partial_j \sigma(w, X) \right] - \mathbb{E}_X \left[(\langle c\sigma, \mu \rangle \partial_i \sigma(w, X)) \right] \mathbb{E}_X \left[(\langle c\sigma, \mu \rangle \partial_j \sigma(w, X)) \right] \right)_{i,j \in [p]} \tag{4.38}$$

By doing something similar to what we did with $A_m(t)$, we expect that for the term

$$C_m(t) = \frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr} \left(\Sigma_{k,m}^{\frac{1}{2}}(w)^T H_w \varphi(c, w) \Sigma_{k,m}^{\frac{1}{2}}(w) \right), \mu_s^m \right\rangle ds,$$

we'll obtain

$$C_m(t) \xrightarrow[n \to \infty]{\mathcal{L}} 0,$$

in the case where $\alpha > 0$, and

$$C_m(t) \xrightarrow[n \to \infty]{\mathcal{L}} \gamma \int_0^t \langle \mathrm{Tr} \left(S(\mu, \tilde{W}, \tilde{c})^T H_w \varphi(\tilde{c}, \tilde{W}) \right), \mu_s \rangle ds$$

in the case where $\alpha = 0$. For $E_m(t)$, we expect:

$$E_m(t) \xrightarrow[n \to \infty]{\mathcal{L}} \sqrt{2\tau} \int_0^t \langle \mathrm{Tr} \left(H_w \varphi(\tilde{c}, \tilde{W}) \right), \mu_s \rangle ds.$$

Next, Lemma 4.15 tells us that the martingale terms $B_m(t)$ and $D_m(t)$ will go to 0 in L^1.

Lemma 4.15 *Given $t \geq 0$, we have:*

$$\lim_{m \to \infty} \mathbb{E} \left[|B_m(t)| \right] = 0,$$

and

$$\lim_{m \to \infty} \mathbb{E} \left[|D_m(t)| \right] = 0.$$

Proof. Let $t \geq 0$. By definition, we know that

$$B_m(t) = -\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}} \sum_{k=1}^m \int_0^t \nabla \varphi(W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m}) dB_s^k.$$

By repeating a previous calculation

$$\langle B_m(t) \rangle = \left\langle \frac{\gamma}{m^{2+\alpha}} \sum_{k=1}^m \int_0^N \nabla \varphi(W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m}) dB_s^k \right\rangle \leq C \frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}} \int_0^N \sum_{k=1}^m \sum_{i_1, i_2 = 1}^p |\Sigma_{k,m}(W_s^{k,m})_{i_1, i_2}| ds. \tag{4.39}$$

Let $B_m^*(t) := \sup_{s \leq t} |B_m(t)|$. By Burkholder-Davis-Gundy (BDG) inequality:

$$\mathbb{E}[B_m^*(t)] \leq C \mathbb{E} \left[\langle B_m(t) \rangle^{\frac{1}{2}} \right]. \tag{4.40}$$

Replacing equation (4.39) in (4.40), and then bounding φ's gradient norm:

$$\begin{aligned}\mathbb{E}[B_m^*(t)] &\leq C\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\mathbb{E}\left[\left(\int_0^t \sum_{k=1}^m \sum_{i_1,i_2=1}^p |\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}|ds\right)^{\frac{1}{2}}\right] \\ &\leq C\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\left(\mathbb{E}\left[\int_0^t \sum_{k=1}^m \sum_{i_1,i_2=1}^p |\Sigma_{k,m}(W_s^{k,m})_{i_1,i_2}|ds\right]\right)^{\frac{1}{2}} \qquad \text{By Jensen's inequality.}\end{aligned}$$

Now, by Cauchy-Schwarz:

$$\mathbb{E}[B_m^*(t)] \leq C\frac{\gamma^{\frac{1}{2}}}{m^{1+\frac{\alpha}{2}}}\left(\mathbb{E}\left[\int_0^t \sum_{k=1}^m p\|\Sigma_{k,m}(W_s^{k,m})\|ds\right]\right)^{\frac{1}{2}}$$

By rewriting this last expression, we get:

$$\begin{aligned}\mathbb{E}[B_m^*(t)] &\leq C\frac{\gamma^{\frac{1}{2}}\sqrt{p}}{m^{1+\frac{\alpha}{2}}}\left(\mathbb{E}\left[\int_0^N \sum_{k=1}^m \|\Sigma_{k,m}(W_s^{k,m})\|ds\right]\right)^{\frac{1}{2}} \\ &\leq C\frac{\gamma^{\frac{1}{2}}\sqrt{p}}{m^{1+\frac{\alpha}{2}}}\left(\mathbb{E}\left[\int_0^N \sum_{k=1}^m \frac{c_k^2}{m}L(W_s^m)ds\right]\right)^{\frac{1}{2}} \qquad \text{By Lemma 4.3} \\ &\leq C\frac{\gamma^{\frac{1}{2}}\sqrt{p}}{m^{1+\frac{\alpha}{2}}}\left(\int_0^N \sum_{k=1}^m \frac{\mathbb{E}\left[c_k^2 L(W_s^m)\right]}{m}ds\right)^{\frac{1}{2}} \qquad \text{By Fubini}\end{aligned}$$

By applying Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned}\frac{\mathbb{E}\left[c_k^2 L(W_s^m)\right]}{m} &\leq C\mathbb{E}[c_k^4]^{\frac{1}{4}}\frac{\mathbb{E}[L(W_s^m)^2]^{\frac{1}{2}}}{m} \\ &\leq C\left(\mathbb{E}\left[\frac{L(W_s^m)^2}{m^2}\right]\right)^{\frac{1}{2}} \qquad \text{Because } c_k\text{'s 4th moment are bounded.} \\ &\leq C \qquad \text{By Lemma 4.2.}\end{aligned}$$

Replacing this in our last inequality for $\mathbb{E}[B_m^*(t)]$:

$$\begin{aligned}\mathbb{E}[B_m^*(t)] &\leq C\frac{\gamma^{\frac{1}{2}}\sqrt{p}}{m^{1+\frac{\alpha}{2}}}\left(\int_0^N \sum_{k=1}^m ds\right)^{\frac{1}{2}} \\ &\leq C\frac{\gamma^{\frac{1}{2}\sqrt{N}}\sqrt{p}}{m^{\frac{1+\alpha}{2}}}.\end{aligned}$$

Finally, we take the limit as $m \to \infty$, we get:

$$\lim_{m\to\infty}\mathbb{E}[B_m^*(t)] = 0.$$

At last, since $|B_m(t)| \leq B_m^*(t)$, we conclude:

$$\lim_{m\to\infty}\mathbb{E}[|B_m(t)|] = 0,$$

which is exactly what we wanted to prove. Note that the proof for $D_m(t)$ is analogous, since instead of Σ we have the identity. Having this, we conclude the demonstration. □

Before concluding this section, we'll introduce the notation:

$$\langle c\sigma(w,X), \mu_s^m \rangle := \int_{\mathbb{R}\times\mathbb{R}^p} c\sigma(w,X)\mu_s^m(dc,dw).$$

By taking into account Lemma 4.15, it's natural to expect that, if the limit exists in some sense, then it will be given by:

- If $\alpha = 0$:

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(\tilde{c},\tilde{W})^T \tilde{W} \mu_s(d\tilde{c},d\tilde{W})ds \\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[\langle c\sigma(w,X),\mu_s\rangle)\tilde{c}\nabla\varphi(\tilde{W})^T\nabla\sigma(\tilde{W},X)\right]\mu_s(d\tilde{c},d\tilde{W})ds \\ &+ \gamma\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(\mu,\tilde{W},\tilde{c})^T H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c},d\tilde{W})ds \\ &+ \sqrt{2\tau}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c},d\tilde{W})ds.\end{aligned}$$

- If $\alpha > 0$:

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(\tilde{c},\tilde{W})^T \tilde{W} \mu_s(d\tilde{c},d\tilde{W})ds \\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[\langle c\sigma,\mu_s\rangle)\tilde{c}\nabla\varphi(\tilde{W})^T\nabla\sigma(\tilde{W},X)\right] d\mu_s(d\tilde{c},d\tilde{W})ds \\ &+ \sqrt{2\tau}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c},d\tilde{W})ds.\end{aligned}$$

4.4.2. Convergence to the Limit

Having identified our limit candidate, we'll can now proceed to prove convergence to this candidate in some sense. We'll only study the case when $\alpha = 0$. The case when $\alpha > 0$ will be a straightforward extension of this other case. We define $F : \mathcal{P}(\mathbb{R}^p) \to \mathbb{R}$ such that, for $\mu \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^2))$,

$$\begin{aligned}F(\mu) = \Big|&\langle \varphi, \mu_t - \mu_0 \rangle + \lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(\tilde{W})^T\tilde{W} d\mu_s(d\tilde{c},d\tilde{W})ds && (4.41)\\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[\langle c\sigma,\mu_s\rangle)\tilde{c}\nabla\varphi(\tilde{W})^T\nabla\sigma(\tilde{W},X)\right] d\mu_s(d\tilde{c},d\tilde{W})ds && (4.42)\\ &+ \gamma\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(\mu,\tilde{W},\tilde{c})^T H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c},d\tilde{W})ds && (4.43)\\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c},d\tilde{W})ds\Big|. && (4.44)\end{aligned}$$

The idea will be to use this function F to prove the convergence in the PDEs on the last subsection, when m goes to infinity. We'll prove that $F(\mu^m)$ goes to 0 as m goes to infinity. Since F is positive, this will help us prove that the limit μ satisfies the PDE inside the module of (4.82). Having done this analysis, a direct corollary will be the fact that the

PDE has a solution. Then a natural question will be: Is this solution unique? We'd like the solution to be unique, since in that case we could prove that the convergence is not only for a sub-sequence of the process of empirical measures, but of the whole sequence. The next section will be devoted to this study.

In order to make our proof clear, we'll define the following function, which is given by the difference that arises when we try to approximate the matrices $\Sigma_{k,m}$ by the mean-field operator $S(w,\mu)$, for $m \in \mathbb{N}$. The difference between both operator will be summed up in a reminder, which will be called $R_{k,m}$. More precisely, we define for $k \in \{1 \dots m\}$, the function $R_{k,m} : \mathcal{P}(\mathbb{R}^p) \times \mathbb{R} \to \mathbb{R}^{p\times p}$ by:

$$\begin{aligned}(R_{k,m}(\mu,w))_{i,j} = c_k^2 \mathbb{E}_{X,Y}\left[\left(\frac{Y^2}{m} - \frac{2Y f^{\mu^m}(X)}{m}\right)\partial_i\sigma(W_s^{k,m},X)\partial_j\sigma(W_s^{k,m},X)\right] \\ -c_k^2\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\partial_i\sigma(W_s^{k,m},X)\right]\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\partial_j\sigma(W_s^{k,m},X)\right],\end{aligned} \tag{4.45}$$

where we introduced the notation for the neural network:

$$f^{\mu}(X) := \frac{1}{\sqrt{m}}\sum_{k=1}^{m} c_k\sigma(W^{k,m},X). \tag{4.46}$$

In the following, we proceed to state a Lemma that will allow us to control the norm of $R_{k,m}(\mu,w)$ for $\mu \in \mathcal{P}(\mathbb{R}^p)$, $w \in \mathbb{R}$.

Lemma 4.16 *For every $\varepsilon > 0$, there exists $\tilde{m}$ such that $\forall m \geq \tilde{m}$,*

$$\|R(\mu^m, W_s^{k,m})\|_{Frob} \leq \varepsilon,$$

for each $k \in \{1,\dots,m\}$, almost surely.

Proof. Let $\varepsilon > 0$. Given $m \in \mathbb{N}$, $k \in \{1,\dots,m\}$, and $i,j \in \{1,\dots,p\}$. Recall that we note the neural network by

$$f^{\mu}(X) := \frac{1}{\sqrt{m}}\sum_{k=1}^{m} c_k\sigma(W^{k,m},X). \tag{4.47}$$

With this notation in mind, we have

$$|R_{k,m}(\mu^m, W_s^{k,m})_{i,j}| = \left|c_k^2\mathbb{E}_{X,Y}\left[\left(\frac{Y^2}{m} - \frac{2Y f^{\mu}(X)}{m}\right)\partial_i\sigma(W_s^{k,m},X)\partial_j\sigma(W_s^{k,m},X)\right]\right. \tag{4.48}$$

$$\left. -c_k^2\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\partial_i\sigma(W_s^{k,m},X)\right]\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\partial_j\sigma(W_s^{k,m},X)\right]\right|. \tag{4.49}$$

$$\leq c_k^2\mathbb{E}_{X,Y}\left[\left(\frac{Y^2}{m} + \frac{2Y|f^{\mu}(X)|}{m}\right)|\partial_i\sigma(W_s^{k,m},X)||\partial_j\sigma(W_s^{k,m},X)|\right] \tag{4.50}$$

$$+ c_k^2\mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}|\partial_i\sigma(W_s^{k,m},X)|\right]\mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}|\partial_j\sigma(W_s^{k,m},X)|\right] \tag{4.51}$$

$$\leq Cc_k^2\mathbb{E}_{X,Y}\left[\frac{Y^2}{m} + \frac{2|Y||f^{\mu}(X)|}{m}\right] + Cc_k^2\mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}\right]\mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}\right]. \tag{4.52}$$

Now, since σ is a bounded function by our assumptions, for $x \in \mathcal{X}$ we have that for some $C > 0$

$$\left| \frac{f^{\mu}(x)}{\sqrt{m}} \right| \leq C \frac{1}{m} \sum_{i=1}^{m} |c_i|.$$

Replacing this in equation (4.52):

$$\begin{aligned} |R_{k,m}(\mu^m, W_s^{k,m})_{i,j}| &\leq C c_k^2 \mathbb{E}_{X,Y} \left[\frac{Y^2}{m} + \frac{2|Y| \frac{1}{m} \sum_{i=1}^{m} |c_i|}{\sqrt{m}} \right] + C c_k^2 \mathbb{E}_{X,Y} \left[\frac{|Y|}{\sqrt{m}} \right] \mathbb{E}_{X,Y} \left[\frac{|Y|}{\sqrt{m}} \right] \\ &\leq C c_k^2 \frac{C}{m} + C c_k^2 \frac{C}{m^{\frac{3}{2}}} \sum_{i=1}^{m} |c_i| + \frac{C}{\sqrt{m}} c_k^2 \mathbb{E}_{X,Y} \left[\frac{|Y|}{\sqrt{m}} \right] \end{aligned}$$

By taking expectation and Cauchy Schwarz, since c_k's have finite second moment, we get:

$$\mathbb{E}[|R_{k,m}(\mu^m, W_s^{k,m})_{i,j}|] \leq \frac{C \mathbb{E}[c_k^2]}{m} + \frac{C}{m^{\frac{3}{2}}} \mathbb{E}[c_k^4]^{\frac{1}{2}} \mathbb{E} \left[\left(\sum_{i=1}^{m} |c_i| \right)^2 \right]^{\frac{1}{2}} + \frac{C}{\sqrt{m}} \mathbb{E}[c_k^2] \tag{4.53}$$

$$\leq \frac{C}{m} + \frac{C}{m^{\frac{3}{2}}} \mathbb{E} \left[\left(\sum_{i=1}^{m} |c_i| \right)^2 \right]^{\frac{1}{2}} + \frac{C}{\sqrt{m}} \tag{4.54}$$

By Cauchy Schwarz on the sum, we know that:

$$\left(\sum_{i=1}^{m} |c_i| \right)^2 \leq m \sum_{i=1}^{m} c_i^2.$$

Replacing this in equation (4.54):

$$\mathbb{E}[|R_{k,m}(\mu^m, W_s^{k,m})_{i,j}|] \leq \frac{C}{m} + \frac{C}{m} \left(\sum_{i=1}^{m} \mathbb{E}\left[c_i^2\right] \right)^{\frac{1}{2}} + \frac{C}{\sqrt{m}} \tag{4.55}$$

$$\leq \frac{C}{m} + \frac{C}{m^{\frac{1}{2}}} + \frac{C}{\sqrt{m}}. \tag{4.56}$$

Finally, by taking m as big as necessary, we get:

$$\mathbb{E}[|R_{k,m}(\mu^m, W_s^{k,m})_{i,j}|] \leq \varepsilon \quad \text{a.s,}$$

which is what we wanted to conclude. □

The previous Lemma will be useful to prove that the limit of $\mathbb{E}[F(\mu^m)]$ when m goes to infinity is in fact 0. We state this in the following Lemma.

