Mean Field Analysis of Neural Networks: A Central Limit Theorem

Justin Sirignano* and Konstantinos Spiliopoulos†‡

August 29, 2018

Abstract

Machine learning has revolutionized fields such as image, text, and speech recognition. There's also growing interest in applying machine and deep learning methods in science, engineering, medicine, and finance. Despite their immense success in practice, there is limited mathematical understanding of neural networks. We mathematically study neural networks in the asymptotic regime of simultaneously (A) large network sizes and (B) large numbers of stochastic gradient descent training iterations. We rigorously prove that the neural network satisfies a central limit theorem. Our result describes the neural network's fluctuations around its mean-field limit. The fluctuations have a Gaussian distribution and satisfy a stochastic partial differential equation.

1 Introduction

Neural networks have achieved immense practical success over the past decade in fields such as image, text, and speech recognition. Neural networks are nonlinear machine learning models whose parameters are estimated from data using stochastic gradient descent. Our result characterizes neural networks with a single hidden layer in the asymptotic regime of large network sizes and large numbers of stochastic gradient descent iterations. A law of large numbers was previously proven in [25]. This paper rigorously proves a central limit theorem (CLT) for the empirical distribution of the neural network parameters. The central limit theorem describes the fluctuations of the finite empirical distribution of the neural network parameters around its mean-field limit.

The mean-field limit (proven in [25]) is a law of large numbers for the empirical measure of the neural network parameters as $N \to \infty$. It satisfies a deterministic nonlinear partial differential equation. The mean-field limit of course is only accurate in the limit $N \to \infty$, and the central limit theorem provides a first-order correction in N. The central limit theorem quantifies the fluctuations of the finite N empirical measure around its mean-field limit. It satisfies a linear stochastic partial differential equation (SPDE) driven by a Gaussian process. In particular, our result shows that the trained neural network behaves as $\mu_t^N \approx \bar{\mu}_t + \frac{1}{\sqrt{N}}\bar{\eta}_t$ where μ_t^N is the empirical measure of the parameters for a neural network with N hidden units, $\bar{\mu}_t$ is the mean-field limit, and $\bar{\eta}_t$ is the Gaussian correction from the central limit theorem.

The proof requires a linearization of the nonlinear pre-limit evolution equation for the empirical distribution of the neural network parameters. This linearization produces several remainder terms which must be shown to vanish in the limit (similar to a perturbation analysis for PDEs). The SPDE for the CLT $\bar{\eta}_t$ is linearized around the nonlinear PDE for the mean-field limit $\bar{\mu}_t$. The CLT SPDE and mean-field limit PDE are therefore coupled. We must also show that the pre-limit evolution equation (which is in discrete time since stochastic gradient descent is a discrete-time algorithm) converges to a continuous-time limit.

The proof relies upon weak convergence analysis for interacting particle systems. The convergence analysis is technically challenging since the fluctuations of the empirical distribution is a signed-measure-valued process and its limit process turns out to be distribution-valued in the appropriate space. Unfortunately, the space of signed measures endowed with the weak topology is in general not metrizable (ee [8] and [27] for

*Department of Industrial & Systems Engineering, University of Illinois at Urbana Champaign, Urbana, E-mail: jasirign@illinois.edu

†Department of Mathematics and Statistics, Boston University, Boston, E-mail: kspiliop@math.bu.edu

‡K.S. was partially supported by the National Science Foundation (DMS 1550918)

further discussion of the space of signed measures). We study the convergence of the fluctuations as a process taking values in the dual space of an appropriate Sobolev space. We prove that the pre-limit fluctuation process is relatively compact in that space and that any limit point is unique in that space. In particular, we will use the dual space $W^{-J,2} = W^{-J,2}(\Theta)$ of the Sobolev space $W_0^{J,2}(\Theta)$ with Θ a bounded subset of the appropriate Euclidean space and where J is sufficiently large; see Section 2 for a detailed description. Since the pre-limit evolution equation has discrete updates, we study convergence in the Skorokhod space $D_{W^{-J,2}}([0,T])$. ($D_S([0,T])$ is the set of maps from $[0,T]$ into S which are right-continuous and which have left-hand limits.)

Most of the literature on central limit theorems for interacting particle systems considers continuous-time systems, see for example [12, 21, 27]. In contrast, in this article the pre-limit process is in discrete time and converges to a continuous-time limit process after an appropriate time rescaling. At a practical level, this shows that the relation between the number of particles ("hidden units" in the language of neural networks) and the number of stochastic gradient steps should be of the same order to have convergence and statistically good behavior. At a more mathematical level, this passage from discrete to continuous time produces a number of additional remainder terms that must be shown to vanish at the correct rate in order for a CLT to hold. We resolve all these issues for one-layer neural network models, rigorously establishing and characterizing the fluctuations limit.

Weak convergence analysis has been widely used in other fields (for example, see [14], [15], [5], [6], [7], [4], and [17] for a non-exhaustive list). In fact, mean field analysis has been actively used for many years to study biological neural networks and physical systems of interacting particles; see for example [9], [19], [23], [29], [26], and the references therein. Recently, [25], [30], [22], and [24] study mean-field limits of machine learning algorithms, including neural networks. In this paper, we rigorously establish a central limit theorem for neural networks trained with stochastic gradient descent.

Consider the one-layer neural network

$$g_\theta^N(x) = \frac{1}{N}\sum_{i=1}^{N} c^i \sigma(w^i \cdot x), \tag{1.1}$$

where for every $i \in \{1,\cdots,N\}$, $c^i \in \mathbb{R}$ and $x, w^i \in \mathbb{R}^d$. For notational convenience we shall interpret $w^i \cdot x = \sum_{j=1}^d w^{i,j}x^j$ as the standard scalar inner product. The neural network model has parameters $\theta = (c^1,\dots,c^N,w^1,\dots,w^N) \in \mathbb{R}^{(1+d)N}$, which must be estimated from data.

The neural network (1.1) takes a linear function of the original data, applies an element-wise nonlinear operation using the function $\sigma : \mathbb{R} \to \mathbb{R}$, and then takes another linear function to produce the output. The activation function $\sigma(\cdot)$ is a nonlinear function such as a sigmoid or tanh function. The quantity $\sigma(w^i \cdot x)$ is referred to as the i-th "hidden unit", and the vector $\big(\sigma(w^1 \cdot x),\dots,\sigma(w^N \cdot x)\big)$ is called the "hidden layer". The number of units in the hidden layer is N.