Lemma 4.17 *For fixed φ and t, and given $\mu^m \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R} \times \mathbb{R}^p)$, we have that:*

$$\lim_{m \to \infty} \mathbb{E}[F(\mu^m)] = 0.$$

Proof. We begin by recalling the dynamics of the empirical measure we wrote in equation

(4.23):

$$\begin{aligned}\langle \varphi, \mu_t^m - \mu_0^m \rangle = &\int_0^t \left\langle \nabla\varphi(c,w)^T h^{k,m}(w) ds, \mu_s^m \right\rangle ds - \lambda \int_0^t \left\langle \nabla\varphi(c,w)^T w, \mu_s^m \right\rangle ds \\ &\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}} \int_0^t \left\langle \nabla\varphi(c,w)^T \Sigma_{k,m}^{\frac{1}{2}}(w), \mu_s^m \right\rangle dB_s^k \\ &+ \frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr}\left(\Sigma_{k,m}(w)^T H_w \varphi(c,w)\right), \mu_s^m \right\rangle ds \\ &+ \int_0^t \left\langle \sqrt{2\tau}\nabla\varphi(c,w)^T, \mu_s^m \right\rangle d\tilde{B}_s^{k,m} + \int_0^t \left\langle 2\tau\, \mathrm{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right), \mu_s^m \right\rangle ds.\end{aligned}$$

By expanding the first term in the right hand side, and summing it on the left side, we obtain:

$$\begin{aligned}\langle \varphi, \mu_t^m - \mu_0^m \rangle + &\int_0^t \left\langle \nabla\varphi(c,w)^T \mathbb{E}_{X,Y}\left[\frac{f^{\mu_s^m}(X)}{\sqrt{m}} \nabla\sigma(w,X)\right] ds, \mu_s^m \right\rangle ds \\ &+ \lambda \int_0^t \left\langle \nabla\varphi(c,w)^T w, \mu_s^m \right\rangle ds = \int_0^t \left\langle \nabla\varphi(c,w)^T \mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}} \nabla\sigma(w,X)\right] ds, \mu_s^m \right\rangle ds \\ &\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}} \int_0^t \left\langle \nabla\varphi(c,w)^T \Sigma_{k,m}^{\frac{1}{2}}(w), \mu_s^m \right\rangle dB_s^k \\ &+ \frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr}\left(\Sigma_{k,m}(w)^T H_w \varphi(c,w)\right), \mu_s^m \right\rangle ds \\ &+ \int_0^t \left\langle \sqrt{2\tau}\nabla\varphi(c,w)^T, \mu_s^m \right\rangle d\tilde{B}_s^{k,m} + \int_0^t \left\langle 2\tau\, \mathrm{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right), \mu_s^m \right\rangle ds.\end{aligned}$$

Next, note that by adding and subtracting the integral of the mean field operator we defined in equation (4.38):

$$\frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr}\left(S(c,w,\mu_s^m)^T H_w \varphi(c,w)\right), \mu_s^m \right\rangle ds,$$

and mixing it with the matrix $R_{k,m}(\mu^m, W_s^{k,m}$ we defined in equation (4.45) to obtain:

$$\begin{aligned}\langle \varphi, \mu_t^m - \mu_0^m \rangle + &\int_0^t \left\langle \nabla\varphi(c,w)^T \mathbb{E}_{X,Y}\left[\frac{f^{\mu_s^m}(X)}{\sqrt{m}} \nabla\sigma(w,X)\right] ds, \mu_s^m \right\rangle ds \\ &+ \lambda \int_0^t \left\langle \nabla\varphi(c,w)^T w, \mu_s^m \right\rangle ds - \frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr}\left(S(c,w,\mu_s^m) H_w\varphi(c,w)\right), \mu_s^m \right\rangle ds \\ &- \int_0^t \left\langle 2\tau\, \mathrm{Tr}\left(H_w\varphi(c_k, W_s^{k,m})\right), \mu_s^m \right\rangle ds = \int_0^t \left\langle \nabla\varphi(c,w)^T \mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}} \nabla\sigma(w,X)\right] ds, \mu_s^m \right\rangle ds \\ &\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}} \int_0^t \left\langle \nabla\varphi(c,w)^T \Sigma_{k,m}^{\frac{1}{2}}(w), \mu_s^m \right\rangle dB_s^k + \frac{\gamma}{2m^\alpha} \int_0^t \left\langle \mathrm{Tr}\left(R(c,w,\mu_s^m)^T H_w \varphi(c,w)\right), \mu_s^m \right\rangle ds \\ &+ \int_0^t \left\langle \sqrt{2\tau}\nabla\varphi(c,w)^T, \mu_s^m \right\rangle d\tilde{B}_s^{k,m}.\end{aligned}$$

Next, by taking module of both sides we get:

$$\begin{aligned}\Big|\langle\varphi,\mu_t^m-\mu_0^m\rangle &+\int_0^t\left\langle\nabla\varphi(c,w)^T\mathbb{E}_{X,Y}\left[\frac{f^{\mu_s^m}(X)}{\sqrt{m}}\nabla\sigma(w,X)\right]ds,\mu_s^m\right\rangle ds\\ &+\lambda\int_0^t\left\langle\nabla\varphi(c,w)^Tw,\mu_s^m\right\rangle ds-\frac{\gamma}{2m^\alpha}\int_0^t\left\langle\mathrm{Tr}\left(S(c,w,\mu_s^m)H_w\varphi(c,w)\right),\mu_s^m\right\rangle ds\\ &-\int_0^t\left\langle 2\tau\,\mathrm{Tr}\left(H_w\varphi(c_k,W_s^{k,m})\right),\mu_s^m\right\rangle ds\Big|=\left|\int_0^t\left\langle\nabla\varphi(c,w)^T\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla\sigma(w,X)\right]ds,\mu_s^m\right\rangle ds\right.\\ &\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\int_0^t\left\langle\nabla\varphi(c,w)^T\Sigma_{k,m}^{\frac{1}{2}}(w),\mu_s^m\right\rangle dB_s^k+\frac{\gamma}{2m^\alpha}\int_0^t\left\langle\mathrm{Tr}\left(R(c,w,\mu_s^m)^TH_w\varphi(c,w)\right),\mu_s^m\right\rangle ds\\ &\left.+\int_0^t\left\langle\sqrt{2\tau}\nabla\varphi(c,w)^T,\mu_s^m\right\rangle d\tilde{B}_s^{k,m}\right|.\end{aligned}$$

Note that the left side of this equation corresponds exactly to the definition of $F(\mu^m)$ in equation (4.82). We'll note

$$B_m(t):=\frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\int_0^t\left\langle\nabla\varphi(c,w)^T\Sigma_{k,m}^{\frac{1}{2}}(w),\mu_s^m\right\rangle dB_s^k\text{ and }E_m(t):=\int_0^t\left\langle\sqrt{2\tau}\nabla\varphi(c,w)^T,\mu_s^m\right\rangle d\tilde{B}_s^{k,m}.$$

Whit this, we can write the following equality:

$$\begin{aligned}F(\mu^m)=\Bigg|&\underbrace{\frac{1}{m}\sum_{k=1}^m\int_0^t\nabla\varphi(c_k,W_s^{k,m})^T\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}c_k\nabla\sigma(W_s^{k,m},X)\right]ds}_{(1)}\\ &+\underbrace{B_m(t)}_{(2)}+\underbrace{\frac{\gamma}{2m}\int_0^t\sum_{k=1}^m\mathrm{Tr}\left(R(\mu_s^m,W_s^{k,m},c_k)^TH_w\varphi(c_k,W_s^{k,m})\right)ds}_{(3)}+\underbrace{E_m(t)}_{(4)}\Bigg|\end{aligned}$$

By applying the expectation w.r.t the Brownian filtration on the left side, we can apply the triangular inequality in the right one and obtain:

$$\mathbb{E}[F(\mu^m)]\leq\mathbb{E}[|(1)|]+\mathbb{E}[|(2)|]+\mathbb{E}[|(3)|].$$

Let's see that the limit of each of these terms when m goes to infinity is 0. Let0s begin with

(1). We have:

$$
\begin{aligned}
|(1)| &= \left| \frac{1}{m} \sum_{k=1}^{m} \int_0^t \nabla \varphi(c_k, W_s^{k,m})^T \mathbb{E}_{X,Y} \left[\frac{Y}{\sqrt{m}} c_k \nabla \sigma(W_s^{k,m}, X) \right] ds \right| && \text{By definition} \\
&\leq \frac{1}{m} \sum_{k=1}^{m} \int_0^t \|\nabla \varphi(c_k, W_s^{k,m})\| \mathbb{E}_{X,Y} \left[\frac{|Y|}{\sqrt{m}} |c_k| \|\nabla \sigma(W_s^{k,m}, X)\| \right] ds && \text{By bounding with the module} \\
&\leq \frac{C}{m} \sum_{k=1}^{m} \int_0^t \frac{|c_k|}{\sqrt{m}} \mathbb{E}_{X,Y} \left[|Y| \right] ds && \text{Bounding } |\sigma| \text{ and } |\varphi| \\
&\leq \frac{C}{m} \sum_{k=1}^{m} \int_0^t \frac{|c_k|}{\sqrt{m}} \mathbb{E}_{X,Y} \left[Y^2 \right]^{\frac{1}{2}} ds && \text{por C-S} \\
&\leq \frac{C}{m^{\frac{1}{2}+1}} \int_0^t \sum_{k=1}^{m} |c_k| ds && \text{because} \mathbb{E}[Y^2] < \infty.
\end{aligned}
$$

Now, we can use Cauchy Schwarz on the sum and use that we are considering c to have a bounded second momentum. This way:

$$\mathbb{E}[|(1)|] \leq \frac{Ct}{\sqrt{m}}. \tag{4.57}$$

Now, by taking $m \to \infty$ we conclude

$$\lim_{m \to \infty} \mathbb{E}[|(1)|] = 0. \tag{4.58}$$

We bound (2) and (4) by using Lemma 4.15, which tells us that the limit when m goes to infinity of this term was equal to 0. At last, we have (3). We have:

$$
\begin{aligned}
|(3)| &= \left| \frac{\gamma}{2m} \int_0^t \sum_{k=1}^{m} \operatorname{Tr} \left(R(\mu_s^m, W_s^{k,m}, c_k)^T H_w \varphi(c_k, W_s^{k,m}) \right) ds \right| && \text{By definition} \\
&\leq \frac{\gamma}{2m} \int_0^t \sum_{k=1}^{m} \left| \operatorname{Tr} \left(R(\mu_s^m, W_s^{k,m}, c_k)^T H_w \varphi(c_k, W_s^{k,m}) \right) \right| ds && \text{bounding by the module} \\
&\leq \frac{\gamma}{2m} \int_0^t \sum_{k=1}^{m} \|R(\mu_s^m, W_s^{k,m}, c_k)\| \|H_w \varphi(c_k, W_s^{k,m})\| ds && \text{because } \operatorname{Tr}(AB) \leq \operatorname{Tr}\left(AA^T\right)^{\frac{1}{2}} \operatorname{Tr}\left(BB^T\right)^{\frac{1}{2}} \\
&\leq \frac{C\gamma}{2m} \int_0^t \sum_{k=1}^{m} \|R(\mu_s^m, W_s^{k,m}, c_k)\| ds && \text{because } H_w \varphi \text{ has bounded norm.}
\end{aligned}
$$

Now, by taking expectation, we can apply Lemma 4.16, which tells us that $\mathbb{E}[\|R(\mu_s^m, W_s^{k,m}, c_k)\|]$ goes to 0 when m goes to infinity, we can bound this last term, and get:

$$\mathbb{E}[|(3)|] \leq \frac{Ct\gamma\varepsilon}{2}.$$

By taking the right ε we conclude:

$$\lim_{m \to \infty} \mathbb{E}[|(3)|] = 0. \tag{4.59}$$

This way:

$$\lim_{m\to\infty}\mathbb{E}[F(\mu^m)] \leq \lim_{m\to\infty}\mathbb{E}[|(1)|] + \lim_{m\to\infty}\frac{\gamma}{2}\mathbb{E}[|(2)|] + \lim_{m\to\infty}\mathbb{E}[|(3)|] + \lim_{m\to\infty}\mathbb{E}[|(4)|] = 0,$$

with which we conclude:

$$\lim_{m\to\infty}\mathbb{E}[F(\mu^m)] = 0,$$

which is exactly what we wanted to prove.

□

Having Lemma 4.16, we must prove that $F(\mu^m)$ converges to $F(\mu)$, with μ being a limit. Nevertheless, we have two problems:

1. We don't know of the existence of such μ.
2. The function for the pair (c, w) given by $c\sigma(w, x)$ is not bounded, and hence in the case where the convergence existed, we couldn't directly use the definition of weak convergence of probability measures.

In the following Lemma we attempt to solve both problems to prove our desired convergence. For simplicity, by an abuse of notation we'll denote any sub-sequence of $(\mu^m)_m$ as $(\mu^m)_m$.

Lemma 4.18 *Let φ and $t \geq 0$ fixed. Then, given a convergent-in-law sub-sequence of μ^m to a measure μ in the space of continuous paths over measures, we have:*

$$\lim_{m\to\infty}\mathbb{E}[F(\mu^m)] = \mathbb{E}[F(\mu)]$$

Proof. For simplicity, we'll prove the case when $\lambda = 0$. The extension to the case when $\lambda > 0$ is straightforward. In the first place, it's important to recall that by Proposition 4.1, the laws of the process of empirical measures are tight in $\mathcal{C}([0,T], \mathcal{M}(\mathbb{R}\times\mathbb{R}^p))$. Hence, there exists a sub-sequence of empirical measures such that their laws are convergent to a fixed distribution, which we'll denote byπ. Let $\mu \sim \pi$ be a probability measure. Then $\mu_t^m \to \mu_t$ in law for all $t \geq 0$. We'll denote the sub-sequence equally by μ^m.

Now, let $\varepsilon > 0$ an let M be a positive constant. We define a modification of $F(\mu) : \mathcal{C}([0,T], \mathcal{M}(\mathbb{R}\times\mathbb{R}^p)) \to \mathbb{R}$ defined in equation (4.82), which we'll denote by $F_M(\mu) : \mathcal{C}([0,T], \mathcal{M}(\mathbb{R}\times\mathbb{R}^p)) \to \mathbb{R}$, which cuts the module of last layer coefficients of the neural network, which the reader may recall are denoted by c, at M:

$$F_M(\mu) = \Big|\langle\varphi, \mu_t - \mu_0\rangle - \int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\mathbb{E}_{X,Y}\left[f_M^\mu(X)((c\wedge M)\vee(-M))\nabla\varphi(c,W)^T\nabla\sigma(W,X)\right]\mu_s(dc,dW)ds \tag{4.60}$$

$$+\gamma\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\mathrm{Tr}\left(S_M(\mu,W,c)^T H_w\varphi(c,W)\right)\mu_s(dc,dW)ds \tag{4.61}$$

$$+\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\mathrm{Tr}\left(H_w\varphi(c,W)\right)\mu_s(dc,dW)ds\Big|, \tag{4.62}$$

where

$$f_M^\mu(X) := \int ((c \wedge M) \vee (-M))\sigma(W,X)\mu_s(dc,dW),$$

and

$$\begin{aligned} S_M(\mu,W,c)_{i,j} = ((c \wedge M) \vee (-M))^2 \Big(&\mathbb{E}_X\left[f_M^\mu(X)^2 \partial_i\sigma(W,X)\partial_j\sigma(W,X)\right] \\ &-\mathbb{E}_X\left[(f_M^\mu(X)\partial_i\sigma(W,X)\right]\mathbb{E}_X\left[(f_M^\mu(X)\partial_j\sigma(W,X)\right]\Big). \end{aligned}$$

The definition at (4.62) allows $F_M(\mu)$ to be a continuous function of μ, since it's bounded. For simplicity, we'll also denote:

$$c^M := ((c \wedge M) \vee (-M)) \text{ and } f^\mu(X) := \int c\sigma(W,X)\mu_s(dc,dW).$$

Now, to begin our study, note that by the triangular inequality we get

$$|\mathbb{E}[F(\mu^m)] - \mathbb{E}[F(\mu)]| \leq |\mathbb{E}[F(\mu^m) - F_M(\mu^m)]| + |\mathbb{E}[F_M(\mu^m)] - F_M(\mu)]| + |\mathbb{E}[F_M(\mu) - F(\mu)]|. \tag{4.63}$$

We must prove that the three terms in the right-hand-side converge to 0 as m diverges to infinity to conclude the proof of the Lemma. We begin with the first term. By the inverse triangular inequality, we have:

$$\begin{aligned} \mathbb{E}[F(\mu^m) - F_M(\mu^m)]| = \mathbb{E}\Bigg[\Big|\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} &\mathbb{E}_{X,Y}\left[(f_M^{\mu_s^m}(X)c^M - f^\mu(X)c)\nabla\varphi(c,W)^T\nabla\sigma(W,X)\right]\mu_s^m(dc,dW)ds \\ &+\gamma\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left((S(\mu^m,W,c) - S_M(\mu^m,W,c))^T H_w\varphi(c,W)\right)\mu_s^m(dc,dW)ds\Big|\Bigg] \\ \leq \mathbb{E}\Bigg[\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} &\mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)c^M - f^{\mu_s^m}(X)c|\|\nabla\varphi(c,W)\|\|\nabla\sigma(W,X)\|\right]\mu_s^m(dc,dW) \\ &+\gamma\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\|S(\mu^m,W,c) - S_M(\mu^m,W,c)\|\|H_w\varphi(c,W)\|\mu_s^m(dc,dW)ds\Bigg]. \end{aligned}$$

Recall that, by our hypothesis, the norm of the gradient and the Hessian of both σ and φ are bounded. Hence,

$$\begin{aligned} |\mathbb{E}[F(\mu^m) - F_M(\mu^m)]| \leq \mathbb{E}\Bigg[C\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} &\mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)c^M - f^{\mu_s^m}(X)c|\right]\mu_s^m(dc,dW)ds \qquad (4.64) \\ &+C\gamma\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\|S(\mu_s^m,W,c) - S_M(\mu_s^m,W,c)\|\mu_s^m(dc,dW)ds\Bigg]. \qquad (4.65) \end{aligned}$$

Let's study the first term of equation (4.65). Now, by triangular inequality, for $X \in \mathcal{X}$

$$\begin{aligned} |f_M^{\mu_s^m}(X)c^M - f^{\mu_s^m}(X)c| &\leq |f_M^{\mu_s^m}(X)c^M - f_M^{\mu_s^m}(X)c| + |f_M^{\mu_s^m}(X)c - f^{\mu_s^m}(X)c| \qquad (4.66) \\ &\leq |f_M^{\mu_s^m}(X)||c^M - c| + |c||f_M^{\mu_s^m}(X) - f^{\mu_s^m}(X)|. \qquad (4.67) \end{aligned}$$

Hence:

$$\int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)c^M - f^{\mu_s^m}(X)c|\right]\mu_s^m(dc,dW) \leq \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)||c^M - c|\right]\mu_s^m(dc,dW) + \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[|c||f_M^{\mu_s^m}(X) - f^{\mu_s^m}(X)|\right]\mu_s^m(dc,dW),$$

Now, by Cauchy-Schwarz inequality:

$$\int \mathbb{E}_{X,Y}\left[|f_M^{\mu^m}(X)||c^M - c|\right]\mu_s^m(dc,dW) \leq \left(\int |c^M - c|^2\mu_s^m(dc)\right)^{\frac{1}{2}}\left(\int \mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)|\right]^2\mu_s^m(dc,dW)\right)^{\frac{1}{2}}. \tag{4.68}$$

Since $c^M - c \leq |c|\mathbb{1}_{|c|>M}$, and $|f_M^{\mu^m}(X)| \leq \int |c|\mu_s^m(dc)$ since σ is bounded, we get

$$\int \mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)||c^M - c|\right]\mu_s^m(dc,dW) \leq \left(\int |c|^2\mathbb{1}_{|c|>M}\mu_s^m(dc)\right)^{\frac{1}{2}}\left(\int\left(\int |c|\mu_s^m(dc)\right)^2\mu_s^m(dc,dW)\right)^{\frac{1}{2}} \tag{4.69}$$

$$\leq C\left(\int |c|^2\mathbb{1}_{|c|>M}\mu_s^m(dc)\right)^{\frac{1}{2}}, \tag{4.70}$$

because c's second moment is bounded. Since $c \in L^1(\mathbb{R},\mu_s(dc))$, if M is big enough we can conclude:

$$\int \mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X)||c^M - c|\right]\mu_s^m(dc,dW) \leq \varepsilon. \tag{4.71}$$

On the other hand, by using Cauchy-Schwarz and the hypothesis that c's second moment is bounded:

$$\int \mathbb{E}_{X,Y}\left[|c||f_M^{\mu_s^m}(X) - f^{\mu^m}(X)|\right]\mu_s^m(dc,dW) \leq C\left(\int \mathbb{E}_{X,Y}\left[|f_M^{\mu_s^m}(X) - f^{\mu_s^m}(X)|\right]^2\mu_s^m(dc,dW)\right)^{\frac{1}{2}}$$
$$\leq C\left(\int |c^M - c|^2\mu_s^m(dc,dW)\right)^{\frac{1}{2}},$$

and again, since $c \in L^1(\mathbb{R},\mu_s(dc))$, if M is big enough we can conclude:

$$\int \mathbb{E}_{X,Y}\left[|c||f_M^{\mu_s^m}(X) - f^{\mu_s^m}(X)|\right]\mu_s^m(dc,dW) \leq \varepsilon. \tag{4.72}$$

By putting together equations (4.71) and (4.72), we can conclude that for if M is big enough

$$C\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[|f_M^{\mu^m}(X)c^M - f^{\mu_s^m}(X)c|\right]\mu_s^m(dc,dW)ds \leq \varepsilon. \tag{4.73}$$

Now, for the second term of equation (4.65) we do the same procedure, and obtain that for M big enough,

$$C\gamma\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p} \|S(\mu^m,W,c) - S_M(\mu^m,W,c)\|\mu_s^m(dc,dW)ds \leq \varepsilon. \tag{4.74}$$

Then, with equations (4.73) and (4.74), we conclude there exists $\tilde{m}$ such that for all $m \geq \tilde{m}$:

$$|\mathbb{E}[F(\mu^m)] - F_M(\mu^m)]| \leq \frac{\varepsilon}{3} \tag{4.75}$$

The second term we want to bound,

$$|\mathbb{E}[F_M(\mu^m)] - F_M(\mu)]|,$$

also converges to 0, since μ^m is weakly convergent to μ, and therefore by being evaluated w.r.t continuous bounded functions, we obtain our desired convergence. Then, there exists $\tilde{\tilde{m}}$ such that for every $m \geq \max\left\{\tilde{m}, \tilde{\tilde{m}}\right\}$, we have:

$$|\mathbb{E}[F(\mu^m)] - F_M(\mu^m)]| \leq \frac{\varepsilon}{3} \text{ and also } |\mathbb{E}[F_M(\mu^m)] - F_M(\mu)]| \leq \frac{\varepsilon}{3} \tag{4.76}$$

At last, let's study what happens with the third term. It's not difficult to note that by inverse triangular inequality:

$$\begin{aligned}|\mathbb{E}[F_M(\mu) - F(\mu)]| \leq & \left|\int_0^t \langle \mathbb{E}_{X,Y}\left[\left(f_M^{\mu_s}(X)c^M - f^{\mu_s}(X)c\right)\varphi(c,W)^T \nabla\sigma(W,X)\right], \mu_s(dc,dW)\rangle ds\right| \\ & + \left|\gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(\left(S(\mu_s,W,c) - S_M(\mu_s,W,c)\right)^T H_w\varphi(c,W)\right) d\mu_s(dc,dW)ds\right|.\end{aligned}$$

Now, we enter the module to the inside of the integrals and bounding the gradients of φ and σ. With this, we eliminate the dependence on X, Y and t, and therefore of the expectation. We obtain:

$$\begin{aligned}|\mathbb{E}[F_M(\mu) - F(\mu)]| \leq & C\int_0^t \int \left(|f_M^{\mu_s}(X)||c^M| - |f^{\mu_s}(X)||c|\right)\mu_s(dc,dW)ds \\ & + C\int_0^t \int \langle\left(f_M^{\mu_s}(X)^2|c^M|^2 - f^{\mu_s}(X)^2|c|^2\right)\mu_s(dc,dW)ds \\ & + C\int_0^t \int \langle\left(f_M^{\mu_s}(X)^2|c^M|^2 - f^{\mu_s}(X)^2|c|^2\right)\mu_s(dc,dW)ds.\end{aligned}$$

By noting that the two last terms are the same:

$$\begin{aligned}|\mathbb{E}[F_M(\mu) - F(\mu)]| \leq & C\int_0^t \int \left(|f_M^{\mu_s}(X)||c^M| - |f^{\mu_s}(X)||c|\right)\mu_s(dc,dW)ds \\ & + C\int_0^t \int \langle\left(f_M^{\mu_s}(X)^2|c^M|^2 - f^{\mu_s}(X)^2|c|^2\right)\mu_s(dc,dW)ds.\end{aligned}$$

By adding and subtracting the same quantities as we did for the first term in equation (4.63), we can see that the next steps are exactly the same as the ones we took for the former, except for the fact that we are now working with μ_s instead of μ_s^m. We conclude, then that for large enough M, the inequality below is satisfied:

$$|\mathbb{E}[F_M(\mu) - F(\mu)]| \leq \frac{\varepsilon}{3}. \tag{4.77}$$

Hence, by putting together (4.75), (4.76) and (4.77), we deduce that there exists taking $\bar{M}$,

such that if we take $M \geq \bar{M}$, there exists $\tilde{m}$ such that for all $m \geq \tilde{m}$:

$$|\mathbb{E}[F(\mu^m)] - \mathbb{E}[F(\mu)]| \leq |\mathbb{E}[F(\mu^m) - F_M(\mu^m)]| + |\mathbb{E}[F_M(\mu^m)] - F_M(\mu)]| + |\mathbb{E}[F_M(\mu) - F(\mu)]| \leq \varepsilon,$$

which can also be written as

$$\lim_{m\to\infty} \mathbb{E}[F(\mu^m)] = \mathbb{E}[F(\mu)],$$

which corresponds exactly to what we wanted to prove. □

Having proved Lemmas 4.17 and 4.18, we can conclude that any sub-sequence μ^{m_k} of μ^m, convergent in law to μ, will satisfy:

$$\mathbb{E}[F(\mu)] = 0. \tag{4.78}$$

Recall that we defined F as $F : \mathcal{P}(\mathbb{R}^p) \to \mathbb{R}$ such that, for $\mu \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^2))$,

$$\begin{aligned} F(\mu) = \Big| &\langle \varphi, \mu_t - \mu_0 \rangle + \lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(W)^T W \, d\mu_s(dc, dW) ds && (4.79)\\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[f^{\mu_s}(X) c \nabla\varphi(W)^T \nabla\sigma(W, X) \right] d\mu_s(dc, dW) ds && (4.80)\\ &+ \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(\mu_s, W, c)^T H_w \varphi(W) \right) d\mu_s(dc, dW) ds && (4.81)\\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w \varphi(W) \right) d\mu_s(dc, dW) ds \Big| . && (4.82) \end{aligned}$$

From (4.82) it's straightforward to note that F is positive. Then, by the result in equation (4.78), we can conclude:

$$F(\mu) = 0 \text{ c.s..}$$

This means that, when $\alpha > 0$, the limiting path in the space of measure for the law of the parameters of the neural network satisfies the limiting PDE given by:

$$\begin{aligned} \langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(W)^T W \, d\mu_s(dc, dW) ds \\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[f^{\mu_s}(X) c \nabla\varphi(W)^T \nabla\sigma(W, X) \right] d\mu_s(dc, dW) ds \\ &+ \lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(c, W)^T W \mu_s(dc, dW) ds \\ &- \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(\mu_s, W, c)^T H_w \varphi(W) \right) d\mu_s(dc, dW) ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w \varphi(W) \right) d\mu_s(dc, dW) ds, \end{aligned}$$

where we defined $f^\mu(X) := \int_{\mathbb{R}\times\mathbb{R}^p} c\sigma(W, X)\mu(dc, dW)$.

4.5. Uniqueness of solutions for the PDE limit

Let $f^\mu(X)$ denote $f^\mu(X) := \int_{\mathbb{R}\times\mathbb{R}^p} c\sigma(W, X)\mu(dc, dW)$. In the last section found that the limit in law of the process of paths over the space of empirical measures $(\mu_s^m)_{t\in[0,T]}^{m\in\mathbb{N}}$ satisfies

the equation (4.83):

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(W)^T W \, d\mu_s(dc, dW)ds \\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[f^{\mu_s}(X) c \nabla\varphi(W)^T \nabla\sigma(W, X) \right] d\mu_s(dc, dW)ds \\ &+ \lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(c, W)^T W \mu_s(dc, dW)ds \\ &- \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(\mu_s, W, c)^T H_w \varphi(W) \right) d\mu_s(dc, dW)ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w \varphi(W) \right) d\mu_s(dc, dW)ds, \end{aligned} \tag{4.83}$$

A direct consequence of the previous section is the fact that the limiting PDE has a solution, which is given by the limit of the process of empirical measures. For this reason, we devote this section to the proof that the PDE has, in fact, a unique solution.

The PDE found in equation (4.83) corresponds to a Non-Linear McKean - Vlasov equation. This equations where first studied by Henry McKean in 1963 in his seminal paper *A class of Markov processes associated with nonlinear parabolic equations* ([18]) and have been subject to deep studies ever since.

Even though the previous result is part of what we were looking for, it's not the end of our quest. This is because, even though it proves convergence of the empirical measure processes to a measure that solves a PDE, nothing is assuring us that such a limiting measure is unique. But, if we could prove that such PDE has a unique solution, then we could prove that this limit is unique. Note that the since the limiting equation is non-linear, proving the uniqueness of solutions is not trivial.

Let's start by remembering the limiting PDE. We proved that if $\alpha = 0$, then

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(W)^T W \, d\mu_s(dc, dW)ds \\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[f^{\mu_s}(X) c \nabla\varphi(W)^T \nabla\sigma(W, X) \right] d\mu_s(dc, dW)ds \\ &+ \lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(c, W)^T W \mu_s(dc, dW)ds \\ &- \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(\mu_s, W, c)^T H_w \varphi(W) \right) d\mu_s(dc, dW)ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w \varphi(W) \right) d\mu_s(dc, dW)ds, \end{aligned} \tag{4.84}$$

and if $\alpha > 0$, then

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0 \rangle = &-\lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(W)^T W d\mu_s(dc, dW)ds \\ &+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[f^{\mu_s}(X)c\nabla\varphi(W)^T\nabla\sigma(W,X)\right] d\mu_s(dc, dW)ds \\ &+ \lambda \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \nabla\varphi(c, W)^T W \mu_s(dc, dW)ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(W)\right) d\mu_s(dc, dW)ds \end{aligned} \tag{4.85}$$

Once again, we'll only have to deal with the case when $\alpha = 0$, the case $\alpha > 0$ will be a direct consequence, since it just suffices to put $\gamma = 0$ in (4.84). In order to prove uniqueness of equation, we must first prove that the limiting equation's terms have the needed regularity for the analogue linear McKean-Vlasov equation to have a unique solution. This is presented in the following:

Lemma 4.19 *For fixed $c \in \mathbb{R}$, consider the function $S(\mu, c, W) : \mathcal{P}(\mathbb{R}^p)\times\mathbb{R}\times\mathbb{R}^p \to \mathcal{M}_{p,p}(\mathbb{R})$ given by*

$$S(\mu, w, c) = c^2\left(\mathbb{E}_X\left[\langle c\sigma, \mu\rangle^2\partial_i\sigma(w, X)\partial_j\sigma(w, X)\right] - \mathbb{E}_X\left[(\langle c\sigma, \mu\rangle\partial_i\sigma(w, X)\right]\mathbb{E}_X\left[(\langle c\sigma, \mu\rangle\partial_j\sigma(w, X)\right]\right)_{i,j\in[p]}$$

Then, the function $S^{\frac{1}{2}}(\mu, c, W)$, which corresponds to taking the square root of the diagonal matrix in the diagonal matrix decomposition of $S(\mu, c, W)$, is Lipschitz in the pair (W, μ), where $S^{\frac{1}{2}}(\mu, c, W)$ corresponds to the square-root matrix of $S(\mu, c, W)$.

Proof. Our proof is based on the arguments in *de Bortoli et al.* [17] and *Stroock and Varadhan* [27].

Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}\times\mathbb{R}^p)$ with the same marginal on c, and

$$\varphi_S(t) = S(t\mu_1 + (1-t)\mu_2, c, t2_1 + (1-t)w_2, c),$$

for fixed $c \in \mathbb{R}$. We denote $\mu_t := t\mu_1 + (1-t)\mu_2$ and $w_t := tw_1 + (1-t)w_2$. This proof will be divided in two parts. In the first one, we'll prove:

$$|\varphi_S''| \leq C\left(|w_2 - w_1|^2 + \mathcal{W}(\mu_1, \mu_2)^2\right), \tag{4.86}$$

and in the second one we'll conclude.
·**First Part:** Since we aim to bound φ_S's second derivative, we must begin by showing that φ_S is a $C^2([0,1], \mathbb{R})$ function. For this, we'll use the Dominated Convergence Theorem.

Following [17], we define:

$$g(t, x) = f^{\mu_t}(x)c\nabla\sigma(w_t, x) \text{ and } \tilde{g}(t, x) = \mathbb{E}_x[g(t, x)] - g(t, x).$$

Note that

$$|g(t, x)| \leq \langle|\tilde{c}||\sigma(\tilde{w}, x)|, \mu_t\rangle|c|||\nabla\sigma(w_t, x)|||,$$

and using the fact that both σ and the norm of $\nabla\sigma(w_t, x)$ are bounded, we obtain:

$$|g(t,x)| \leq C\langle|\tilde{c}|, \mu_t\rangle|c| \leq C\mathbb{E}_{\tilde{c}}[\tilde{c}^2]|c| \leq C|c|. \tag{4.87}$$

Additionally,

$$\begin{aligned}\|\partial_t g(t,x)\| &= |c\nabla\sigma(w_t,x)\left(\langle\tilde{c}\sigma(\tilde{w},X),\mu_1\rangle - \langle\tilde{c}\sigma(\tilde{w},X),\mu_2\rangle\right) + \langle\tilde{c}\sigma(\tilde{w},X),\mu_t\rangle c\mathcal{H}_w\sigma(w_t,x)(w_2-w_1)| \\ &\leq C|c|\,|\langle\tilde{c}\sigma(\tilde{w},X),\mu_1\rangle - \langle\tilde{c}\sigma(\tilde{w},X),\mu_2\rangle| + C\|w_2-w_1\|. \end{aligned}\tag{4.88}$$

Now, let ν denote an optimal coupling between μ_1 and μ_2 (See [32]). We get:

$$|\langle\tilde{c}\sigma(\tilde{w},X),\mu_1\rangle - \langle c\sigma(w,X),\mu_2\rangle| \leq \int |c\sigma(w,X) - c'\sigma(w',X)|\nu(dc,dw,dc',dw'),$$

and using the fact that given x, $w \to c\sigma(w,x)$ is Lipschitz:

$$|\langle\tilde{c}\sigma(\tilde{w},X),\mu_1\rangle - \langle c\sigma(w,X),\mu_2\rangle| \leq C\int |(c,w)-(c',w')|\nu(dc,dw,dc',dw').$$

Now we apply Jensen's inequality and obtain:

$$|\langle\tilde{c}\sigma(\tilde{w},X),\mu_1\rangle - \langle c\sigma(w,X),\mu_2\rangle| \leq C\mathcal{W}_2(\mu_1,\mu_2).$$

Replacing this in equation (4.88):

$$\|\partial_t g(t,x)\| \leq C|c|\left(\mathcal{W}_2(\mu_1,\mu_2) + \|w_2-w_1\|\right). \tag{4.89}$$

Now, for $|\partial_t^2 g(t,x)|$, we use the previous result inequality for $\|\partial_t g(t,x)\|$:

$$\begin{aligned}\|\partial_t^2 g(t,x)\| &\leq \|c\mathcal{H}_w\sigma(w_t,x)(w_2-w_1)\left(\langle\tilde{c}\sigma(\tilde{w},X),\mu_1-\mu_2\rangle\right) \\ &\quad +\langle\tilde{c}\sigma(\tilde{w},X),\mu_1-\mu_2\rangle c\mathcal{H}_w\sigma(w_t,x)(w_2-w_1) \\ &\quad +\langle\tilde{c}\sigma(\tilde{w},X),\mu_t\rangle c\partial_t\mathcal{H}_w\sigma(w_t,x)(w_2-w_1)\| \\ &\leq |2\langle\tilde{c}\sigma(\tilde{w},X),\mu_1-\mu_2\rangle c\mathcal{H}\sigma(w_t,x)(w_2-w_1) + \langle\tilde{c}\sigma(\tilde{w},X),\mu_t\rangle c\langle\tilde{c}\sigma(\tilde{w},X),\mu_t\rangle c\partial_t\mathcal{H}_w\sigma(w_t,x)(u \\ &\leq C|c|2W_2(\mu_1,\mu_2)\|w_2-w_1\| + C|c|\|w_2-w_1\|^2,\end{aligned}$$

and using the classical inequality $2ab \leq a^2+b^2$:

$$|\partial_t^2 g(t,x)| \leq C|c|\left(\mathcal{W}_2^2(\mu_1,\mu_2) + \|w_2-w_1\|^2\right). \tag{4.90}$$

With this, by the Theorem of Dominated Convergence, we conclude that $\tilde{f}(t,x) \in \mathcal{C}^2([0,1],\mathbb{R})$, and with that, given that

$$\varphi_S(t) = \mathbb{E}_{X,Y}[\tilde{g}(t,x)^T\tilde{g}(t,x)],$$

we can conclude that $\varphi_S(t) \in \mathcal{C}^2([0,1],\mathbb{R})$ (applying once again the Theorem of Dominated Convergence). At last, we note that:

$$\begin{aligned}\|\varphi_S^{''}(t)\| &= \|\partial_t\mathbb{E}_X[\tilde{g}(t,x)^T\tilde{g}(t,x)]\| \\ &= \|\mathbb{E}_X[(\partial_t\tilde{g}(t,x))^T\tilde{g}(t,x) + g(t,x)^T(\partial_t^2 g(t,x))]\| \\ &\leq C|c|\left(\|w_2-w_1\|^2 + \mathcal{W}_2^2(\mu_1,\mu_2)\right).\end{aligned}$$

·**Second Part:** Now let's prove, using the first part, that $S^{\frac{1}{2}}(c, w, \mu)$ is Lipschitz.