The objective function is

$$L(\theta) = \mathbb{E}_{Y,X}[(Y - g_\theta^N(X))^2], \tag{1.2}$$

where the data (Y,X) is assumed to have a joint distribution $\pi(dx,dy)$. We shall write $\mathcal{X}$ and $\mathcal{Y}$ for the state spaces of X and Y, respectively. The parameters $\theta = (c^1,\dots,c^N,w^1,\dots,w^N)$ are estimated using stochastic gradient descent:

$$\begin{aligned}
c^i_{k+1} &= c^i_k + \frac{\alpha}{N}(y_k - g^N_{\theta_k}(x_k))\sigma(w^i_k \cdot x_k),\\
w^{i,j}_{k+1} &= w^{i,j}_k + \frac{\alpha}{N}(y_k - g^N_{\theta_k}(x_k))c^i_k\sigma'(w^i_k \cdot x_k)x^j_k, \quad j = 1,\cdots,d,
\end{aligned} \tag{1.3}$$

where α is the learning rate and $(x_k,y_k) \sim \pi(dx,dy)$. Stochastic gradient descent minimizes (1.2) using a sequence of noisy (but unbiased) gradient descent steps $\nabla_\theta[(y_k - g^N_{\theta_k}(x_k))^2]$. Note that typically $\nabla_\theta[(y - g_\theta^N(x))^2]$ is not a priori globally Lipschitz nor globally bounded as a function of θ. Stochastic gradient descent typically converges more rapidly than gradient descent for large datasets. For this reason, stochastic gradient descent is widely used in machine learning.

Define the empirical measure

$$\nu_k^N(dc, dw) = \frac{1}{N}\sum_{i=1}^N \delta_{c_k^i, w_k^i}(dc, dw).$$

The neural network's output can be re-written in terms of the empirical measure:

$$g_{\theta_k}^N(x) = \left\langle c\sigma(w \cdot x), \nu_k^N \right\rangle.$$

$\langle f, h\rangle$ denotes the inner product of f and h. For example, $\left\langle c\sigma(w \cdot x), \nu_k^N \right\rangle = \int c\sigma(w \cdot x)\nu_k^N(dc, dw)$.

The scaled empirical measure is

$$\mu_t^N = \nu_{\lfloor Nt \rfloor}^N.$$

The scaled empirical measure μ^N is a random element of the Skorokhod space $D_E([0,T]) = D([0,T]; E)$[1] with $E = \mathcal{M}(\mathbb{R}^{1+d})$.

We shall work on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all the random variables are defined. The probability space is equipped with a filtration that is right continuous and contains all $\mathbb{P}$-null sets.

We impose the following conditions.

Assumption 1.1. We have that

- The activation function $\sigma \in C_b^\infty(\mathbb{R})$.
- The data $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is compactly supported.
- The sequence of data samples (x_k, y_k) is i.i.d.
- The random initialization (c_0^i, w_0^i) is i.i.d, generated from a distribution with compact support.

1.1 Law of Large Numbers

[25] proves the mean-field limit $\mu^N \xrightarrow{p} \bar{\mu}$ as $N \to \infty$. The convergence theorems of [25] are summarized below.

Theorem 1.2. *Assume Assumption 1.1. The scaled empirical measure μ_t^N converges in distribution to $\bar{\mu}_t$ in $D_E([0,T])$ as $N \to \infty$. For every $f \in C_b^2(\mathbb{R}^{1+d})$, $\bar{\mu}$ satisfies the measure evolution equation*

$$\langle f, \bar{\mu}_t \rangle = \langle f, \bar{\mu}_0 \rangle + \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c'\sigma(w' \cdot x), \bar{\mu}_s\rangle\big) \left\langle \nabla(c\sigma(w \cdot x)) \cdot \nabla f, \bar{\mu}_s \right\rangle \pi(dx, dy) \right) ds, \quad (1.4)$$

where $\nabla f = (\partial_c f, \nabla_w f)$.

Remark 1.3. Since weak convergence to a constant implies convergence in probability, Theorem 1.2 leads to the stronger result of convergence in probability

$$\lim_{N\to\infty} \mathbb{P}\left\{ d_E(\mu^N, \bar{\mu}) \geq \delta \right\} = 0$$

for every $\delta > 0$ and where d_E is the metric for $D_E([0,T])$.

Corollary 1.4. *Assume Assumption 1.1. Suppose that $\bar{\mu}_0$ admits a density $p_0(c, w)$ and there exists a solution to the nonlinear partial differential equation*

$$\begin{aligned}
\frac{\partial p(t,c,w)}{\partial t} &= -\alpha \int_{\mathcal{X}\times\mathcal{Y}} \left(\big(y - \langle c'\sigma(w' \cdot x), p(t, c', w')\rangle\big) \frac{\partial}{\partial c}\big[\sigma(w \cdot x) p(t,c,w)\big] \right) \pi(dx, dy) \\
&\quad -\alpha \int_{\mathcal{X}\times\mathcal{Y}} \left(\big(y - \langle c'\sigma(w' \cdot x), p(t, c', w')\rangle\big) x \cdot \nabla_w \big[c\sigma'(w \cdot x) p(t,c,w)\big] \right) \pi(dx, dy), \\
p(0, c, w) &= p_0(c, w).
\end{aligned}$$

Then, we have that the solution to the measure evolution equation (1.4) is such that

$$\bar{\mu}_t(dc, dw) = p(t, c, w)dcdw.$$

[1] $D_S([0,T])$ is the set of maps from $[0,T]$ into S which are right-continuous and which have left-hand limits.

1.2 Main Result: A Central Limit Theorem

In this paper, we prove a central limit theorem for one-layer neural networks as the size of the network and the number of training steps become large. The central limit theorem quantifies the speed of convergence of the finite neural network to its mean-field limit as well as how the finite neural network fluctuates around the mean-field limit for large N.

Define the fluctuation process

$$\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t).$$

We prove that $\eta^N \xrightarrow{d} \bar{\eta}$, where $\bar{\eta}$ satisfies a stochastic partial differential equation. This result characterizes the fluctuations of the finite empirical measure μ^N around its mean-field limit $\bar{\mu}$ for large N. The limit $\bar{\eta}$ has a Gaussian distribution. We study the convergence of η_t^N in the space $D_{W^{-J,2}}([0,T])$, where $W^{-J,2} = W^{-J,2}(\Theta)$ is the dual of the Sobolev space $W_0^{J,2}(\Theta)$ with $\Theta \subset \mathbb{R}^{1+d}$ a bounded domain. These spaces are described in detail in Section 2.

Theorem 1.5. *Assume Assumption 1.1 and let* $J \geq 3\lceil \frac{d+1}{2} \rceil + 7$. *Let* $0 < T < \infty$ *be given. The sequence* $\{\eta_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$ *is relatively compact in* $D_{W^{-J,2}}([0,T])$. *The sequence of processes* $\{\eta_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$ *converges in distribution in* $D_{W^{-J,2}}([0,T])$ *to the process* $\{\bar{\eta}_t, t \in [0,T]\}$, *which, for every* $f \in W_0^{J,2}(\Theta)$, *satisfies the stochastic partial differential equation*

$$\begin{aligned}\langle f, \bar{\eta}_t \rangle &= \langle f, \bar{\eta}_0 \rangle + \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle\big) \langle \nabla(c\sigma(w\cdot x))\cdot \nabla f, \bar{\eta}_s\rangle \, \pi(dx,dy)ds \\ &\quad - \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \langle c\sigma(w\cdot x), \bar{\eta}_s\rangle \langle \nabla(c\sigma(w\cdot x))\cdot\nabla f, \bar{\mu}_s\rangle \, \pi(dx,dy)ds + \langle f, \bar{M}_t\rangle. \qquad (1.5)\end{aligned}$$

$\bar{M}_t$ *is a mean-zero Gaussian process; see Lemma 5.2 for its covariance structure. Finally, the stochastic evolution equation (1.5) has a unique solution in* $W^{-J,2}$, *which implies that* $\bar{\eta}$ *is unique.*

The CLT SPDE (1.5) is coupled with the mean-field limit PDE (1.4). (1.4) is a deterministic nonlinear PDE while (1.5) is a stochastic linear PDE. The SPDE (1.5) is linear in $\bar{\eta}$ and driven by a Gaussian process; therefore, the CLT $\bar{\eta}_t$ itself is a Gaussian process.

Theorem 1.5 indicates that for large N the empirical distribution of the neural network's parameters behaves as

$$\mu^N \approx \bar{\mu} + \frac{1}{\sqrt{N}}\bar{\eta},$$

where $\bar{\eta}$ has a Gaussian distribution.

1.3 Outline of Paper

The Sobolev spaces which we study convergence in are presented in Section 2. The pre-limit evolution equation for the fluctuation process η^N is derived in Section 3. Section 4 proves relative compactness. Section 5 derives the limiting SPDE (1.5). Uniqueness of the SPDE (1.5) is proven in Section 6. Section 7 collects these results and proves Theorem 1.5. Conclusions are in Section 8.

2 Sobolev Spaces

We study convergence in a Sobolev space [1]. Weighted Sobolev spaces have been previously used to study central limit theorems of mean field systems in papers such as [12], [21] and [27]. Weights are not necessary in this paper since η_t^N and μ_t^N are compactly supported uniformly with respect to $N \in \mathbb{N}$ and $t \in [0,T]$ (see Lemma 4.2).

Let $\Theta \subset \mathbb{R}^D$ be a bounded domain with $D = d+1$. For any integer $J \in \mathbb{N}$, consider the space of real valued functions f with partial derivatives up to order J which satisfy

$$\|f\|_J = \left(\sum_{|k| \leq J} \int_\Theta \left| D^k f(x) \right|^2 dx \right)^{1/2} < \infty.$$

Define the space $W_0^{J,2}(\Theta)$ as the closure of functions of class $C_0^\infty(\Theta)$ in the norm defined above. $C_0^\infty(\Theta)$ is the space of all functions in $C^\infty(\Theta)$ with compact support. (The space $W_0^{J,2}(\Theta)$ is frequently also denoted by $H_0^J(\Theta)$ in the literature.) $W_0^{J,2}(\Theta)$ is a Hilbert space (see Theorem 3.5 and Remark 3.33 in [1]) and has the inner product

$$\langle f, g \rangle_J = \sum_{|k| \leq J} \int_\Theta D^k f(x) D^k g(x) dx.$$

When $J = 0$, we write $\langle f, g \rangle_0 = \langle f, g \rangle$. $W^{-J,2}(\Theta)$ denotes the dual space of $W_0^{J,2}(\Theta)$ that is equipped with the norm

$$\|f\|_{-J} = \sup_{g \in W_0^{J,2}(\Theta)} \frac{| \langle f, g \rangle |}{\|g\|_J}.$$

We will study convergence in the Sobolev space corresponding to $J \geq 3\lceil \frac{D}{2} \rceil + 7$. From Lemma 4.2, we have that μ_t^N and η_t^N are compactly supported. In particular, there exists a compact set $K = [-C_o, C_o]^D \subset \mathbb{R}^D$ such that μ_t^N and η_t^N vanish outside the compact set K for every $N \in \mathbb{N}$ and $t \in [0, T]$. We choose $\Theta = (-B, B)^D$ where $B = 3\sqrt{D}C_o$. Note that C_o, and thus the domain Θ, may depend upon fixed parameters of the problem such that T, α, $\pi(dx, dy)$, and $\bar{\mu}_0$, but what is important is that the bounded set Θ is fixed and does not change with $N \in \mathbb{N}$ or $t \in [0, T]$.

Sometimes, we may write for simplicity $W^{-J,2}$ in place of $W^{-J,2}(\Theta)$ and $W_0^{J,2}$ in place of $W_0^{J,2}(\Theta)$.

3 Preliminary Calculations

The goal of this section is to write $\langle f, \eta_t^N \rangle$, with η_t^N being the fluctuation process and $f \in C_b^2(\mathbb{R}^{1+d})$ a test function, in a way that allows us to take limits. In particular, our goal is describe the evolution of $\langle f, \eta_t^N \rangle$ in terms of the equation (3.5). In order to do this, we need some preliminary computations.

We consider the evolution of the empirical measure ν_k^N via test functions $f \in C_b^2(\mathbb{R}^{1+d})$. A Taylor expansion yields

$$\begin{aligned}
\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle &= \frac{1}{N} \sum_{i=1}^N \Big(f(c_{k+1}^i, w_{k+1}^i) - f(c_k^i, w_k^i) \Big) \\
&= \frac{1}{N} \sum_{i=1}^N \partial_c f(c_k^i, w_k^i)(c_{k+1}^i - c_k^i) + \frac{1}{N} \sum_{i=1}^N \nabla_w f(c_k^i, w_k^i)^\top (w_{k+1}^i - w_k^i) \\
&\quad + \frac{1}{N} \sum_{i=1}^N \partial_c^2 f(\bar{c}_k^i, \bar{w}_k^i)(c_{k+1}^i - c_k^i)^2 + \frac{1}{N} \sum_{i=1}^N (c_{k+1}^i - c_k^i) \nabla_{cw} f(\bar{c}_k^i, \bar{w}_k^i)(w_{k+1}^i - w_k^i) \\
&\quad + \frac{1}{N} \sum_{i=1}^N (w_{k+1}^i - w_k^i)^\top \nabla_w^2 f(\bar{c}_k^i, \bar{w}_k^i)(w_{k+1}^i - w_k^i),
\end{aligned} \tag{3.1}$$