Note that:

$$\|\varphi''(t)\| = \sup_{x\in\mathbb{R}^p} \langle x, \varphi''(t)x\rangle \leq C\left(\|w_2 - w_1\|^2 + \mathcal{W}_2^2(\mu_1, \mu_2)\right).$$

By using a direct modification of Theorem 5.2.3 and Lemma 3.2.3 of [27], as done in [17], we conclude than $S^{\frac{1}{2}}$ is Lipschitz. □

Remark By similar arguments, it's clear to prove that $\Sigma^m(W^{k,m})$ is Lipschitz for W^m for all $m \in \mathbb{N}$. This allows us to prove that the SDE in the core of this work has a solution.

Note that in the associated McKean-Vlasov equation, it won't be $S^{\frac{1}{2}}$ but $S^{\frac{1}{2}} + \mathcal{I}$ that will be on the stochastic process's diffusion coefficient, but since $\mathcal{I}^{\frac{1}{2}} = \mathcal{I}$, and the identity is trivially Lispchitz, a direct consequence is the following Lemma.

Lemma 4.20 *For fixed $c \in \mathbb{R}$, the function $S^{\frac{1}{2}}(\mu, c, W) + \mathcal{I}$ is Lipschitz in the pair (W, μ), where $S^{\frac{1}{2}}(\mu, c, W)$ corresponds to the square-root matrix of $S(\mu, c, W)$.*

We already have our desired conditions for the diffusion term. Now, let's prove the desired conditions for the drift.

Lemma 4.21 *Given $c \in \mathbb{R}$, the function*

$$b(c, W, \mu) = -\lambda W + \mathbb{E}_{X,Y}\left[\langle c\sigma(\cdot, X), \mu_t\rangle c\nabla\sigma(W, X)\right]\nabla\varphi(W)$$

is Lipschitz in the pair (W, μ).

Proof. Note that the first term is linear on W, so it's enough to prove that the second term is Lipschitz. For this, let $w_1, w_2 \in \$$, $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^2)$. Since c is fixes, we'll omit b's dependence on our notation. We have:

$$\begin{aligned}
\|b(w_1, \mu_1) - b(w_2, \mu_2)\ | &\leq \|\mathbb{E}_{X,Y}\left[\langle c\sigma(\cdot, X), \mu_1\rangle c\nabla\sigma(w_1, X)\nabla\varphi(w_1)\right. \\
&\quad \left. -\langle c\sigma(\cdot, X), \mu_2\rangle c\nabla\sigma(w_2, X)\nabla\varphi(w_2)\right]\| \\
&\leq \underbrace{\|\mathbb{E}_{X,Y}\left[\langle c\sigma(\cdot, X), \mu_1\rangle\left(c\nabla\sigma(w_1, X)\nabla\varphi(w_1) - c\nabla\sigma(w_2, X)\nabla\varphi(w_2)\right)\right]\|}_{A} \\
&\quad + \underbrace{\|\mathbb{E}_{X,Y}\left[\left(\langle c\sigma(\cdot, X), \mu_1\rangle - \langle c\sigma(\cdot, X), \mu_2\rangle\right)c\nabla\sigma(w_2, X)\nabla\varphi(w_2)\right]\|}_{B}.
\end{aligned}$$

Let's see that A y B have bounds such that $b(c, W, \mu)$ is Lipschitz.

Given that $\mathcal{H}_w\sigma$ has a bounded norm, we obtain

$$|c\nabla\sigma(w_1, x) - c\nabla\sigma(w_2, x)| \leq L|c|\|w_1 - w_2\|.$$

This way, we get:

$$A \leq C|c|\|w_1 - w_2\|. \tag{4.91}$$

On the other hand,

$$B \leq C|c|\mathbb{E}_{X,Y}[|\langle c\sigma(\cdot, X), \mu_1\rangle - \langle c\sigma(\cdot, X), \mu_2\rangle|]. \tag{4.92}$$

Now, let Π be an optimal coupling between μ_1 y μ_2. Then:

$$\begin{aligned}
B &\leq C|c|\mathbb{E}_{X,Y}\left[\int c\sigma(w, X) - c'\sigma(w', X)d\Pi(dc, dw, dc', dw')\right] \\
&\leq C|c|\mathbb{E}_{X,Y}\left[\int \|(c, w) - (c', w')\|d\Pi(dc, dw, dc', dw')\right] && \text{because } c\sigma \text{ is uniformly Lipschitz} \\
&\leq C|c|\left(\int \|(c, w) - (c', w')\|^2 d\Pi(dc, dw, dc', dw')\right)^{\frac{1}{2}} && \text{by Cauchy-Schwarz} \\
&\leq C|c|\mathcal{W}_2(\mu_1, \mu_2). && (4.93)
\end{aligned}$$

By putting together (4.91) y (4.93):

$$|b(c, w_1, \mu_1) - b(c, w_2, \mu_2)| \leq C|c|\left(\|w_1 - w_2\| + \mathcal{W}_2(\mu_1, \mu_2)\right),$$

which is exactly what we wanted to prove. □

With the two Lemmas we just proved, we already know that the coefficients $b(c, w, \mu)$ and $S^{\frac{1}{2}}(c, w, \mu)$ are Lipschitz in (w, μ). Now, given $(\mu_t)_{t\in[0,T]}$, a continuous path in measure space, consider the linear equation (4.94):

$$\begin{aligned}
\langle \varphi, \nu_t - \nu_0\rangle = &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} b(x, \mu_s)\nabla\varphi(c, W)\nu_s(dc, dW)ds \\
&+ \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(S(x, \mu_s)^T \mathcal{H}_w\varphi(c, W)\right)\nu_s(dc, dW)ds \\
&+ \sqrt{2\tau}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(c, W)\right)d\nu_s(dc, dW)ds.
\end{aligned} \tag{4.94}$$

This equation is different from our non-linear McKean-Vlasov equation in (4.83) because the first term is linear, hence it's a linear McKean Vlasov equation, or a Focker Planck equation. By proving that this equation has a solution in our context, we'll be able to extend this results to the non-linear case. For this, we'll need following Theorem, which gives us an existence and uniqueness results for the linear PDE (4.94).

Lemma 4.22 *If the coefficients b and $S^{\frac{1}{2}}$ are Lipschitz in (w, μ), then the linear Focker-Planck equation has a unique solution.*

Proof. See [33], Theorem 1.1, or [16] Theorem 2.2. □

A direct consequence 4.22 is the following proposition.

Proposition 4.2 *The linear Focker-Planck equation associated to our problem,*

$$\begin{aligned}
\langle \varphi, \nu_t - \nu_0\rangle = &- \int_0^t \langle b(x, \mu_s)\nabla\varphi(\cdot), \nu_s\rangle ds + \int_0^t \langle \mathrm{Tr}\left(S(x, \mu_s)^T \mathcal{H}_w\varphi(W, x)\right), \nu_s\rangle ds \\
&+ \sqrt{2\tau}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right)d\nu_s(d\tilde{c}, d\tilde{W})ds.
\end{aligned}$$

has a unique solution.

PROOF. It suffices to use Lemmas 4.20 and 4.21, which prove that both coefficients b and $S^{\frac{1}{2}}$ are Lipschitz. We conclude by applying the results in Lemma 4.22. □

We finish this section with our uniqueness result. For this, we'll need the following Lemma, which gives us uniqueness of solutions (in Law) of the corresponding SDE of our non-linear McKean Vlasov Equation.

Proposition 4.3 *Let $\sigma : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be a bounded function with the norm of it's gradient and it's Hessian bounded. Let f^μ denote the shallow neural network given by:*

$$f^\mu(X) := \int_{\mathbb{R}\times R^p} c\sigma(W, X)\mu(dc, dW).$$

Let $c \in \mathbb{R}$ be normal coefficients that are not trained, but initialized as in the assumptions. Then, the solution of the stochastic differential equation:

$$\begin{aligned} W_t = W_0 &+ \int_0^t \int_{\mathbb{R}^p} \mathbb{E}_{X,Y}\left[\left(\langle c\sigma, \mu_t\rangle\right) c\nabla\sigma(\cdot, X)\nabla\varphi(W_s)\right]\mu_s(dc, dw)ds \\ &+ \int_0^t S^{\frac{1}{2}}(c, W_s, \mu_s)dBs + \sqrt{2\tau}\int_0^t d\tilde{B}s, \end{aligned} \tag{4.95}$$

which correspond to Stochastic Gradient Descent in the Mean Field setting in continuous time, is unique in Law.

PROOF. This proof is based on the arguments presented in Theorem 1.1 and Lemma 1.3 in Snitzman's book [16], and also in [34]. We denote

$$\tilde{b}(W_s, c, w, \mu) = \mathbb{E}_{X,Y}\left[\left(\langle c\sigma, \mu_t\rangle\right) c\nabla\sigma(\cdot, X)\nabla\varphi(W_s)\right].$$

Note that as a consequence of the gradient of φ being Lipschitz, $\tilde{b}(W_s, c, w, \mu)$ is also Lipschitz in W_s. Also, by Lemma 4.21 we know that it's also Lipschitz in the par (w, μ). Now, let $(\mu_t^1)_{t\in[0,T]}, (\mu_t^2)_{t\in[0,T]} \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R} \times \mathbb{R}^d)$, such that their first marginal (i.e the one corresponding to c) is the same. We define:

$$W_t^1 = W_0 + \int_0^t \int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^1)\mu_s^1(dc, dw)ds + \int_0^t S^{\frac{1}{2}}(c, W_s^1, \mu_s^1)dBs + \sqrt{2\tau}\int_0^t d\tilde{B}s,$$

and

$$W_t^2 = W_0 + \int_0^t \int_{\mathbb{R}^p} b(W_s^2, c, w, \mu_s^1)\mu_s^2(dc, dw)ds + \int_0^t S^{\frac{1}{2}}(c, W_s^2, \mu_s^2)dBs + \sqrt{2\tau}\int_0^t d\tilde{B}s.$$

We'll add a supremum in the interior of the Wasserstein's metric, which also results in a well-defined complete metric in $\mathcal{P}_2(\mathcal{C}([0, T], \mathcal{P}(\mathbb{R} \times \mathbb{R}^d))$. This metric is called Kantorovich - Rubinstein's metric. In this case:

$$\mathcal{W}(\mu_s^1, \mu_s^2) \leq \mathcal{W}(\mu_t^1, \mu_t^2)$$

if $s \leq t$. We begin with the following straightforward calculation:

$$\mathbb{E}\left[\sup_{t\leq T}\|W_t^1 - W_t^2\|\|^2\right] \leq C\underbrace{\int_0^T \mathbb{E}\left[\left|\int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^1)\mu_s^1(dc, dw) - \int_{\mathbb{R}^p} b(W_s^2, c, w, \mu_s^2)\mu_s^2(dc, dw)\right|^2\right] ds}_{(1)}$$
$$+ \underbrace{C\mathbb{E}\left[\sup_{t\leq T}\left(\int_0^t (S^{\frac{1}{2}}(c, W_s^1, \mu_s^1) - S^{\frac{1}{2}}(c, W_s^2, \mu_s^2))dBs\right)^2\right]}_{(2)}.$$

Let's study (1) and (2) separately, beginning with (1). We have:

$$\begin{aligned}
&\int_0^t \left|\int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^1)\mu_s^1(dc, dw) - \int_{\mathbb{R}^p} b(W_s^2, c, w, \mu_s^2)\mu_s^2)\right| ds \\
&\leq \int_0^t \left|\int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^1)\mu_s^1(dc, dw) - \int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^2)\mu_s^1(dc, dw)\right| ds \\
&+ \int_0^t \left|\int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^2)\mu_s^1(dc, dw) - \int_{\mathbb{R}^p} b(W_s^2, c, w, \mu_s^2)\mu_s^2)\right| ds. \qquad (4.96)
\end{aligned}$$

Let $\tilde{\mu}_s$ be any coupling between μ_s^1 and μ_s^2. Then, for the first term in (4.96), note that:

$$\begin{aligned}
&\int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^1)\mu_s^1(dc, dw) - \int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^2)\mu_s^1(dc, dw) \\
&\qquad = \int_{\mathbb{R}^p} |b(W_s^1, c, w^1, \mu_s^1) - b(W_s^1, c, w^2, \mu_s^2)|\tilde{\mu}_s(dc, dw^1, dw^2) \\
&\qquad \leq K\int_{\mathbb{R}^p} |c|(\|w^1 - w^2\| + \mathcal{W}_2(\mu_s^1, \mu_s^2))\tilde{\mu}_s(dc, dw^1, dw^2) \\
&\qquad \leq K\left(\int_{\mathbb{R}^p} (\|w^1 - w^2\| + \mathcal{W}_2(\mu_s^1, \mu_s^2))^2\tilde{\mu}_s(dc, dw^1, dw^2)\right)^{\frac{1}{2}},
\end{aligned}$$

by Lemma 4.21 and then using Cauchy-Schwarz. By remembering that $(a+b)^2 \leq 2(a^2+b^2)$, we obtain:

$$\begin{aligned}
&\int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^1)\mu_s^1(dc, dw) - \int_{\mathbb{R}^p} b(W_s^1, c, w, \mu_s^2)\mu_s^1(dc, dw) \\
&\qquad \leq K\left(\int_{\mathbb{R}^p} \|w^1 - w^2\|^2\tilde{\mu}_s(dc, dw^1, dw^2)\right)^{\frac{1}{2}} + K\mathcal{W}_2(\mu_s^1, \mu_s^2) \\
&\qquad \leq K\mathcal{W}_2(\mu_s^1, \mu_s^2).
\end{aligned}$$

On the other hand, for the second term in (4.96):

$$\begin{aligned}\left|\int_{\mathbb{R}^p} b(W_s^1,c,w,\mu_s^2)\mu_s^1(dc,dw) - \int_{\mathbb{R}^p} b(W_s^2,c,w,\mu_s^2)\mu_s^2(dc,dw)\right| & \\ \leq \int_{\mathbb{R}^p} |b(W_s^1,c,w^1,\mu_s^2) - b(W_s^2,c,w^2,\mu_s^2)|\tilde{\mu}_s(dc,dw_1^1dw^2)ds & \\ \leq \int_{\mathbb{R}^p} |b(W_s^1,c,w^1,\mu_s^2) - b(W_s^1,c,w^2,\mu_s^2)|\tilde{\mu}_s(dc,dw_1^1dw^2)ds & \\ + \int_{\mathbb{R}^p} |b(W_s^1,c,w^2,\mu_s^2) - b(W_s^2,c,w^2,\mu_s^2)|\tilde{\mu}_s(dc,dw_1^1dw^2)ds & \\ \leq K\mathcal{W}(\mu_s^1,\mu_s^2) + K\sup_{u\leq s}\|W_u^1 - W_u^2\|^2, &\end{aligned}$$

by using the same techniques we use for the first term. Going back to (1), we conclude:

$$\begin{aligned}\int_0^T \mathbb{E}\left[\left|\int_{\mathbb{R}^p} b(W_s^1,c,w,\mu_s^2)\mu_s^1(dc,dw) - \int_{\mathbb{R}^p} b(W_s^2,c,w,\mu_s^2)\mu_s^2(dc,dw)\right|^2\right] ds & \\ \leq C\int_0^T \mathbb{E}\left[K\mathcal{W}(\mu_s^1,\mu_s^2)^2 + K\sup_{s\leq t}|W_s^1 - W_s^2|^2\right] ds & \\ \leq C\int_0^T \mathbb{E}\left[K\mathcal{W}(\mu_s^1,\mu_s^2)^2\right] + KC\int_0^T \mathbb{E}\left[\sup_{u\leq s}\|W_u^1 - W_u^2\|^2\right] ds. &\end{aligned}$$

Having this estimate, we'll now study (2). We have:

$$(2) = C\mathbb{E}\left[\sup_{t\leq T}\left(\int_0^t (S^{\frac{1}{2}}(c,W_s^1,\mu_s^1) - S^{\frac{1}{2}}(c,W_s^2,\mu_s^2))dBs\right)^2\right],$$

and by using BDG's inequality, and then using that by Lemma 4.21, $S^{\frac{1}{2}}$ is Lipschitz:

$$\begin{aligned}(2) &= C\mathbb{E}\left[\int_0^t (S^{\frac{1}{2}}(c,W_s^1,\mu_s^1) - S^{\frac{1}{2}}(c,W_s^2,\mu_s^2))^2\right] ds & \\ &\leq K\mathbb{E}\left[\int_0^t (\|W_s^1 - W_s^2\|^2 + \mathcal{W}(\mu_s^1,\mu_s^2)^2)\right] ds & \text{Because } S^{\frac{1}{2}} \text{ is Lipschitz.} \\ &\leq K\mathbb{E}\left[\int_0^t \|W_s^1 - W_s^2\|^2\right] ds + K\mathbb{E}\left[\int_0^t \mathcal{W}(\mu_s^1,\mu_s^2)^2\right] ds & \\ &\leq K\int_0^t \mathbb{E}\left[\sup_{u\leq s}\|W_u^1 - W_u^2\|^2\right] + K\int_0^t \mathbb{E}\left[\mathcal{W}(\mu_s^1,\mu_s^2)^2\right] ds & \text{By Fubini's theorem.} \\ &\leq K\int_0^t \mathbb{E}\left[\sup_{u\leq s}\|W_u^1 - W_u^2\|^2\right] + K\int_0^t \mathbb{E}\left[\mathcal{W}(\mu_s^1,\mu_s^2)^2\right] ds & \end{aligned}$$

By putting all together, we obtain:

$$\mathbb{E}\left[\sup_{s\leq t}\|W_t^1 - W_t^2\|^2\right] \leq K\int_0^t \mathbb{E}\left[\sup_{u\leq s}\|W_u^1 - W_u^2\|^2\right] ds + K\int_0^t \mathbb{E}\left[\mathcal{W}(\mu_s^1,\mu_s^2)^2\right] ds. \quad (4.97)$$

Now, by applying Gronwall's Lemma:

$$\mathbb{E}\left[\sup_{s\leq t}\|W_t^1 - W_t^2\|^2\right] \leq Ke^T\left(\int_0^t \mathbb{E}\left[\mathcal{W}(\mu_s^1,\mu_s^2)^2\right] ds\right). \quad (4.98)$$

Now, let $\Phi : \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}\times\mathbb{R}^d) \to \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}\times\mathbb{R}^d)$, such that for $(m_t) \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}\times\mathbb{R}^d)$, $\Phi(m)$ corresponds to the law of the process:

$$X_t = X_0 + \int_0^t \int_{\mathbb{R}^p} b(X_s, c, w, m_s)\mu_s(dc, dw)ds + \int_0^t S^{\frac{1}{2}}(c, X_s, m_s)dBs + \sqrt{2\tau}\int_0^t d\tilde{B}_s.$$

Note that if X_t is a solution of 4.95, then it's law it's a fixed point for Φ. Using this definition, then the result in equation 4.98 allows us to conclude:

$$\mathbb{E}\left[\mathcal{W}(\Phi(\mu_t^1), \Phi(\mu_t^2))^2\right] \leq Ke^T\left(\int_0^t \mathbb{E}\left[\mathcal{W}(\mu_s^1, \mu_s^2)^2\right] ds\right). \tag{4.99}$$

Since a solution is also a fixed point, we can iterate and obtain:

$$\mathbb{E}\left[\mathcal{W}(\Phi^k(\mu_t^1), \Phi^k(\mu_t^2))^2\right] \leq \frac{Ce^{Tk}}{k!}\left(\mathbb{E}\left[\mathcal{W}(\mu_t^1, \mu_t^2)^2\right] ds\right), \tag{4.100}$$

and hence conclude, by making k sufficiently big, that in fact $\mu_1 = \mu_2$, i.e that the solutions for the SDE are unique in Law. □

Theorem 4.3 *The Focker-Planck equation*

$$\begin{aligned}\langle\varphi, \mu_t - \mu_0\rangle = &-\lambda\int_0^t \nabla\varphi^T\tilde{W}\mu_s(d\tilde{c}, d\tilde{W})ds + \int_0^t \mathbb{E}_{X,Y}\left[\left(\langle c\sigma, \mu_t\rangle\right)\langle c\nabla\sigma(\cdot, X)\nabla\varphi, \mu_s\rangle\right] ds \\ &+ \gamma\int_0^t \langle \mathrm{Tr}\left(S(x,\mu_s)^T\mathcal{H}_w\varphi(\cdot)\right), \mu_s\rangle ds + \sqrt{2\tau}\int_0^T\int_{\mathbb{R}\times\mathbb{R}^p} \mathrm{Tr}\left(H_w\varphi(\tilde{W})\right) d\mu_s(d\tilde{c}, d\tilde{W})ds\end{aligned}$$

with μ_0 as the initial condition, has a unique solution.

Proof. Let μ be a solution of the non-linear PDE equation, which we know that exists as a consequence of the previous section.

On the other hand, let ν be a solution to the unique linear PDE equation, which we know that exists as a consequence of Lemma 4.2. That is, ν is the unique solution of

$$\begin{aligned}\langle\varphi, \nu_t - \nu_0\rangle = &-\lambda\int_0^t \nabla\varphi(\tilde{W})^T\tilde{W}\nu(d\tilde{c}, d\tilde{W})ds + \int_0^t \langle b(x,\mu_s)\nabla\varphi(\cdot), \nu_s\rangle ds \\ &+ \int_0^t \langle \mathrm{Tr}\left(S(x,\mu_s)^T\mathcal{H}_w\varphi(\cdot)\right), \nu_s\rangle ds + \sqrt{2\tau}\int_0^t \langle \mathrm{Tr}\left(\mathcal{H}_w\varphi(\cdot)\right), \nu_s\rangle ds.\end{aligned} \tag{4.101}$$

Since the linear equation (4.5) has a unique solution, and we know μ_t is a solution to the nonlinear equation, then μ_t has to be the only solution to the equation , therefore there exists a unique stochastic process X_t such that $\mathcal{L}(X_t) = \nu_t = \mu_t$.

Then $(\mathcal{L}(X_t))_{t\in[0,T]}$ is a fixed point, and hence X_t solves the corresponding non-linear Stochastic Differential Equation. Since the solutions of our SDE are unique in Law, by Proposition 4.3, $(\mathcal{L}(X_t))_{t\in[0,T]} = (\mu_t)_{t\in_{[0,T]}}$ has to be the unique solution to the associated non-linear PDE, which concludes our theorem. □

Having proved uniqueness of solutions, we can state the following Theorem, which is one of our main results.

Theorem 4.4 *Let $\alpha > 0$, $\lambda \in [0,1)$, $\gamma \geq 0$, and μ_t^m denote the empirical measure process that represents the weights of a shallow neural network, who's parameters are trained in continuous time by the dynamics:*

$$dW_t^{k,m} = h^{k,m}(W_t^m)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m},$$

where $W_t^{k,m}$ denotes one neuron in the hidden layer. Let μ_0 denote the initialization distribution for the pair (C,W). Then, in the limit as m goes to infinity, the empirical measure converges in Law to the unique solution of the non-linear Focker Planck Equation:

- *If $\alpha = 0$:*

$$\begin{aligned}\langle \varphi, \mu_t - \mu_0\rangle = &-\lambda \int_0^t \nabla\varphi(\tilde{W})^T\tilde{W}\mu_s(d\tilde{c},d\tilde{W})ds + \int_0^t \mathbb{E}_{X,Y}\left[\left(\langle c\sigma, \mu_t\rangle\right)\langle c\nabla\sigma(\cdot,X)\nabla\varphi, \mu_s\rangle\right]ds \\ &+ \gamma\int_0^t \langle \mathrm{Tr}\left(S(x,\mu_s)^T\mathcal{H}_w\varphi(\cdot)\right), \mu_s\rangle ds + \sqrt{2\tau}\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\mathrm{Tr}\left(H_w\varphi(\tilde{W})\right)d\mu_s(d\tilde{c},d\tilde{W})ds.\end{aligned} \tag{4.102}$$

- *If $\alpha > 0$:*

$$\langle \varphi, \mu_t - \mu_0\rangle = -\lambda \int_0^t \nabla\varphi(\tilde{W})^T\tilde{W}\mu_s(d\tilde{c},d\tilde{W})ds + \sqrt{2\tau}\int_0^t\int_{\mathbb{R}\times\mathbb{R}^p}\mathrm{Tr}\left(H_w\varphi(\tilde{W})\right)d\mu_s(d\tilde{c},d\tilde{W})ds. \tag{4.103}$$

Perhaps the most surprising result of this equation is the fact that it states that the empirical measure does not see training data in this scale, which is reflected by the absence of Y in the equation. Note that this is what one would expect from the results in [22]. Yet, in order to precisely arrive to this results, one has to put special parameters in this equation.

Another remarkable property is the fact that, in the general case, an analytic solution is not straightforward: the fact that Langevin Dynamics add an extra regularization on the parameters makes the analysis harder. Another important fact is that by the result in equation (4.103), we conclude that in this setting the limiting measure does not "see" Stochastic Gradient Descent, but only a regularized version of Gradient Descent.

4.6. The solution with Xavier Initialization and a Construction of the NTK

Having proved our main result in the last section, in this section we study the different solutions in different cases. We'll study the special case of Xavier Initialization, Xavier initialization is one of the most widely used initializations for neural networks. It consists in initializing parameters (c,w) independently, and the last layer as centered gaussians.

Through this section, we'll only consider the case when $\lambda = \tau = 0$. This means we are training the neural network with the classic stochastic gradient descent method. The extension to Langevin Dynamics will be a direct extension.

Note that in this cases, Theorem 4.4 states that the limiting measure follows the dynamics:

$$\langle \varphi, \mu_t - \mu_0 \rangle = \int_0^t \mathbb{E}_{X,Y} \left[(\langle c\sigma, \mu_t \rangle) \langle c \nabla \sigma(\cdot, X) \nabla \varphi, \mu_s \rangle \right] ds + \gamma \int_0^t \langle \mathrm{Tr} \left(S(x, \mu_s)^T \mathcal{H}_w \varphi(\cdot) \right), \mu_s \rangle ds + \sqrt{2\tau} \int_0^T \int_{\mathbb{R} \times \mathbb{R}^p} \mathrm{Tr} \left(H_w \varphi(\tilde{W}) \right) d\mu_s(d\tilde{c}, d\tilde{W}) ds. \tag{4.104}$$

In our setting, Xavier initialization is equivalent to considering the initialization:

$$c \sim \mathcal{N}(0, 1),$$

and W independent to c, as a centered distribution with it's four first moments being finite. We get the following result, which also appears in [11] and in [35]. Note that it's statement cna be directly linked to the results in [22], which state that in the Nerual Tangent Kernel Regime, the parameters tend to stay close to it's initialization distribution. This proves this fact in a different way.

Corolary 4.1 *If $\lambda = \tau = 0$ and $\mathbb{E}[c] = 0$, then equation 4.5 becomes:*

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle. \tag{4.105}$$

Proof. Note that if $\mathbb{E}[c] = 0$, then all terms are 0: Since at initialization the parameters are independent, if we replace μ_s by μ_0, then all terms integrated by μ_0 are split into two terms, one of them being 0, and by multiplying all the parameters are also 0. Hence μ_0 solves the PDE. To conclude this is the unique solution, we apply Theorem 4.3, which states the uniqueness of solutions.

□

This result confirms what we were expecting: as m grows, the parameters enter the Lazy Regime, as described in [22]. The authors of the technical note [11] and of [35] also arrive to this result, yet they use different techniques: They pass from discrete to continuous time using Taylor approximations (which, in fact, corresponds to applying Itô's Lemma in a discrete setting), and they go through another setting first in [35]. On the other hand, they don't get the general PDE for the empirical measure.

Now, one can ask: So, what is the NTK? The answer to this question lies in Corollary . If we consider $\varphi = \sigma$, then we get that given $x_1, x_2 \in \mathcal{X}$, the following convergence is satisfied in law:

$$\langle c^2 \nabla \sigma(\tilde{W}, x_1)^T \nabla \sigma(\tilde{W}, x_2), \mu_s^m \rangle \overset{m \to \infty}{\longrightarrow} \langle c^2 \nabla \sigma(\tilde{W}, x_1)^T \nabla \sigma(\tilde{W}, x_2), \mu_0 \rangle.$$

Considering this, we define the continuous version of the **Neural Tangent Kernel**, first defined in [19], associated to our neural network as the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that

$$K(x_1, x_2) = \langle c^2 \nabla \sigma(\tilde{W}, x_1)^T \nabla \sigma(\tilde{W}, x_2), \mu_0 \rangle. \tag{4.106}$$

Another important remark is that, as it can be seen in Sirignano and Spiliopoulos' work

in [11], the analysis proves that only $\alpha \geq \frac{1}{2}$ allows one to study the limiting dynamic. This is true in the general case, because the regularization terms does not allow us to study the equation when the step size is bigger.

Chapter 5

The Neural Network with Xavier Initialization and SGD training

In the last section we gave a complete study of the dynamic of the empirical measure in the *Lazy Regime* of Neural Networks trained by Stochastic Gradient Descent. In the present chapter, our aim will be to study the dynamic of the neural network itself. Nevertheless, as the reader may note, the study of empirical measures is not sufficient for this objective: Since in our parametrization of f^m does not depend exactly on the empirical measure but on a scaling of it, it could potentially be non-convergent.

As it can be seen in the results in the last section, the case of non- centered initializations of c and $\alpha < \frac{1}{2}$ is not tractable: The non linear terms of the Focker Planck Equation are non-convergent in this scale. For this reason, through this section we consider the case of $\alpha \geq \frac{1}{2}$. We'll also consider $\tau = \lambda = 0$, i.e the usual stochastic gradient descent setting.

As in the final part of Chapter 4, we'll consider the Xavier initialization setting, which was first studied in [36]. In our setting is equivalent to considering

$$c \sim \mathcal{N}(0, 1),$$

and W independent to c, as a centered distribution with it's four first moments being finite. As we said before, this particular setting was studied before in [11], yet our techniques are different from their, since we based our analysis in the continuous time equivalents of stochastic gradient descent.
In section 1, we'll remember the dynamics we found in the last chapter, and we'll give special attention to a a non-linear term of our resulting dynamic, which will determine if there exists (or not) a convergence of the evaluations of against test functions when m, i.e the number of neurons, goes to infinity. After this, in section 2 we'll study the limiting dynamic and convergence to it. Then, in section 3 we'll study the convergence of f_t (the limiting process of the neural network) to a certain type of minimum of the loss. We'll also discuss the consequences of this fact from different perspectives.

5.1. Identifying the limit

Let $\alpha \geq \frac{1}{2}$ and $\varphi \in \mathcal{C}_b^2(\mathbb{R}^p)$. As we saw in Chapter 4, if the step-size of our SGD is $\frac{\gamma}{m^\alpha}$, then by Itô's Lemma, for $m \in \mathbb{N}$, $k \in \{1, \dots, m\}$:

$$\begin{aligned}\varphi(c_k, W_t^{k,m}) = \varphi(c_k, W_0^{k,m}) &+ \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m) ds \\ &+ \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}} \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(c_k, W_s^{k,m}) dB_s^k \\ &+ \frac{\gamma}{2m^\alpha} \int_0^t \mathrm{Tr}\left(\Sigma_{k,m}^{\frac{1}{2}}(c_k, W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m}) \Sigma_{k,m}^{\frac{1}{2}}(c_k, W_s^{k,m})\right) ds.\end{aligned} \tag{5.1}$$

Remember we are not considering the case of Langevin Dynamics. Let $f_t^m(x)$ denote:

$$f_t^m(x) = \frac{1}{\sqrt{m}} \sum_{k=1}^m c_k \sigma(W_t^{k,m}, x).$$

Then, with this notation, we know that:

$$\begin{aligned} f_t^m - f_0^m = &\underbrace{\frac{1}{\sqrt{m}} \sum_{k=1}^m c_k \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m) ds}_{(A)} \\ &+ \underbrace{\frac{\gamma^{\frac{1}{2}}}{m^{\frac{1+\alpha}{2}}} \sum_{k=1}^m c_k \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^{k,m}) dB_s^k}_{(B)} \\ &+ \underbrace{\frac{\gamma}{2m^{\alpha+\frac{1}{2}}} \sum_{k=1}^m c_k \int_0^t \mathrm{Tr}\left(\Sigma_{k,m}(W_s^{k,m})^T H_w\varphi(c_k, W_s^{k,m})\right) ds}_{(C)}. \end{aligned} \tag{5.2}$$

Let's analyze each term so that we can propose a a limiting equation. We'll begin with (A). We have:

$$\begin{aligned}(A) &= \frac{1}{\sqrt{m}} \sum_{k=1}^m c_k \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T h^{k,m}(W_s^m) ds \\ &= \frac{1}{\sqrt{m}} \sum_{k=1}^m \frac{c_k^2}{\sqrt{m}} \int_0^t \nabla\varphi(c_k, W_s^{k,m})^T \mathbb{E}_{X,Y}[(Y - f_s^m(X)) \nabla\sigma(W_s^{k,m})] ds \\ &= \frac{1}{m} \sum_{k=1}^m \int_0^t \mathbb{E}_{X,Y}[(Y - f_s^m(X)) c_k^2 \nabla\varphi(c_k, W_s^{k,m})^T \nabla\sigma(W_s^{k,m}, X)] ds \\ &= \langle \int_0^t \mathbb{E}_{X,Y}[(Y - f_s^m(X)) c_k^2 \nabla\varphi(c_k, W_s^{k,m})^T \nabla\sigma(W_s^{k,m}, X)] ds, \mu_s^m \rangle,\end{aligned}$$

where μ_s^m is the empirical measure process we studied in Chapter 4. By using Fubini (which is possible, since all measures are finite and the function is integrable), we get:

$$(A) = \int_0^t \mathbb{E}_{X,Y}[(Y - f_s^m(X)) \int_{\mathbb{R}\times\mathbb{R}^p} c^2 \nabla\varphi(w)^T \nabla\sigma(w, X)\mu_s^m(dc, dw)]ds. \tag{5.3}$$

Before continuing with (B), let's study (C). Note that:

$$\begin{aligned}(C) &= \frac{\gamma}{2m^{\alpha+\frac{1}{2}}} \sum_{k=1}^m c_k \int_0^t \operatorname{Tr}\left(\Sigma_{k,m}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\right) ds \\ &= \frac{1}{m^{\alpha-\frac{1}{2}}} \frac{\gamma}{2m} \sum_{k=1}^m c_k \int_0^t \operatorname{Tr}\left(\Sigma_{k,m}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\right) ds,\end{aligned}$$

and by the last chapter, we know that $\frac{\gamma}{2m}\sum_{k=1}^m c_k \int_0^t \operatorname{Tr}\left(\Sigma_{k,m}(W_s^{k,m})^T H_w\varphi(W_s^{k,m})\right) ds$ converges as $m \to \infty$. Hence, we expect the limit of (C) to exist if $\alpha \geq \frac{1}{2}$. If not, the study seems to be quite harder, and it may even result in this term being divergent when m goes to infinity. Thus, we'll only consider the case when $\alpha \geq \frac{1}{2}$ from now on. Note that according to our analysis in Chapter 4, if $\alpha = \frac{1}{2}$,

$$(C) \overset{m\to\infty}{\longrightarrow} \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \operatorname{Tr}\left(S(\mu_s, W, c)^T H_w\varphi(W)\right) d\mu_s(dc, dW)ds, \tag{5.4}$$

and if $\alpha > \frac{1}{2}$, then

$$(C) \overset{m\to\infty}{\longrightarrow} 0. \tag{5.5}$$

At last, if we consider the case when $\alpha \geq \dfrac{1}{2}$, then by the analysis we did in Chapter 4, we expect:

$$(B) \to 0. \tag{5.6}$$

Remark It would be interesting to know what happens to this term when $\alpha \geq 0$. If it has a limit, we conjecture to be a white noise -driven martingale. We left this as a study for the future.