for points $\bar{c}_k^i, \bar{w}_k^i$ in the segments connecting c_{k+1}^i with c_k^i and w_{k+1}^i with w_k^i, respectively. [25] has shown that the parameters are uniformly bounded (in both $0 \leq k \leq NT$ and N):

$$|c_k^i| + \|w_k^i\| < C_o. \tag{3.2}$$

Using the relation (1.3), equation (3.1) becomes

$$
\begin{aligned}
\left\langle f,\nu^N_{k+1}\right\rangle-\left\langle f,\nu^N_{k}\right\rangle &= \frac{1}{N^2}\sum_{i=1}^N\partial_c f(c^i_k,w^i_k)\alpha(y_k-g^N_{\theta_k}(x_k))\sigma(w^i_k\cdot x_k)\\
&\quad+\frac{1}{N^2}\sum_{i=1}^N\alpha(y_k-g^N_{\theta_k}(x_k))c^i_k\sigma'(w^i_k\cdot x_k)\nabla_w f(c^i_k,w^i_k)\cdot x_k+\frac{G^N_k}{N^2}.
\end{aligned}
$$

where $\frac{G^N_k}{N^2}$ is an $O\left(N^{-2}\right)$ term with

$$
\begin{aligned}
G^N_k &= N^2\bigg(\frac{1}{N}\sum_{i=1}^N\partial^2_c f(\bar c^i_k,\bar w^i_k)(c^i_{k+1}-c^i_k)^2+\frac{1}{N}\sum_{i=1}^N(c^i_{k+1}-c^i_k)\nabla_{cw}f(\bar c^i_k,\bar w^i_k)(w^i_{k+1}-w^i_k)\\
&\quad+\frac{1}{N}\sum_{i=1}^N(w^i_{k+1}-w^i_k)^\top\nabla^2_w f(\bar c^i_k,\bar w^i_k)(w^i_{k+1}-w^i_k)\bigg).
\end{aligned}
$$

Note that $|G^N_k|<C\sum_{|\alpha|=2}\sup_{c,w\in K}|D^\alpha f(c,w)|$ due to the uniform bound $|c^i_k|+\left\|w^i_k\right\|<C_o$, (X,Y) having compact support, and the relation (1.3). $K\subset\mathbb{R}^{1+d}$ is the compact set $K=[-C_o,C_o]^{1+d}$.

We next define the drift and martingale components:

$$
\begin{aligned}
D^{1,N}_k &= \frac{1}{N}\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\nu^N_k\right\rangle\big)\left\langle\sigma(w\cdot x)\partial_c f,\nu^N_k\right\rangle\pi(dx,dy),\\
D^{2,N}_k &= \frac{1}{N}\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\nu^N_k\right\rangle\big)\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f,\nu^N_k\right\rangle\pi(dx,dy),\\
\left\langle f,M^{1,N}_k\right\rangle &= \frac{1}{N}\alpha\big(y_k-\left\langle c\sigma(w\cdot x_k),\nu^N_k\right\rangle\big)\left\langle\sigma(w\cdot x_k)\partial_c f,\nu^N_k\right\rangle-D^{1,N}_k,\\
\left\langle f,M^{2,N}_k\right\rangle &= \frac{1}{N}\alpha\big(y_k-\left\langle c\sigma(w\cdot x_k),\nu^N_k\right\rangle\big)\left\langle c\sigma'(w\cdot x_k)x\cdot\nabla_w f,\nu^N_k\right\rangle-D^{2,N}_k.
\end{aligned}
$$

Combining the different terms together, we then obtain

$$
\left\langle f,\nu^N_{k+1}\right\rangle-\left\langle f,\nu^N_{k}\right\rangle = D^{1,N}_k+D^{2,N}_k+\left\langle f,M^{1,N}(t)\right\rangle+\left\langle f,M^{2,N}(t)\right\rangle+O\left(N^{-2}\right).
$$

Next, we define the scaled versions of $D^{1,N},D^{2,N},M^{1,N}$ and $M^{2,N}$:

$$
\begin{aligned}
D^{1,N}(t) &= \sum_{k=0}^{\lfloor Nt\rfloor-1}D^{1,N}_k,\qquad D^{2,N}(t)=\sum_{k=0}^{\lfloor Nt\rfloor-1}D^{2,N}_k,\\
\left\langle f,M^{1,N}(t)\right\rangle &= \sum_{k=0}^{\lfloor Nt\rfloor-1}\left\langle f,M^{1,N}_k\right\rangle,\qquad \left\langle f,M^{2,N}(t)\right\rangle=\sum_{k=0}^{\lfloor Nt\rfloor-1}\left\langle f,M^{2,N}_k\right\rangle.
\end{aligned}
$$

We also define

$$
\left\langle f,M^N_t\right\rangle=\left\langle f,M^{1,N}(t)\right\rangle+\left\langle f,M^{2,N}(t)\right\rangle.
$$

$D^{1,N}(t)$ and $D^{2,N}(t)$ can be approximated by integrals:

$$
\begin{aligned}
\sum_{k=0}^{\lfloor Nt\rfloor-1}D^{1,N}_k &= \sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\nu^N_k\right\rangle\big)\left\langle\sigma(w\cdot x)\partial_c f,\nu^N_k\right\rangle\pi(dx,dy)ds\\
&= \sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\mu^N_s\right\rangle\big)\left\langle\sigma(w\cdot x)\partial_c f,\mu^N_s\right\rangle\pi(dx,dy)ds\\
&= \int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\mu^N_s\right\rangle\big)\left\langle\sigma(w\cdot x)\partial_c f,\mu^N_s\right\rangle\pi(dx,dy)ds+V^{1,N}_t,
\end{aligned}
$$

where $V_t^{1,N}$ is a remainder term defined below. Similarly,

$$\sum_{k=0}^{\lfloor Nt\rfloor-1} D_k^{2,N} = \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle\big) \langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N\rangle \pi(dx,dy)ds + V_t^{2,N}.$$