By combining equations (5.3), (5.4), (5.5) and (5.6), and replacing φ by σ, we expect the following dynamic for the limiting neural network:

- If $\alpha = \frac{1}{2}$:

$$\begin{aligned}f_t - f_0 = &\int_0^t \int_{\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w)^T \nabla\sigma(w, X)]\mu_s(dc, dw)ds \\ &+ \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \operatorname{Tr}\left(S(\mu, \tilde{W}, \tilde{c})^T H_w\sigma(\tilde{W})\right) d\mu_s(d\tilde{c}, d\tilde{W})ds.\end{aligned} \tag{5.7}$$

- If $\alpha > \frac{1}{2}$:

$$f_t - f_0 = \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X)) \int_{\mathbb{R}^p} c^2 \nabla\sigma(w)^T \nabla\sigma(w, X)\mu_s(dc, dw)]ds. \tag{5.8}$$

In both cases, the empirical measure's dynamic will be, because of the results in chapter 4,

for every $\varphi \in \mathcal{C}_b^2(\mathbb{R}^p)$:

$$\langle \varphi, \mu_t - \mu_0 \rangle = \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}\left[\langle c\sigma, \mu_s\rangle)\tilde{c}^2 \nabla\varphi(\tilde{W})^T \nabla\sigma(\tilde{W}, X)\right] d\mu_0(d\tilde{c}, d\tilde{W})ds = 0,$$

since c's distribution is assumed to be centered in Xavier's initialization. This way, we can re-write the dynamics in equations (5.9) and (5.10) as

- If $\alpha = \frac{1}{2}$:

$$\begin{aligned} f_t - f_0 = &\int_0^t \int_{\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w)^T \nabla\sigma(w, X)]\mu_0(dc, dw)ds \\ &+ \gamma \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \operatorname{Tr}\left(S(\mu, W, c)^T H_w \sigma(W)\right) d\mu_0(dc, dW)ds. \end{aligned} \tag{5.9}$$

- If $\alpha > \frac{1}{2}$:

$$f_t - f_0 = \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X)) \int_{\mathbb{R}^p} c^2 \nabla\sigma(w)^T \nabla\sigma(w, X)\mu_0(dc, dw)]ds. \tag{5.10}$$

We'll analyze only the case when $\alpha = \frac{1}{2}$, and the case when $\alpha > \frac{1}{2}$ will be a direct consequence. To prove convergence, well begin by defining η_0 to be a white noise in $\mathcal{L}^2(\mathbb{R}^p)$, such that it's covariance is given by the limiting measure μ_0. This means, by definition, that η_0 is a Gaussian Process indexed by $f \in \mathcal{L}^2(\mathbb{R}^p)$, and such that given f, g test functions:

$$\mathbb{E}[\langle f, \eta_0\rangle\langle g, \eta_0\rangle] = \int f(w)g(w)\mu_0(dw).$$

Now, let T_0^m define the optimal transport from μ_0 to μ_0^m, which exists because the former measure has a density. We define η_0^m by:

$$\langle \varphi, \eta_0^m\rangle := \langle \varphi \circ T_0^m, \eta_0\rangle. \tag{5.11}$$

Then, conditionally on the initialization, η_0^m is a white noise of covariance μ_0^m. We'll see this in the following

Lemma 5.1 *Let η_0 be a white noise of covariance μ_0, and let T_0^m be the optimal transport map between μ_0 and μ_0^m. Then the process defined by*

$$\langle \varphi, \eta_0^m\rangle := \langle \varphi \circ T_0^m, \eta_0\rangle$$

corresponds to a white noise of covariance μ_0^m.

Proof. By definition of T_0^m, for every $g \in L^2(\mu_0^m), goT_0^m \in L^2(\mu_0^m)$, we have that

$$\langle g, \eta_0^m\rangle_{g\in L(\mu_0^m)} = \langle goT_0^m, \eta_0\rangle_{g\in L(\mu_0^m)}$$

is a sub-Gaussian vector of $\langle g, \eta_0\rangle_{g\in L(\mu_0)}$. Hence, η_0^m is a Gaussian Process. Now, let f_0^m denote the neural network at initialization. Since η_0's mean is 0, we obtain:

$$\mathbb{E}(\langle g, \eta_0^m\rangle | f_0^m) = \mathbb{E}(\langle goT_0^m, \eta_0\rangle | f_0^m) = 0.$$

On the other hand:

$$\mathbb{E}(\langle g, \eta_0^m\rangle^2 | f_0^m) = \mathbb{E}(\langle go T_0^m, \eta_0\rangle^2 | f_0^m) = \int g^2(T_0^m(w))\mu_0(dw) = \int g^2(w)\mu_0^m(dw).$$

Then, η_0^m's covariance is μ_0^m. □

Now, for each $s \in \mathbb{R}$ we define T_s^m as the optimal transport from μ_0 and μ_s^m. Note that T_0^m exists because μ_0 has a density. Then, we can write equation (5.2) using T_s^m, obtaining:

$$\begin{aligned} f_t^m - f_0^m = &\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s^m(X))c^2 \nabla(\sigma \circ T_s^m)(w)^T \nabla(\sigma \circ T_s^m)(w, X)]\mu_0(dc, dw)ds \\ &- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(T_s^m(c), T_s^m(w), \mu_0)^T H_w(\sigma \circ T_s^m)(w)\right.\right. \mu_0(dc, dw)ds + R_t^m, \end{aligned} \tag{5.12}$$

where $f_0^m = \langle \sigma, \eta_0^m\rangle$, and we defined:

$$\begin{aligned} R_t^m = &\frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c^3 \left\{ \mathrm{Tr}\left(\mathbb{E}_{X,Y}[\frac{Y}{\sqrt{m}}\nabla\sigma(w)\nabla\sigma(w)^T]^T H_w(\sigma \circ T_s^m)(w)\right)\right. \\ &- \mathrm{Tr}\left(\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla(\sigma \circ T_s^m(w))\right]\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla(\sigma \circ T_s^m)(w)^T\right]^T H_w(\sigma \circ T_s^m)(w)\right) \\ &+ \mathrm{Tr}\left(\mathbb{E}_{X,Y}\left[(\langle(\sigma \circ T_s^m), \mu_0\rangle)\nabla(\sigma \circ T_s^m(w))\right]\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla(\sigma \circ T_s^m)(w)^T\right]^T H_w(\sigma \circ T_s^m)(w)\right) \\ &\left. + \mathrm{Tr}\left(\mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla(\sigma \circ T_s^m(w))\right]\mathbb{E}_{X,Y}\left[\langle(\sigma \circ T_s^m), \mu_0\rangle\nabla(\sigma \circ T_s^m)(w)^T\right]^T H_w(\sigma \circ T_s^m)(w)\right)\right\}\mu_0(dc, dw)ds \\ &\frac{\gamma}{m}\sum_{k=1}^m \int_0^t \Sigma^{\frac{1}{2}}(W_s^{k,m})dB_s^{k,m}. \end{aligned}$$

But, why does it make sense to say that $f_0^m = \langle \sigma, \eta_0^m\rangle$ models the problem correctly ? We can see this in the following way. Let η_0^m by defined by the following signed measure:

$$\eta_0^m := \frac{1}{\sqrt{m}}\sum_{k=1}^m c_k \delta_{W^{k,m}}.$$

Then, we can state the following

Lemma 5.2 *Let η_0^m be defined by*

$$\eta_0^m := \frac{1}{\sqrt{m}}\sum_{k=1}^m c_k \delta_{W^{k,m}}.$$

Then, conditionally on the initialization, η_0^m is a white noise with covariance μ_0.

If the lemma is true, then modeling the problem as we did in equation 5.12 makes sense.

Proof. In the first place, note that since all c_k's are Gaussian, η_0^m is a Gaussian process. On

the other hand, given $g \in L^2$, note that:

$$\mathbb{E}[\langle g, f\rangle | f_0^m] = 0,$$

because the c_k's are centered, and at last:

$$\mathbb{E}[\langle g, f\rangle^2 | f_0^m] = \frac{1}{m}\sum_{k=1}^{m} \underbrace{\mathbb{E}[c_k^2 | f_0^m]}_{=1} g(W^{k,m}) = \langle g, \mu_0^m\rangle^2,$$

which concludes that η_0^m's covariance is indeed μ_0^m. □

Knowing this, we can model the initial condition as we did in (5.12), knowing that it makes sense mathematically. We begin our path to prove convergence by showing that $\mathbb{E}[|R_t^m|]$ goes to 0 as m goes to infinity.

Lemma 5.3 *R_t^m satisfies:*

$$\lim_{m\to\infty} \mathbb{E}[|R_t^m|] = 0.$$

Proof. The expectation of the last term of R_t^m converges to 0 by 4.15. Let's study the other terms.

We begin by bounding the module of R_t^m using Cauchy-Schwartz' inequality on each term.

$$\begin{aligned}
|R_t^m| \leq \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \Bigg\{ & |c|^3 \left\| \mathbb{E}_{X,Y}[\frac{Y}{\sqrt{m}} \nabla\sigma(w)\nabla\sigma(w)^T \|_{Frob} \|H_w(\sigma\circ T_s^m)(w)\|_{Frob} \right. \\
& + |c|^3 \left\| \mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla|(\sigma\circ T_s^m(w))|\right] \mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla|(\sigma\circ T_s^m)(w)|^T\right] \right\|_{Frob} \|H_w(\sigma\circ T_s^m)(w)\|_{Frob} \\
& \left. + |c|^3 \left\| \mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\langle|(\sigma\circ T_s^m)|, \mu_0\rangle \nabla(\sigma\circ T_s^m(w))\right] \mathbb{E}_{X,Y}\left[\frac{Y}{\sqrt{m}}\nabla(\sigma\circ T_s^m)(w)^T\right] \right\|_{Frob} \|H_w(\sigma\circ T_s^m)(w)\|_{Frob}
\end{aligned}$$

Now, by using that the norm of σ's hessian is bounded, and bounding each norm of the expectation by the integral of the norm:

$$\begin{aligned}
|R_t^m| \leq \frac{C\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \Bigg\{ & |c|^3 \mathbb{E}_{X,Y}[\frac{|Y|}{\sqrt{m}} \left\|\nabla\sigma(w)\right\| \|\nabla\sigma(w)^T\|] \\
& + |c|^3 \mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}\|\nabla(\sigma\circ T_s^m(w))\|\right] \mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}\|\nabla(\sigma\circ T_s^m)(w)\|\right] \\
& + |c|^3 \mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}\langle(\sigma\circ T_s^m), \mu_0\rangle\|\nabla(\sigma\circ T_s^m(w))\|\right] \mathbb{E}\left[\frac{|Y|}{\sqrt{m}}\|\nabla(\sigma\circ T_s^m)(w)\|\right] \\
& + |c|^3 (\mathbb{E}_{X,Y}\left[\frac{|Y|}{\sqrt{m}}\|\nabla(\sigma\circ T_s^m(w))\|\right] \mathbb{E}_{X,Y}\left[\langle(\sigma\circ T_s^m), \mu_0\rangle\|\nabla(\sigma\circ T_s^m)(w)\|\right] \Bigg\} \mu_0(dc, dw)ds.
\end{aligned}$$

We can now bound the module of σ and the norm of σ's gradient and bound each expectation

of Y by using the fact that it has bounded second moment. This way:

$$\begin{aligned}|R_t^m| &\leq \frac{C\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \left\{ |c|^3 \frac{C}{\sqrt{m}} + |c|^3 \frac{C}{m} + |c|^3 \frac{C}{m} + |c|^3 \frac{C}{m} \right\} \mu_0(dc,dw)ds \\ &\leq \frac{C\gamma}{2}\left\{ \frac{1}{\sqrt{m}} + \frac{1}{m} + \frac{1}{m} + \frac{1}{m} \right\} t,\end{aligned}$$

with which we can conclude.

$$\lim_{m\to\infty} \mathbb{E}[|R_t^m|] = 0.$$

□

Now, let $f_0 = \langle \sigma, \eta_0 \rangle$, where η_0 is a white noise with covariance μ_0, and let f_t be a solution in $L^2(\mathbb{R}^p, \mathbb{P})$ of the equation with initial condition f_0:

$$\begin{aligned}f_t - f_0 &= \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - \langle \sigma, \eta_s \rangle) c^2 \nabla\sigma(w)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds \\ &- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c^3 \Big(\mathrm{Tr}\left(\mathbb{E}_X[(\langle \sigma, \mu_0 \rangle) \nabla\sigma(w)\sigma(w)^T]^T H_w\sigma(w) \right. \\ &- \left. \mathbb{E}_X\left[\langle \sigma, \mu_0 \rangle \nabla\sigma\right] \mathbb{E}_X\left[\langle \sigma, \mu_0 \rangle \nabla\sigma^T \right]^T H_w\sigma(w) \right) \Big) \mu_0(dc,dw)ds.\end{aligned}$$

We'd like to prove that in some sense (hopefully in $L^2(\mathbb{R}^p, \mathbb{P})$) f_t^m converges to f_t. We'll see that this is true, for which the transport T_s^m we previously defined will prove to be very useful. This technique was first used in [24] and [25], where the measurably of T_s^m with respect to the filtration was also proved. We believe this technique can be used in more general settings, as the ones where a white noise may appear in the limiting equations.

Theorem 5.1 *Let η_0 be a white noise with covariance μ_0, and let f_t be a solution of the equation in $L^2(\mathcal{X})$:*

$$\begin{aligned}f_t(x) - f_0(x) &= \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X)) c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds \\ &- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w\sigma(w) \right) \right) \mu_0(dc,dw)ds., \qquad (5.13)\end{aligned}$$

with $f_0 = \langle \sigma, \eta_0 \rangle$. Then, for every $t \geq 0$:

$$\lim_{m\to\infty} \|f_t^m - f_t\| = 0.$$

Proof. Let's remember both dynamics, for f_t^m and f_t. We define f_t^m as the solution of the

problem:

$$f_t^m(x) = f_0^m(x) + \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s^m(X))c^2 \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw)ds$$
$$- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c, T_s^m(w), \mu_s^m)^T H_w(\sigma \circ T_s^m)(w,x)\right)\right)\mu_0(dc,dw)ds + R_t^m,$$

with initial condition $f_0^m = \langle \sigma, \eta_0^m \rangle$, where η_0^m is white noise in $L^2(\mathcal{X})$ with covariance μ_0^m. On the other hand, f_t is the solution of :

$$f_t(x) = f_0(x) + \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds$$
$$- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w\sigma(w)\right)\right)\mu_0(dc,dw)ds,$$

with initial condition $f_0 = \langle \sigma, \eta_0 \rangle$, where η_0 is a White Noise in $L^2(\mathcal{X})$ with covariance μ_0. Then, by subtracting the equations for f_t and f_t^m:

$$\begin{aligned} f_t(x) - f_t^m(x) &= f_0(x) - f_0^m(x) \\ &+ \Big(\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s^m(X))c \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw)ds \Big) \\ &- \Big(\frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w\sigma(w)\right)\right)\mu_0(dc,dw)ds \\ &- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c, T_s^m(w), \mu_s^m)^T H_w(\sigma \circ T_s^m)(w,x)\right)\right)\mu_0(dc,dw)ds \Big). \end{aligned}$$

By adding and subtracting $\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw)ds$ inside the second term:

$$\begin{aligned} f_t(x) - f_t^m(x) &= f_0(x) - f_0^m(x) \\ &+ \Big(\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw)ds \Big) \\ &+ \Big(\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw)ds \\ &- \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s^m(X))c^2 \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw)ds \Big) \\ &- \Big(\frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w\sigma(w)\right)\right)\mu_0(dc,dw)ds \\ &- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c, T_s^m(w), \mu_s^m)^T H_w(\sigma \circ T_s^m)(w,x)\right)\right)\mu_0(dc,dw)ds \Big). \end{aligned}$$

and now by adding and subtracting $\frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w(\sigma \circ T_s^m)(w,x)\right)\right)\mu_0(dc,dw)ds$

in the last term:

$$
\begin{aligned}
f_t(x) - f_t^m(x) &= (f_0(x) - f_0^m(x)) \\
&\quad + \Big(\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_0(dc,dw)ds \\
&\quad - \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla(\sigma\circ T_s^m)(w,x)^T \nabla(\sigma\circ T_s^m)(w,X)]\mu_0(dc,dw)ds \Big) \\
&\quad + \Big(\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla(\sigma\circ T_s^m)(w,x)^T \nabla(\sigma\circ T_s^m)(w,X)]\mu_0(dc,dw)ds \\
&\quad - \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s^m(X))c^2 \nabla(\sigma\circ T_s^m)(w,x)^T \nabla(\sigma\circ T_s^m)(w,X)]\mu_0(dc,dw)ds \Big) \\
&\quad - \Big(\frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w \sigma(w) \right) \right) \mu_0(dc,dw)ds \\
&\quad - \frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w (\sigma\circ T_s^m)(w,x) \right) \right) \mu_0(dc,dw)ds \Big) \\
&\quad + \Big(\frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w (\sigma\circ T_s^m)(w,x) \right) \right) \mu_0(dc,dw)ds \\
&\quad - \frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left(S(c,T_s^m(w),\mu_s^m)^T H_w (\sigma\circ T_s^m)(w,x) \right) \right) \mu_0(dc,dw)ds \Big).
\end{aligned}
$$

By manipulating the different terms we obtain:

$$
\begin{aligned}
f_t(x) - f_t^m(x) &= \underbrace{f_0(x) - f_0^m(x)}_{(1)} \\
&+ \underbrace{\left(\int_0^t \langle \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 (\nabla\sigma(w,x)^T \nabla\sigma(w,X) - \nabla(\sigma\circ T_s^m)(w,x)^T \nabla(\sigma\circ T_s^m)(w,X))], \mu_0 \rangle ds \right)}_{(2)} \\
&+ \underbrace{\left(\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(f_s^m(X) - f_s(X))c^2 \nabla(\sigma\circ T_s^m)(w,x)^T \nabla(\sigma\circ T_s^m)(w,X)]\mu_0(dc,dw)ds \right)}_{(3)} \\
&- \underbrace{\left(\frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left(S(c,w,\mu_0)^T (H_w\sigma(w) - H_w(\sigma\circ T_s^m)(w,x)) \right) \right) \mu_0(dc,dw)ds \right)}_{(4)} \\
&+ \underbrace{\left(\frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left((S(c,w,\mu_0)^T - S(c,T_s^m(w),\mu_s^m))^T H_w (\sigma\circ T_s^m)(w,x) \right) \right) \mu_0(dc,dw)ds \right)}_{(5)}.
\end{aligned}
$$