The remainder terms $V_t^{1,N}$ and $V_t^{2,N}$ are

$$\begin{aligned}
V_t^{1,N} &= -\int_{\frac{\lfloor Nt\rfloor}{N}}^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle\big) \langle \sigma(w\cdot x)\partial_c f, \mu_s^N\rangle \pi(dx,dy)ds,\\
V_t^{2,N} &= -\int_{\frac{\lfloor Nt\rfloor}{N}}^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle\big) \langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N\rangle \pi(dx,dy)ds,\\
V_t^N &= V_t^{1,N} + V_t^{2,N}.
\end{aligned}$$

V_t^N is a cádlág process with jumps at times $\frac{1}{N}, \frac{2}{N}, \ldots, \frac{\lfloor NT\rfloor}{N}$. Furthermore, due to the uniform bound (3.2) and $\mathcal{X}\times\mathcal{Y}$ being a compact set, V_t^N is an $\mathcal{O}(N^{-1})$ remainder term:

$$\sup_{t\in[0,T]} |V_t^N| \leq \frac{C}{N} \sum_{|\alpha|=1} \sup_{c,w\in K} |D^\alpha f(c,w)| \tag{3.3}$$

The scaled empirical measure can be written as the telescoping sum

$$\begin{aligned}
\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle &= \Big\langle f, \nu_{\lfloor Nt\rfloor}^N\Big\rangle - \langle f, \nu_0^N\rangle\\
&= \bigg(\Big\langle f, \nu_{\lfloor Nt\rfloor}^N\Big\rangle - \Big\langle f, \nu_{\lfloor Nt\rfloor-1}^N\Big\rangle\bigg) + \bigg(\Big\langle f, \nu_{\lfloor Nt\rfloor-1}^N\Big\rangle - \Big\langle f, \nu_{\lfloor Nt\rfloor-2}^N\Big\rangle\bigg)\\
&\quad + \ldots + \bigg(\langle f, \nu_1^N\rangle - \langle f, \nu_0^N\rangle\bigg)\\
&= \sum_{k=0}^{\lfloor Nt\rfloor-1} \bigg(\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle\bigg).
\end{aligned}$$

Therefore, the scaled empirical measure satisfies

$$\begin{aligned}
\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle &= \sum_{k=0}^{\lfloor Nt\rfloor-1} \bigg(\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle\bigg)\\
&= \sum_{k=0}^{\lfloor Nt\rfloor-1} \bigg(D_k^{1,N} + D_k^{2,N} + \langle f, M^{1,N}(t)\rangle + \langle f, M^{2,N}(t)\rangle\bigg) + \frac{1}{N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1} G_k^N\\
&= \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle\big) \langle \sigma(w\cdot x)\partial_c f, \mu_s^N\rangle \pi(dx,dy)ds\\
&\quad + \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle\big) \langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N\rangle \pi(dx,dy)ds\\
&\quad + \langle f, M_t^N\rangle + \frac{1}{N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1} G_k^N + V_t^N
\end{aligned} \tag{3.4}$$

Note that $\frac{1}{N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1} G_k^N$ is $\mathcal{O}(N^{-1})$. Define the fluctuation process

$$\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t).$$

Then,

$$
\begin{aligned}
\left\langle f, \eta_t^N \right\rangle - \left\langle f, \eta_0^N \right\rangle &= \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle\big) \left\langle \sigma(w\cdot x)\partial_c f, \eta_s^N \right\rangle \pi(dx,dy) \right) ds \\
&- \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle \langle \sigma(w\cdot x)\partial_c f, \bar{\mu}_s\rangle \pi(dx,dy) \right) ds \\
&+ \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle\big) \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \eta_s^N \right\rangle \pi(dx,dy) \right) ds \\
&- \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle\big) \langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \bar{\mu}_s\rangle \pi(dx,dy) \right) ds \\
&+ \sqrt{N}\left\langle f, M_t^N \right\rangle + \Gamma_t^{1,N} + \Gamma_t^{2,N} + R_t^{1,N} + R_t^{2,N},
\end{aligned}
\tag{3.5}
$$

where

$$
\begin{aligned}
\Gamma_t^{1,N} &= \frac{1}{\sqrt{N}} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} -\alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \eta_s^N \right\rangle \pi(dx,dy) ds \\
\Gamma_t^{2,N} &= \frac{1}{\sqrt{N}} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} -\alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\nabla_w f, \eta_s^N \right\rangle \pi(dx,dy) ds.
\end{aligned}
$$

$R_t^{1,N}$ and $R_t^{2,N}$ are $\mathcal{O}(N^{-1/2})$ remainder terms where

$$
\begin{aligned}
R_t^{1,N} &= N^{-3/2} \sum_{k=0}^{\lfloor Nt \rfloor - 1} G_k^N, \\
R_t^{2,N} &= \sqrt{N} V_t^N.
\end{aligned}
$$

4 Relative Compactness

This section proves the relative compactness of the pre-limit processes $\{\eta_t^N, t\in[0,T]\}_{N\in\mathbb{N}}$ in $D_{W^{-J,2}}([0,T])$ and of $\{\sqrt{N}M_t^N, t\in[0,T]\}_{N\in\mathbb{N}}$ in $D_{W^{-J,2}}([0,T])$.

4.1 Uniform bound on the fluctuations process η^N

The main result of this section is Lemma 4.1 below and it provides a uniform bound with respect to $N\in\mathbb{N}$ and $t\in[0,T]$ for the process η_t^N.

Lemma 4.1. *If $J_1 = 2\lceil \frac{D}{2} \rceil + 4$, then there is a constant $C<\infty$ such that*

$$
\sup_{N\in\mathbb{N}} \sup_{t\in[0,T]} \mathbb{E} \left\| \eta_t^N \right\|_{-J_1}^2 < C. \tag{4.1}
$$

The proof of this lemma requires a number of intermediate results. We develop these estimates now and present the proof of Lemma 4.1 in the end of this section.

Consider the particle system

$$
\begin{aligned}
\tilde{c}_t^i &= c_0^i + \int_0^t \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle)\sigma(\tilde{w}_s^i\cdot x)\pi(dx,dy)ds, \\
\tilde{w}_t^i &= w_0^i + \int_0^t \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle)\tilde{c}_s^i\sigma'(\tilde{w}_s^i\cdot x)x\pi(dx,dy)ds. \\
\tilde{\mu}_t^N &= \frac{1}{N}\sum_{i=1}^N \delta_{(\tilde{c}_t^i, \tilde{w}_t^i)}.
\end{aligned}
$$

The particles $(\tilde{c}^i, \tilde{w}^i)$ are i.i.d. with law $\bar{\mu}$ and $\tilde{\mu}^N \xrightarrow{p} \bar{\mu}$. By the results of [25] we obtain that $\tilde{\mu}^N$ is also compactly supported uniformly in $N \in \mathbb{N}$ and $t \in [0,T]$. We decompose the η_t^N into two terms:

$$\eta_t^N = \sqrt{N}(\mu_t^N - \tilde{\mu}_t^N) + \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t). \tag{4.2}$$

Define $\Xi_t^N = \sqrt{N}(\mu_t^N - \tilde{\mu}_t^N)$. Then,

$$\begin{aligned}
\left\langle f, \Xi_t^N \right\rangle &= \sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle\big) \left\langle \sigma(w\cdot x)\partial_c f, \mu_s^N \right\rangle \pi(dx,dy)ds \\
&+ \sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle\big) \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N \right\rangle \pi(dx,dy)ds \\
&- \sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \left\langle c\sigma(w\cdot x), \bar{\mu}_s \right\rangle\big) \left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&- \sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \left\langle c\sigma(w\cdot x), \bar{\mu}_s \right\rangle\big) \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&+ \sqrt{N}\left\langle f, M_t^N \right\rangle + R_t^{1,N} + R_t^{2,N}.
\end{aligned} \tag{4.3}$$

By chain rule,

$$\begin{aligned}
\left\langle f, \Xi_t^N \right\rangle^2 &= 2\sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \big(y - \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle\big) \left\langle \sigma(w\cdot x)\partial_c f, \mu_s^N \right\rangle \pi(dx,dy)ds \\
&+ 2\sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \big(y - \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle\big) \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N \right\rangle \pi(dx,dy)ds \\
&- 2\sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \big(y - \left\langle c\sigma(w\cdot x), \bar{\mu}_s \right\rangle\big) \left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&- 2\sqrt{N} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \big(y - \left\langle c\sigma(w\cdot x), \bar{\mu}_s \right\rangle\big) \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&+ \sum_{k=0}^{\lfloor Nt\rfloor-1} \left(\left\langle f, \Xi_{\frac{k+1}{N}-}^N + \sqrt{N}M_k^{1,N} + \sqrt{N}M_k^{2,N} \right\rangle^2 - \left\langle f, \Xi_{\frac{k+1}{N}-}^N \right\rangle^2 \right) \\
&+ \tilde{R}_t^{1,N} + \tilde{R}_t^{2,N}.
\end{aligned} \tag{4.4}$$

$\tilde{R}_t^{1,N}$ is the remainder term:

$$\tilde{R}_t^{1,N} = \sum_{k=0}^{\lfloor Nt\rfloor-1} \left(\big(\left\langle f, \Xi_{\frac{k+1}{N}-}^N \right\rangle + G_k^N N^{-3/2} \big)^2 - \left\langle f, \Xi_{\frac{k+1}{N}-}^N \right\rangle^2 \right),$$

where $|G_k^N| < C \sum_{|\alpha|=2} \sup_{c,w\in K} |D^\alpha f(c,w)|$ due to the bound $|c_k^i| + \left\| w_k^i \right\| < C_o$ and $\pi(dx,dy)$ having compact support. $K \subset \mathbb{R}^{1+d}$ is a compact set.

By the Sobolev embedding (Theorem 6.2 in [1]), we have that

$$\sum_{|\alpha|\le 2} \sup_{c,w\in K} |D^\alpha f(c,w)| \le C \left\| f \right\|_L \tag{4.5}$$

where $L = \lceil \frac{D}{2} \rceil + 3$.

Therefore,

$$\begin{aligned}
|\tilde{R}_t^{1,N}| &\leq \frac{C_1 \|f\|_L}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \left\langle f, \mu^N_{\frac{k}{N}} - \tilde{\mu}^N_{\frac{k+1}{N}} \right\rangle + C_2 N^{-2} \|f\|_L^2 \\
&\leq \frac{C_1 \|f\|_L}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sup_{c,w \in K} |f(c,w)| + C_2 N^{-2} \|f\|_L \\
&\leq C_1 \|f\|_L^2 + C_2 N^{-2} \|f\|_L^2 \\
&\leq C \|f\|_L^2 .
\end{aligned}$$

$\tilde{R}_t^{2,N}$ is the remainder term:

$$\begin{aligned}
\tilde{R}_t^{2,N} = -2\sqrt{N} \int_{\frac{\lfloor Nt \rfloor}{N}}^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left(y - \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle \right) \left\langle \sigma(w\cdot x)\partial_c f, \mu_t^N \right\rangle \pi(dx,dy)ds \\
- 2\sqrt{N} \int_{\frac{\lfloor Nt \rfloor}{N}}^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left(y - \left\langle c\sigma(w\cdot x), \mu_t^N \right\rangle \right) \left\langle c\sigma'(w\cdot x) x \cdot \nabla_w f, \mu_t^N \right\rangle \pi(dx,dy)ds
\end{aligned}$$

Using Young's inequality, compactness of $\mathcal{X}\times\mathcal{Y}$, and the bound (3.2),

$$\begin{aligned}
|\tilde{R}_t^{2,N}| &\leq C_1 \int_{\frac{\lfloor Nt \rfloor}{N}}^t \left\langle f, \Xi_s^N \right\rangle^2 ds + C_2 N \int_{\frac{\lfloor Nt \rfloor}{N}}^t \|f\|_L^2 \, ds \\
&\leq C_1 \int_0^t \left\langle f, \Xi_s^N \right\rangle^2 ds + C_2 \|f\|_L^2 .
\end{aligned}$$

Hence, we have obtained that

$$|\tilde{R}_t^{1,N}| + |\tilde{R}_t^{2,N}| \leq C_1 \int_0^t \left\langle f, \Xi_s^N \right\rangle^2 ds + C_2 \|f\|_L^2 . \tag{4.6}$$

We then notice that

$$\begin{aligned}
&\sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbb{E}\Big[\sqrt{N} \left\langle f, \Xi^N_{\frac{k+1}{N}-} \right\rangle \left\langle f, M_k^{1,N} + M_k^{2,N} \right\rangle \Big] \\
= \; & N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbb{E}\Big[\left\langle f, \mu^N_{\frac{k+1}{N}-} - \tilde{\mu}^N_{\frac{k+1}{N}} \right\rangle \left\langle f, M_k^{1,N} + M_k^{2,N} \right\rangle \Big] \\
= \; & N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbb{E}\Big[\left\langle f, \nu_k^N \right\rangle \mathbb{E}\Big[\left\langle f, M_k^{1,N} + M_k^{2,N} \right\rangle \Big| \mathcal{F}_k^N \Big]\Big] \\
& - N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbb{E}\Big[\mathbb{E}\Big[\left\langle f, \tilde{\mu}^N_{\frac{k+1}{N}} \right\rangle \mathbb{E}\Big[\left\langle f, M_k^{1,N} + M_k^{2,N} \right\rangle \Big| \mathcal{F}_k^N \Big] \Big| \mathcal{F}_0^N \Big]\Big] \\
= \; & 0,
\end{aligned} \tag{4.7}$$

where $\mathcal{F}_k^N$ is the $\sigma-$algebra generated by $(c_0^i, w_0^i)_{i=1}^N$ and $(x_j, y_j)_{j=0}^{k-1}$. In the fourth line we use the conditional independence of $\left\langle f, M_k^{1,N} + M_k^{2,N} \right\rangle$ and $\tilde{\mu}^N_{\frac{k+1}{N}}$ given the initial values $\{w_0^i, c_0^i\}_{i=1}^N$. Also, since μ_t^N only changes at discrete times due to jumps, $\mu^N_{\frac{k+1}{N}-} = \nu_k^N$.