Let's begin with (1). We have, conditionally on the initialization, that:

$$
\begin{aligned}
\mathbb{E}[\int_X (f_0(X) - f_0^m(X))^2] &= \mathbb{E}[(\langle \sigma(X) - \sigma\circ T_0^m(X), \eta_0 \rangle^2] \\
&= \mathbb{E}\left[\int_X \int (c\sigma(w,X) - c\sigma(T_0^m(w),X)^2 \mu_0(dc,dw) \right] \\
&\leq C\mathbb{E}\left[\int (c\|w - T_0^m(w)\|^2 \mu_0(dc,dw) \right] \\
&\leq C\mathcal{W}_2^2(\mu_0^m, \mu_0),
\end{aligned}
$$

therefore we can conclude:

$$\mathbb{E}[\|f_0 - f_0^m\|_{L_2(X)}] \leq C\mathcal{W}_2(\mu_0^m, \mu_0).$$

For (2), we note that:

$$\begin{aligned}
|(2)| &= \Big| \int_0^t \langle \mathbb{E}_{X,Y}[(Y - f_s(X))c(\nabla\sigma(w,x)^T \nabla\sigma(w,X) \\
&\quad -\nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X))], \mu_0\rangle ds \Big| \\
&= \Big| \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X))\langle c(\nabla\sigma(w,x)^T \nabla\sigma(w,X) \\
&\quad -\nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)), \mu_0\rangle] ds \Big| \\
&\overset{C-S}{\leq} \int_0^t \|(Y - f_s(X)\|_2 \underbrace{\|\langle c(\nabla\sigma(w,x)^T \nabla\sigma(w,\tilde{X}) - \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,\tilde{X})), \mu_0\rangle\|_{L^2(\tilde{X})}}_{(\star)} ds
\end{aligned}$$

Let's focus on $(\star)$ for a moment. We have:

$$\begin{aligned}
(\star) &= \|\langle c(\nabla\sigma(w,x)^T \nabla\sigma(w,\tilde{X}) - \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,\tilde{X})), \mu_0\rangle\|_{L^2(\tilde{X})} \\
&\leq \|\langle c(\nabla\sigma(w,x)^T \nabla\sigma(w,\tilde{X}) - \nabla\sigma(w,x)^T \nabla(\sigma \circ T_s^m)(w,\tilde{X})), \mu_0\rangle\| \\
&+ \|\langle c(\nabla\sigma(w,x)^T \nabla(\sigma \circ T_s^m)(w,\tilde{X})) - \nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,\tilde{X})), \mu_0\rangle\| \\
&\leq C\|w - T_s^m(w)\|_2 \\
&\leq C\mathcal{W}(\mu_0, \mu_s^m).
\end{aligned}$$

By replacing this in (2):

$$\begin{aligned}
|(2)| &\leq C \int_0^t \mathbb{E}_{X,Y}[|Y - f_s(X)|\mathcal{W}(\mu_0, \mu_s^m) ds \\
&\overset{C-S}{\leq} C \int_0^t \|Y - f_s(X)\|_{L_2(X)} \mathcal{W}(\mu_0, \mu_s^m) ds.
\end{aligned}$$

For (3), by directly using Cauchy-Schwartz inequality and the fact that $\nabla\sigma$'s norm is bounded, we obtain:

$$\begin{aligned}
|(3)| &= \Big| \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(f_s^m(X) - f_s(X))c\nabla(\sigma \circ T_s^m)(w,x)^T \nabla(\sigma \circ T_s^m)(w,X)]\mu_0(dc,dw) ds \Big| \\
&\leq C \int_0^t \|f_s^m - f_s\|_{L_2(X)} ds.
\end{aligned}$$

By noting that $S(c,w,\mu)$ is bounded, we do the same we did in (2) at (4) and obtain:

$$|(4)| \leq \frac{C\gamma}{2} \int_0^t \mathcal{W}(\mu_s^m, \mu_0) ds.$$

At last, by the same type of manipulations, we obtain:

$$|(5)| \leq \frac{C\gamma}{2} \int_0^t \mathcal{W}(\mu_s^m, \mu_0)^2 ds.$$

By putting all our bounds together, we can conclude:

$$\|f_s - f_s^m\|_{L_2(X)} \leq C\mathcal{W}_2^2(\mu_0^m, \mu_0) + C\int_0^t \mathcal{W}(\mu_s^m, \mu_0) + C\int_0^t \|f_s^m - f_s\|_{L_2(X)}ds \\ + \frac{C\gamma}{2}\int_0^t \mathcal{W}(\mu_s^m, \mu_0)ds + \frac{C\gamma}{2}\int_0^t \mathcal{W}(\mu_s^m, \mu_0)^2 ds. \tag{5.14}$$

We define:

$$L_m(t) = \mathcal{W}_2^2(\mu_0^m, \mu_0) + \int_0^t \mathcal{W}(\mu_s^m, \mu_0) + +\frac{\gamma}{2}\int_0^t \mathcal{W}(\mu_s^m, \mu_0)ds + \frac{\gamma}{2}\int_0^t \mathcal{W}(\mu_s^m, \mu_0)^2 ds,$$

therefore:

$$\|f_t - f_t^m\|_{L_2(X)} \leq CL_m(t) + C\int_0^t \|f_s^m - f_s\|_{L_2(X)}ds.$$

And by applying Gronwall's inequality:

$$\|f_t - f_t^m\|_{L_2(X)} \leq CL_m(t)e^t, \tag{5.15}$$

and by making m to infinity, since $L_m(t) \to 0$ as $m \to \infty$, we can conclude:

$$\lim_{m\to\infty} \|f_t - f_t^m\|_{L_2(X)} = 0.$$

□

As a conclusion, when $\alpha = \frac{1}{2}$ the limiting dynamic of f will be:

$$f_t(x) - f_0(x) = \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2\nabla\sigma(w,x)^T\nabla\sigma(w,X)]\mu_0(dc,dw)ds \\ - \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_0)^T H_w\sigma(w)\right)\right)\mu_0(dc,dw)ds,$$

and when $\alpha > \frac{1}{2}$:

$$f_t(x) - f_0(x) = \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2\nabla\sigma(w,x)^T\nabla\sigma(w,X)]\mu_0(dc,dw)ds,$$

where the initial condition for both equations is $f_0 = \langle\sigma, \eta_0\rangle$.

5.2. Studying the Limiting Dynamic with Xavier initialization

We devote this section to the study of the limiting equations we found in the last section. As in the last section, we focus our study in the case when when the initialization of our parameters is centered. By Theorem 5.1, the limit ODE that f_t satisfies is:

$$f_t(x) = f_0(x) + \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c\nabla\sigma(w,x)^T\nabla\sigma(w,X)]\mu_0(dc,dw)ds.$$

In this section, we attempt to answer the question: Does f_t converges to a minima of some kind when t grows? As we'll see, the answer will be yes. Nevertheless, it won't be in the

sense we'd like it to be: The convergence will not be for our loss

$$L[f] = \mathbb{E}_{X,Y}[(Y - f_s(X))^2],$$

but for a modified version of it, in a different space than the one we'd expect (which would be $L^2(\mathcal{X})$). For this mission, we'll study the results presented in [19], where the authors study the limiting object when the number of neurons go to infinity by means of the Neural Tangent Kernel, and we'll revisit them through a different perspective.
Note that, since all measures are finite, we can exchange integrals and rewrite f_t's dynamic, obtaining:

$$f_t(x) = f_0(x) + \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X))\langle c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X), \mu_0\rangle]ds. \tag{5.16}$$

By remembering the definition of the continuous Neural Tangent Kernel we introduced in equation (4.106) of section 3.6,

$$K(x,x') := \langle c^2 \nabla\sigma(w,x)\nabla\sigma(w,x'), \mu_0\rangle,$$

we obtain that:

$$f_t(x) = f_0(x) + \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X))K(X,\cdot)]ds. \tag{5.17}$$

Consider the RKHS definition we gave in the background section. Then, it's not hard to see from the dynamic of f_t, that it follows Kernel Gradient Descent on the Reproducing Kernel Hilbert Space, with the Kernel equal to the Neural Tangent Kernel. That is, f_t follows gradient descent on the RKHS associated to the NTK with respect to the loss:

$$L_K : \mathcal{H}_K \to \mathbb{R}$$

$$L_K[f] = \|Y - f(X))\|^2_{\mathcal{H}_K}.$$

We'd like to prove that this loss converges to some kind of global minima. For this, we'll rely on the strong convexity of L_K. We stress the fact that L_K **does not corresponds to the original loss metric, but to a it's modification in the RKHS defined by the Neural Tangent Kernel**. Before proving convergence, let's give a definition of m-convexity:

Definition 5.1 *An operator $O : \mathcal{H}_K \to \mathbb{R}$ is m-convex if the operator*

$$f \to O[f] - \frac{m}{2}\|f\|^2_{\mathcal{H}_K}$$

is convex. If $m > 0$, we say O is strongly convex.

Lemma 5.4 *If K is a positive definite kernel, the operator $L_K[f] = \|Y - f(X))\|^2_{\mathcal{H}_K}$ is 2-convex in the Hilbert Space $\mathcal{H}_K$.*

Proof. Let $f_1, f_2 \in \mathcal{H}_K$, and $\lambda \in [0,1]$. We denote $\bar{\lambda} = 1 - \lambda$. If the Kernel is positive

definite, it defines an RKHS and hence, we have:

$$\begin{aligned}
L[\lambda f_1 + \tilde{\lambda} f_2] &= \langle Y - \lambda f_1 + \tilde{\lambda} f_2, Y - \lambda f_1 + \tilde{\lambda} f_2 \rangle_{\mathcal{H}_K} \\
&= \langle \lambda(Y - \lambda f_1) + \tilde{\lambda}(Y - f_2), \lambda(Y - \lambda f_1) + \tilde{\lambda}(Y - f_2) \rangle_{\mathcal{H}_K} \\
&= \lambda^2 \|Y - \lambda f_1\| + \tilde{\lambda}^2 \|Y - \lambda f_2\| + 2\lambda\tilde{\lambda}\langle Y - f_1, Y - f_2 \rangle_{\mathcal{H}_K} \\
&= \lambda^2 L[f_1] + \tilde{\lambda}^2 L[f_2] + 2\lambda\tilde{\lambda}\langle Y - f_1, Y - f_2 \rangle_{\mathcal{H}_K} \\
&= \lambda(1 - \tilde{\lambda}) L[f_1] + \tilde{\lambda}(1 - \lambda) L[f_2] + 2\lambda\tilde{\lambda}\langle Y - f_1, Y - f_2 \rangle_{\mathcal{H}_K} \\
&= \lambda L[f_1] - \lambda\bar{\lambda} L[f_1] + \bar{\lambda} L[f_2] - \lambda\bar{\lambda} L[f_2] + 2\lambda\bar{\lambda}\langle Y - f_1, Y - f_2 \rangle_{\mathcal{H}_K} \\
&= \lambda L[f_1] + \bar{\lambda} L[f_2] - \lambda\bar{\lambda}(L[f_1] + L[f_2] - 2\langle Y - f_1, Y - f_2 \rangle_{\mathcal{H}_K}) \\
&= \lambda L[f_1] + \bar{\lambda} L[f_2] - \lambda\bar{\lambda}\|f_1 - f_2\|^2_{\mathcal{H}_K}.
\end{aligned}$$

By noting that:

$$\begin{aligned}
\|f_1 - f_2\|^2_{\mathcal{H}_K} &= \lambda^2 \|f_1\|^2_{\mathcal{H}_K} + 2\lambda\bar{\lambda}\langle f_1, f_2 \rangle_{\mathcal{H}_K} + \bar{\lambda}^2 \|f_2\|^2_{\mathcal{H}_K} \\
&= \lambda \|f_1\|^2_{\mathcal{H}_K} - \lambda\bar{\lambda}\|f_1\|^2_{\mathcal{H}_K} + 2\lambda\bar{\lambda}\langle f_1, f_2 \rangle_{\mathcal{H}_K} + \bar{\lambda}\|f_2\|^2_{\mathcal{H}_K} - \lambda\bar{\lambda}\|f_2\|^2_{\mathcal{H}_K},
\end{aligned}$$

and replacing in L's expression, we obtain:

$$\begin{aligned}
L[\lambda f_1 + \tilde{\lambda} f_2] - \|f_1 - f_2\|^2_{\mathcal{H}_K} &= \lambda L[f_1] + \bar{\lambda} L[f_2] - \lambda\bar{\lambda}\|f_1 - f_2\|^2_{\mathcal{H}_K} \\
&\quad - \lambda\|f_1\|^2_{\mathcal{H}_K} + \lambda\bar{\lambda}\|f_1\|^2_{\mathcal{H}_K} - 2\lambda\bar{\lambda}\langle f_1, f_2 \rangle_{\mathcal{H}_K} - \bar{\lambda}\|f_2\|^2_{\mathcal{H}_K} + \lambda\bar{\lambda}\|f_2\|^2_{\mathcal{H}_K} \\
&= \lambda L[f_1] + \bar{\lambda} L[f_2] - \lambda\|f_1\|^2_{\mathcal{H}_K} - \bar{\lambda}\|f_2\|^2_{\mathcal{H}_K},
\end{aligned}$$

with which we conclude that L is 2-convex. □

In order to prove convergence, we use the following Lemma, which is classic for the prove of convergence of gradient descent methods. For a deeper study of this kind of inequalities, called Lojasiewicz-Simon inequalities, we recommend [37].

Lemma 5.5 (Lojasiewicz-Simon Inequality) *Let $L : \mathcal{H}_K \to \mathbb{R}$ be an k-convex functional, and f_t be such that:*

$$\frac{d}{dt} f_t = -\nabla L[f_t].$$

Then:

$$L[f_t] - L^* \leq \frac{1}{2k} \|\nabla L[f_t]\|^2_{\mathcal{H}_K},$$

where L^* is a global minima of L.

We'll use this Lemma in order to create a Gronwall Inequality for our Loss functional.

Theorem 5.2 *Let L be the loss we defined above, and consider f_t such that*

$$f_t(x) = f_0(x) + \int_0^t \mathbb{E}_{X,Y}[(Y - f_s(X))K(X, \cdot)]ds.$$

Then, if K is a positive definite kernel,

$$\lim_{t \to \infty} L[f_t] = L^*.$$

Proof. As we noted before, we have that:

$$\frac{d}{dt}f_t = -\nabla L[f_t].$$

Then, by Lemma 5.5:

$$L[f_t] - L^* \leq \frac{1}{4}\|\nabla L[f_t]\|^2_{\mathcal{H}_K},$$

and by multiplying by -1:

$$-\|\nabla L[f_t]\|^2_{\mathcal{H}_K} \leq -(L[f_t] - L^*). \tag{5.18}$$

Now, note that:

$$\frac{d}{dt}L[f_t] = \langle \nabla L[f_t], \frac{d}{dt}f_t\rangle_{\mathcal{H}_K} = -\langle \nabla L[f_t], \nabla L[f_t]\rangle_{\mathcal{H}_K} = -\|\nabla L[f_t]\|^2_{\mathcal{H}_K}. \tag{5.19}$$

By using equation 5.18 in 5.19:

$$\frac{d}{dt}L[f_t] \leq -(L[f_t] - L^*),$$

which we can rewrite as:

$$\frac{d}{dt}(L[f_t] - L^*) \leq -(L[f_t] - L^*).$$

Now, by applying Gronwall's inequality:

$$(L[f_t] - L^*) \leq (L[f_0] - L^*)e^{-t}.$$

We conclude by taking the limit $t \to \infty$. □

With this last Theorem, we prove that as time passes, the mean field version of our shallow neural network converges as to a minimizer of L_K as t grows.

Remark The fact that the functional being minimized is L_K, the RKHS version of L can be seen in the following way in practice: Since in practice the NTK is actually given by the train set, this suggests that the neural network's limit will over-fit in the training set as training occurs.

A natural question would be, what happens in the more general case of non-centered initialization for c, where the limiting dynamic is:

$$\begin{aligned} f_t(x) - f_0(x) = &\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} \mathbb{E}_{X,Y}[(Y - f_s(X))c^2 \nabla\sigma(w,x)^T \nabla\sigma(w,X)]\mu_s(dc,dw)ds \\ &- \frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_s)^T H_w \sigma(w)\right)\right)\mu_s(dc,dw)ds. \end{aligned}$$

As we already saw, the first term is minimizing our Loss, but what is the non-linear term doing in this dynamic? The answer lies in part in [17]. The non-linear term is given by:

$$\frac{\gamma}{2}\int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c\left(\mathrm{Tr}\left(S(c,w,\mu_s)^T H_w \sigma(w)\right)\right)\mu_s(dc,dw)ds$$

Consider the case when $S(c, w, \mu_s)^T = \mathcal{I}$. We'd obtain

$$\frac{\gamma}{2} \int_0^t \int_{\mathbb{R}\times\mathbb{R}^p} c \left(\mathrm{Tr}\left(H_w \sigma(w)\right)\right) \mu_s(dc, dw) ds.$$

This corresponds to an entropic regularization of μ. From this, we can conclude that in the mean field, the neural network minimizes a regularized version of the loss, which could be one of the reasons why SGD generalizes better than gradient descent.

Chapter 6

Conclusion

Different parametrizations of neural networks have been studied in the literature, being the most important ones among them the mean field parametrization and the Neural Tangent Kernel (NTK) parametrization. Even though both lines of research have witnessed lots of advances in recent years, some questions still remain open. The results in this work studied the NTK regime by using the methods proposed in [17]. This techniques proved to be a very good toolkit for the study of the limiting dynamics.

In the first place, this work proved that the NTK is a direct consequence of a centered initialization of the parameters. This suggest that for a complete study of the good generalization properties of neural networks, we should be considering the behavior of more general initialization distributions. In the second place, the results in this work provide a novel framework for studying the neural network process, that is, the one that considers signed measures for the parameters of the neural network. In the third place, this work also studies the differences between different training schemes. We found that, as in [17], a regularizing term appears when the network is trained by Stochastic Gradient Descent, yet not in the same way, since it appears in the neural network's dynamic.