We have also used the fact that the conditional expectation

$$
\begin{aligned}
\mathbb{E}\Big[\Big\langle f, M_k^{1,N}\Big\rangle \Big| \mathcal{F}_k^N\Big] &= \mathbb{E}\Big[\frac{1}{N}\alpha\big(y_k - \langle c\sigma(w\cdot x_k), \nu_k^N\rangle\big)\langle \sigma(w\cdot x_k)\partial_c f, \nu_k^N\rangle - D_k^{1,N}\Big|\mathcal{F}_k^N\Big]\\
&= \frac{\alpha}{N^2}\sum_{i=1}^N \mathbb{E}\Big[\Big(\big(y_k - \langle c\sigma(w\cdot x_k), \nu_k^N\rangle\big)\sigma(w_k^i\cdot x_k)\partial_c f(c_k^i, w_k^i)\\
&\quad - \int_{\mathcal{X}\times\mathcal{Y}}\big(y - \langle c\sigma(w\cdot x), \nu_k^N\rangle\big)\sigma(w_k^i\cdot x)\partial_c f(c_k^i, w_k^i)\pi(dx,dy)\Big)\Big|\mathcal{F}_k^N\Big]\\
&= 0.
\end{aligned}
$$

Similarly, $\mathbb{E}\Big[\Big\langle f, M_k^{2,N}\Big\rangle \Big| \mathcal{F}_k^N\Big] = 0$.

Now we can treat the term $\mathbb{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor -1}\Big(\big(\Big\langle f, \Xi^N_{\frac{k+1}{N}-} + \sqrt{N}M_k^{1,N} + \sqrt{N}M_k^{2,N}\Big\rangle\big)^2 - \Big\langle f, \Xi^N_{\frac{k+1}{N}-}\Big\rangle^2\Big)\Big]$ from (4.4) and get

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor -1}\Big(\big(\Big\langle f, \Xi^N_{\frac{k+1}{N}-} + \sqrt{N}M_k^{1,N} + \sqrt{N}M_k^{2,N}\Big\rangle\big)^2 - \Big\langle f, \Xi^N_{\frac{k+1}{N}-}\Big\rangle^2\Big)\Big]\\
=\ &\mathbb{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor -1}\Big(2\sqrt{N}\Big\langle f, \Xi^N_{\frac{k+1}{N}-}\Big\rangle\Big\langle f, M_k^{1,N} + M_k^{2,N}\Big\rangle + N\Big\langle f, M_k^{1,N} + M_k^{2,N}\Big\rangle^2\Big)\Big]\\
=\ &\mathbb{E}\Big[N\sum_{k=0}^{\lfloor Nt\rfloor -1}\Big\langle f, M_k^{1,N} + M_k^{2,N}\Big\rangle^2\Big]\\
=\ &\alpha^2\mathbb{E}\Big[\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}\Big(\big(y_k - \langle c\sigma(w\cdot x_k), \nu_k^N\rangle\big)\langle \sigma(w\cdot x_k)\partial_c f, \nu_k^N\rangle\\
&+\big(y_k - \langle c\sigma(w\cdot x_k), \nu_k^N\rangle\big)\langle c\sigma'(w\cdot x_k)x\cdot\nabla_w f, \nu_k^N\rangle\\
&-\int_{\mathcal{X}\times\mathcal{Y}}\big(y - \langle c\sigma(w\cdot x), \nu_k^N\rangle\big)\langle \sigma(w\cdot x)\partial_c f, \nu_k^N\rangle\pi(dx,dy)\\
&-\int_{\mathcal{X}\times\mathcal{Y}}\big(y - \langle c\sigma(w\cdot x), \nu_k^N\rangle\big)\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \nu_k^N\rangle\pi(dx,dy)\Big)^2\\
<\ &C\big(\sum_{|\alpha|=1}\sup_{c,w\in K}|D^\alpha f(c,w)|\big)^2 \le C\,\|f\|_L^2\,. \qquad (4.8)
\end{aligned}
$$

Next, we employ a decomposition into several terms in order to study the first and third term of (4.4) (and similarly for the terms two and four of (4.4)).

$$
\begin{aligned}
&\sqrt{N}\Big[\big(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle\big)\langle \sigma(w\cdot x)\partial_c f, \mu_s^N\rangle - \big(y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle\big)\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N\rangle\Big]\\
=\ &y\langle \sigma(w\cdot x)\partial_c f, \Xi_s^N\rangle\\
&-\langle c\sigma(w\cdot x), \mu_s^N\rangle\langle \sigma(w\cdot x)\partial_c f, \Xi_s^N\rangle\\
&-\langle c\sigma(w\cdot x), \Xi_s^N\rangle\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N\rangle\\
&-\Big\langle c\sigma(w\cdot x), \sqrt{N}(\tilde{\mu}_s^N - \bar{\mu}_s)\Big\rangle\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N\rangle\,.
\end{aligned}
$$

Therefore, equation (4.4) can be re-written as:

$$
\begin{aligned}
\left\langle f, \Xi_t^N \right\rangle^2 &= 2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha y \left\langle f, \Xi_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&+2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha y \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \Xi_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \sqrt{N}(\tilde{\mu}_s^N - \bar{\mu}_s) \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \Xi_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \sqrt{N}(\tilde{\mu}_s^N - \bar{\mu}_s) \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&+\sum_{k=0}^{\lfloor Nt\rfloor -1} \left(\left\langle f, \Xi_{\frac{k+1}{N}-}^N + \sqrt{N} M_k^{1,N} + \sqrt{N} M_k^{2,N} \right\rangle \right)^2 - \left\langle f, \Xi_{\frac{k+1}{N}-}^N \right\rangle^2 \\
&+\tilde{R}_t^{1,N} + \tilde{R}_t^{2,N}.
\end{aligned}
$$