Even though some training regimes could be completely studied, in particular the ones that arise when $\alpha \geq \frac{1}{2}$, it's also relevant to mention that it seems difficult to study the case when $\alpha \geq \frac{1}{2}$ with the same tools. This is because the non-linear terms appearing in the limiting dynamics seem to go to infinity when the limit of infinite neurons is taken. We conjecture that this study can be made by the use of other tools, and that in the particular case of centered initializations the dynamics will be the same as the ones in this work. This presents an interesting possible line of future work.

It's also important to remark that while finishing this work, the authors found the work in [11], which does a similar study. Nevertheless, the techniques they use are based on a discrete setting, while ours is a continuous setting. On the other hand, our assumptions on the parameters initialization are more general in Section 3.

An interesting future line of research is extending the work done in this thesis to multilayer neural networks, with different training schemes. This work is partially done in the literature, yet a satisfying theory that completely explains the success of deep neural networks and more generally of deep learning is still not in sight.

Bibliography

[1] Mcculloch, W. y Pitts, W., "A logical calculus of ideas immanent in nervous activity," Bulletin of Mathematical Biophysics, vol. 5, pp. 127–147, 1943.

[2] Goodfellow, I., Bengio, Y., y Courville, A., Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

[3] Zhang, C., Bengio, S., Hardt, M., Recht, B., y Vinyals, O., "Understanding deep learning requires rethinking generalization," 2016, doi:10.48550/ARXIV.1611.03530.

[4] Hornik, K., "Approximation capabilities of multilayer feedforward networks," Neural networks, vol. 4, no. 2, pp. 251–257, 1991.

[5] Wright, J. y Ma, Y., High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications. Cambridge University Press, 2022.

[6] Shalev-Shwartz, S. y Ben-David, S., Understanding Machine Learning - From Theory to Algorithms. Cambridge University Press, 2014.

[7] Bottou, L., "Stochastic gradient learning in neural networks," en Proceedings of Neuro-Nîmes 91, (Nimes, France), EC2, 1991, http://leon.bottou.org/papers/bottou-91c.

[8] Ghosal, S. y van der Vaart, A., Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2017, doi:10.1017/9781139029834.

[9] Wainwright, M. J., High-dimensional statistics: A non-asymptotic viewpoint, vol. 48. Cambridge university press, 2019.

[10] Mei, S., Misiakiewicz, T., y Montanari, A., "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit," 2019, doi:10.48550/ARXIV.1902.06015.

[11] Sirignano, J. y Spiliopoulos, K., "Scaling limit of neural networks with the xavier initialization and convergence to a global minimum," 2019, doi:10.48550/ARXIV.1907.04108.

[12] Chizat, L. y Bach, F., "On the global convergence of gradient descent for over-parameterized models using optimal transport," 2018, doi:10.48550/ARXIV.1805.09545.

[13] Rotskoff, G. M. y Vanden-Eijnden, E., "Trainability and accuracy of neural networks: An interacting particle system approach," 2018, doi:10.48550/ARXIV.1805.00915.

[14] Lasry, J.-M. y Lions, P.-L., "Mean field games," Japanese journal of mathematics, vol. 2, no. 1, pp. 229–260, 2007.

[15] Guerra, F. y Toninelli, F. L., "The thermodynamic limit in mean field spin glass models," Communications in Mathematical Physics, vol. 230, pp. 71–79, 2002.

[16] Sznitman, A.-S., Topics in propagation of chaos. Springer, 1991.

[17] De Bortoli, V., Durmus, A., Fontaine, X., y Simsekli, U., "Quantitative propagation of chaos for sgd in wide neural networks," 2020, doi:10.48550/ARXIV.2007.06352.

[18] McKean Jr, H. P., "A class of markov processes associated with nonlinear parabolic equations," Proceedings of the National Academy of Sciences, vol. 56, no. 6, pp. 1907–1911, 1966.

[19] Jacot, A., Gabriel, F., y Hongler, C., "Neural tangent kernel: Convergence and generalization in neural networks," CoRR, vol. abs/1806.07572, 2018, http://arxiv.org/abs/1806.07572.

[20] Chen, Z., Cao, Y., Gu, Q., y Zhang, T., "A generalized neural tangent kernel analysis for two-layer neural networks," 2020.

[21] Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., y Wang, R., "On exact computation with an infinitely wide neural net," Advances in neural information processing systems, vol. 32, 2019.

[22] Chizat, L., Oyallon, E., y Bach, F., "On lazy training in differentiable programming," 2018, doi:10.48550/ARXIV.1812.07956.

[23] Fontaine, X., De Bortoli, V., y Durmus, A., "Convergence rates and approximation results for sgd and its continuous-time counterpart," 2020, doi:10.48550/ARXIV.2004.04193.

[24] Fontbona, J., Guérin, H., y Méléard, S., "Measurability of optimal transportation and convergence rate for landau type interacting particle systems," Probability theory and related fields, vol. 143, no. 3, pp. 329–351, 2009.

[25] Fontbona, J., Guérin, H., y Méléard, S., "Measurability of optimal transportation and strong coupling of martingale measures," Electronic communications in probability, vol. 15, pp. 124–133, 2010.

[26] Karatzas, I. y Shreve, S. E., Brownian motion and stochastic calculus. Springer, 2012.

[27] Stroock, D. y Varadhan, S., Multidimensional Diffusion Processes. Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 1997, https://books.google.de/books?id=DuDsmoyqCy4C.

[28] Billingsley, P., Convergence of probability measures. Wiley Series in Probability and Statistics: Probability and Statistics, New York: John Wiley & Sons Inc., second ed., 1999. A Wiley-Interscience Publication.

[29] Joffe, A. y Metivier, M., "Weak convergence of sequences of semimartingales with applications to multitype branching processes," Advances in Applied Probability, vol. 18, no. 1, p. 20–65, 1986, doi:10.2307/1427238.

[30] Chaintron, L.-P. y Diez, A., "Propagation of chaos: A review of models, methods and applications. . applications," Kinetic and Related Models, vol. 15, no. 6, p. 1017, 2022, doi:10.3934/krm.2022018.

[31] Gärtner, J., "On the mckean-vlasov limit for interacting diffusions," Mathematische Nachrichten, vol. 137, no. 1, pp. 197–248, 1988.

[32] Peyré, G. y Cuturi, M., "Computational optimal transport," 2018, doi:10.48550/ARXIV.1803.00567.

[33] Méléard, S., Asymptotic behaviour of some interacting particle systems; McKean-Vlasov

and Boltzmann models. Springer, 1996.

[34] Chaintron, L.-P. y Diez, A., "Propagation of chaos: a review of models, methods and applications. ii. applications," 2021, doi:10.48550/ARXIV.2106.14812.

[35] Sirignano, J. A. y Spiliopoulos, K., "Asymptotics of reinforcement learning with neural networks," CoRR, vol. abs/1911.07304, 2019, http://arxiv.org/abs/1911.07304.

[36] Glorot, X. y Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks," en Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[37] Kohout, J., "Notes on the lojasiewicz - simon inequality,", https://people.maths.ox.ac.uk/kohout/LojasiewiczNotes.pdf.

1

Annex

1.1. Controlling the moments of the parameters

In this appendix, we'll prove that the 4th moment of the vector of all the weights in the neural network's hidden layer, W_s^m is finite, i.e $\mathbb{E}[\|W_s^m\|^4]] < \infty$.

Lemma 1.1 *Let $\alpha > 0$, $\lambda \in [0, 1)$, $\gamma \geq 0$, and μ_t^m denote the empirical measure process that represents the weights of a shallow neural network, who's parameters are trained in continuous time by the dynamics:*

$$dW_t^{k,m} = h^{k,m}(W_t^m)dt + \frac{\gamma^{\frac{1}{2}}}{m^{\frac{\alpha}{2}}}\Sigma_{k,m}^{\frac{1}{2}}(W_t^{k,m})dB_t^{k,m} + \sqrt{2\tau}d\tilde{B}_s^{k,m}, \tag{1.1}$$

where $W_t^{k,m}$ denotes one neuron in the hidden layer. Let μ_0 denote the initialization distribution for the pair (C, W). Then, we have, for all $s \geq 0$

$$\mathbb{E}[\|W_s^{k,m}\|^2] < \infty \text{ and } \mathbb{E}[\|W_s^{k,m}\|^4] < \infty.$$

Proof. We already proved, in section 3, that the coefficients in equation (1.1) are Lipschitz for W^m. Then, by Ito's Lemma:

$$\begin{aligned}\|W_t^{k,m}\|^2 = \|W_0^{k,m}\|^2 &+ \int_0^t (W_s^{k,m})^T h^{k,m}(W_s^m)ds \\ &+ \int_0^t (W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^m)dB_s^{k,m} + \frac{\gamma}{2}\int_0^t \mathrm{Tr}(\Sigma_{k,m}(W_s^m))ds.\end{aligned} \tag{1.2}$$

Then, by applying expectation :

$$\begin{aligned}\mathbb{E}[\|W_t^{k,m}\|^2] &= \mathbb{E}[\|W_0^{k,m}\|^2] + \mathbb{E}\left[\int_0^t (W_s^{k,m})^T h^{k,m}(W_s^m)ds\right] + \frac{\gamma}{2}\mathbb{E}\left[\int_0^t \mathrm{Tr}(\Sigma_{k,m}(W_s^m))ds\right] \\ &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + \mathbb{E}\left[\int_0^t (W_s^{k,m})^T h^{k,m}(W_s^m)ds\right] + \frac{\gamma}{2}\mathbb{E}\left[\int_0^t \mathrm{Tr}(\Sigma_{k,m}(W_s^m))ds\right] \\ &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + \int_0^t \mathbb{E}\left[\|W_s^{k,m}\|\|h^{k,m}(W_s^m)\|\right]ds + \frac{\gamma}{2}\int_0^t \mathbb{E}\left[|\,\mathrm{Tr}(\Sigma_{k,m}(W_s^m))|\right]ds \\ &\leq \mathbb{E}[\|W_0^{k,m}\|^2] + \int_0^t \mathbb{E}\left[\|W_s^{k,m}\|\|h^{k,m}(W_s^m)\|\right]ds + \frac{\gamma}{2}\int_0^t \mathbb{E}\left[\|\Sigma_{k,m}(W_s^m))\|\right]ds.\end{aligned}$$

Now, by using that the coefficients are Lipschitz:

$$\mathbb{E}[\|W_t^{k,m}\|^2] \leq \mathbb{E}[\|W_0^{k,m}\|^2] + C\int_0^t \mathbb{E}\left[\|W_s^{k,m}\|(\|W_s^m - W_0^m\| + \|W_0^m\|)\right] ds \\ + \frac{\gamma}{2}\int_0^t \mathbb{E}\left[(\|W_s^m - W_0^m\| + \|W_0^m\|)^2\right] ds.$$

Now, since $\|W_s^{k,m}\| \leq \|W_s^m\|$:

$$\mathbb{E}[\|W_t^{k,m}\|^2] \leq \mathbb{E}[\|W_0^m\|^2] + C\int_0^t \mathbb{E}\left[(\|W_s^m - W_0^m\| + \|W_0^m\|)^2\right] ds \\ + \frac{\gamma}{2}\int_0^t \mathbb{E}\left[(\|W_s^m - W_0^m\| + \|W_0^m\|)^2\right] ds.$$

Next, we use that $(a+b)^2 \leq 2a^2 + 2b^2$ and the triangular inequality and obtain:

$$\mathbb{E}[\|W_t^{k,m}\|^2] \leq C\mathbb{E}[\|W_0^m\|^2] + C\int_0^t \mathbb{E}\left[\|W_s^m\|^2\right] ds.$$

Since this applies for every $k \in \{1, \dots, m\}$, we have:

$$\mathbb{E}[\|W_t^m\|^2] \leq C\mathbb{E}[\|W_0^m\|^2] + C\int_0^t \mathbb{E}\left[\|W_s^m\|^2\right] ds.$$

By using Gronwall in this inequality, we conclude:

$$\mathbb{E}[\|W_t^{k,m}\|^2] \leq C\mathbb{E}[\|W_0^m\|^2] < \infty,$$

where the last norm is finite because we assumed the initial distribution to have finite second moments. Now, let's do the same for the 4th moment. For that, we'll use (1.3), which we recall is given by

$$\|W_t^{k,m}\|^2 = \|W_0^{k,m}\|^2 + \int_0^t (W_s^{k,m})^T h^{k,m}(W_s^m) ds \\ + \int_0^t (W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^m) dB_s^{k,m} + \frac{\gamma}{2}\int_0^t \mathrm{Tr}(\Sigma_{k,m}(W_s^m)) ds. \tag{1.3}$$

By applying Itô's Lemma to the function $(1 + \|W_s^{k,m}\|^2)^p$, we obtain:

$$(1 + \|W_s^{k,m}\|^2)^p = (1 + \|W_0^{k,m}\|^2)^p + \int_0^t p(1 + \|W_s^{k,m}\|^2)^{p-1}(W_s^{k,m})^T h^{k,m}(W_s^m) ds \\ + \int_0^t p(1 + \|W_s^{k,m}\|^2)^{p-1}(W_s^{k,m})^T \Sigma_{k,m}^{\frac{1}{2}}(W_s^m) dB_s^{k,m} + \\ \frac{\gamma}{2}\int_0^t \mathrm{Tr}\Big(p(p-1)(1 + \|W_s^{k,m}\|^2)^{p-2}(W_s^{k,m})(W_s^{k,m})^T \Sigma_{k,m}(W_s^m)\Big) ds \\ + \frac{\gamma}{2}\int_0^t \mathrm{Tr}\Big(p(1 + \|W_s^{k,m}\|^2)^{p-1}\Sigma_{k,m}(W_s^m)\Big) ds.$$

Next, we apply expectation on both sides:

$$\mathbb{E}[(1+\|W_s^{k,m}\|^2)^p] = \mathbb{E}[(1+\|W_0^{k,m}\|^2)^p] + \mathbb{E}\left[\int_0^t p(1+\|W_s^{k,m}\|^2)^{p-1}(W_s^{k,m})^T h^{k,m}(W_s^m)ds\right] + \frac{\gamma p(p-1)}{2}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big((1+\|W_s^{k,m}\|^2)^{p-2}(W_s^{k,m})(W_s^{k,m})^T\Sigma_{k,m}(W_s^m)\Big)ds\right] + \frac{p\gamma}{2}\mathbb{E}\left[\int_0^t \mathrm{Tr}\Big((1+\|W_s^{k,m}\|^2)^{p-1}\Sigma_{k,m}(W_s^m)\Big)ds\right].$$

By bounding each integral by the integral of the module:

$$\mathbb{E}[(1+\|W_s^{k,m}\|^2)^p] = \mathbb{E}[(1+\|W_0^{k,m}\|^2)^p] + \mathbb{E}\left[\int_0^t p(1+\|W_s^{k,m}\|^2)^{p-1}\|W_s^{k,m}\|\|h^{k,m}(W_s^m)\|ds\right] + \frac{\gamma p(p-1)}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^{k,m}\|^2)^{p-2}\|W_s^{k,m}\|^2\|\Sigma_{k,m}(W_s^m)\|ds\right] + \frac{p\gamma}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^{k,m}\|^2)^{p-1}\|\Sigma_{k,m}(W_s^m)\|ds\right].$$

Now, by using that the coefficients are Lipschitz and doing the same manipulation of the terms as in the last case:

$$\mathbb{E}[(1+\|W_s^{k,m}\|^2)^p] = C\mathbb{E}[(1+\|W_0^{k,m}\|^2)^p] + \mathbb{E}\left[\int_0^t p(1+\|W_s^{k,m}\|^2)^{p-1}\|W_s^m\|^2 ds\right] + \frac{\gamma p(p-1)}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^{k,m}\|^2)^{p-2}\|W_s^{k,m}\|^2\|W_s^m\|^2 ds\right] + \frac{p\gamma}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^{k,m}\|^2)^{p-1}\|W_s^{k,m}\|^2 ds\right].$$

Next, we bound $\mathbb{E}[\|W_s^{k,m}\|] \leq C\mathbb{E}[\|W_s^m\|]$ and obtain:

$$\mathbb{E}[(1+\|W_s^{k,m}\|^2)^p] = C\mathbb{E}[(1+\|W_0^m\|^2)^p] + \mathbb{E}\left[\int_0^t p(1+\|W_s^m\|^2)^{p-1}\|W_s^m\|^2 ds\right] + \frac{\gamma p(p-1)}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^m\|^2)^{p-2}\|W_s^m\|^2\|W_s^m\|^2 ds\right] + \frac{p\gamma}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^m\|^2)^{p-1}\|W_s^m\|^2 ds\right].$$

Since $\|W_s^m\|^2 \leq 1+\|W_s^m\|^2$, we obtain:

$$\mathbb{E}[(1+\|W_s^{k,m}\|^2)^p] = C\mathbb{E}[(1+\|W_0^m\|^2)^p] + \mathbb{E}\left[\int_0^t p(1+\|W_s^m\|^2)^p ds\right] + \frac{\gamma p(p-1)}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^m\|^2)ds\right] + \frac{p\gamma}{2}\mathbb{E}\left[\int_0^t (1+\|W_s^m\|^2)^p ds\right].$$

This way, we conclude:

$$\mathbb{E}[(1+\|W_s^{k,m}\|^2)^p] = C\mathbb{E}[(1+\|W_0^m\|^2)^p] + C\mathbb{E}\left[\int_0^t (1+\|W_s^m\|^2)^p ds\right].$$

Since this is true for every $k \in \{1, \dots, m\}$:

$$\mathbb{E}[(1 + \|W_s^m\|^2)^p] = C\mathbb{E}[(1 + \|W_0^m\|^2)^p] + C\mathbb{E}\left[\int_0^t (1 + \|W_s^m\|^2)^p ds\right],$$

and by Gronwall:

$$\mathbb{E}[(1 + \|W_s^m\|^2)^p] = C\mathbb{E}[(1 + \|W_0^m\|^2)^p].$$

Therefore, the equation propagates all the initial moments, and in particular the 4th one, which is what we wanted to prove. □