Using the bounds (4.6) and (4.8), equation (4.4) gives

$$
\begin{aligned}
\mathbb{E}\left[\left\langle f, \Xi_t^N \right\rangle^2 \right] &\leq 2\mathbb{E}\Bigg[\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha y \left\langle f, \Xi_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&+2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha y \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \Xi_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \sqrt{N}(\tilde{\mu}_s^N - \bar{\mu}_s) \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \mu_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \Xi_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \Xi_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \\
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle f, \Xi_s^N \right\rangle \left\langle c\sigma(w\cdot x), \sqrt{N}(\tilde{\mu}_s^N - \bar{\mu}_s) \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \tilde{\mu}_s^N \right\rangle \pi(dx,dy)ds \Bigg] \\
&+C_1 \int_0^t \mathbb{E}\left[\left\langle f, \Xi_s^N \right\rangle^2 \right] ds + C_2 \left\| f \right\|_L^2 . \qquad (4.9)
\end{aligned}
$$

We begin with the fourth term in (4.9); the seventh term can be treated completely analogously and is omitted. First, notice that for any $x \in \mathcal{X}$,

$$
\left\langle c\sigma(w\cdot x), \Xi_s^N \right\rangle^2 \leq \left\| c\sigma(w\cdot x) \right\|_{J_1}^2 \left\| \Xi_s^N \right\|_{-J_1}^2 \leq C \left\| \Xi_s^N \right\|_{-J_1}^2 , \qquad (4.10)
$$

due to Assumption 1.1 and to the compactness of Θ.

Using Young's inequality, (4.10), and (4.5) to bound $\left\langle \sigma(w\cdot x)\partial_c f, \tilde{\mu}_s^N\right\rangle^2 \leq C\big(\sum_{|\alpha|=1}\sup_{(c,w)\in K}|D^\alpha f(c,w)|\big)^2 \leq C\left\|f\right\|_L^2$, we obtain

$$\begin{aligned}
&-\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\left\langle f,\Xi_s^N\right\rangle\left\langle c\sigma(w\cdot x),\Xi_s^N\right\rangle\left\langle \sigma(w\cdot x)\partial_c f,\tilde{\mu}_s^N\right\rangle\pi(dx,dy)ds\\
\leq\;& C\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\left(\left\langle f,\Xi_s^N\right\rangle^2+\left\|\Xi_s^N\right\|_{-J_1}^2\left\|f\right\|_L^2\right)\pi(dx,dy)ds.
\end{aligned}$$

Next, we study the fifth term in (4.9); the eighth term can be treated completely analogously and is omitted. The term $\left\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_s^N-\bar{\mu}_s)\right\rangle$ can be re-written as

$$\left\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_t^N-\bar{\mu}_t)\right\rangle = N^{-1/2}\sum_{i=1}^N\left(\tilde{c}_t^i\sigma(\tilde{w}_t^i x)-\left\langle c\sigma(wx),\bar{\mu}_t\right\rangle\right).$$

Since $(\tilde{c}_t^i,\tilde{w}_t^i)$ are i.i.d. random variables with law $\bar{\mu}_t$ and x takes values in the compact set $\mathcal{X}$,

$$\mathbb{E}\left[\left\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_t^N-\bar{\mu}_t)\right\rangle^2\right]\leq C.$$

Using Young's inequality and the fact that $\bar{\mu}$ takes values in a compact set K,

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}-\left\langle f,\Xi_s^N\right\rangle\left\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_s^N-\bar{\mu}_s)\right\rangle\left\langle\sigma(w\cdot x)\partial_c f,\tilde{\mu}_s^N\right\rangle\pi(dx,dy)ds\right]\\
\leq\;& C\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathbb{E}\big[\left\langle f,\Xi_s^N\right\rangle^2\big]+\mathbb{E}\big[\left\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_s^N-\bar{\mu}_s)\right\rangle^2\left\langle\sigma(w\cdot x)\partial_c f,\tilde{\mu}_s^N\right\rangle^2\big]\pi(dx,dy)ds\\
\leq\;& C\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathbb{E}\big[\left\langle f,\Xi_s^N\right\rangle^2\big]+\mathbb{E}\big[\left\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_s^N-\bar{\mu}_s)\right\rangle^2\big](\sup_{c,w\in K}|\partial_c f|)^2\pi(dx,dy)ds\\
\leq\;& C\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\left(\mathbb{E}\big[\left\langle f,\Xi_s^N\right\rangle^2\big]+\left\|f\right\|_L^2\right)\pi(dx,dy)ds.
\end{aligned}$$

Hence, it remains to study the first, second, third and sixth term in (4.9). To do so, we first state the following lemma.

Lemma 4.2. *There is a compact set $K=[-C_o,C_o]^{1+d}\subset\mathbb{R}^{1+d}$ such that η_t^N, Ξ_t^N, and μ_t^N vanish when evaluated on any $A\subset K^c$.*

Proof. Due to the uniform bound (3.2), there exists a compact set $K\subset\mathbb{R}^{1+d}$ such that $\mu_t^N(K^c)=\tilde{\mu}_t^N(K^c)=\bar{\mu}_t(K^c)=0$. It directly follows from the definitions of η_t^N and Ξ_t^N that they also vanish outside of the set K. For example, for $A\in K^c$, $\eta_t^N(A)=\sqrt{N}\big(\mu_t^N(A)-\bar{\mu}_t(A)\big)=0$. □

Due to Lemma 4.2, there is a C_c^∞ "bump" function $b(c,w)$ such that $b(c,w)c\sigma'(wx)$ is in $C_c^\infty(\mathbb{R}^{1+d}\times\mathcal{X})$ and $b(c,w)c\sigma'(wx)=c\sigma'(wx)$ for every $(c,w)\in K$, the compact set defined in Lemma 4.2, and $x\in\mathcal{X}$. Similar statements hold for the terms $\sigma(wx)$ and $c\sigma(wx)$. See [13] for a discussion of bump functions. An example of a bump function is:

$$\begin{aligned}
b(z) &= \frac{h\big(2-\frac{\|z\|}{r}\big)}{h\big(\frac{\|z\|}{r}-1\big)+h\big(2-\frac{\|z\|}{r}\big)},\\
h(v) &= e^{-\frac{1}{v^2}}\mathbf{1}_{v>0}. \qquad (4.11)
\end{aligned}$$

The function $b(z)$ is $C_c^\infty(\mathbb{R}^{1+d})$, vanishes for $\|z\|\geq 2r$, and is one on $\|z\|\leq r$ [13]. For the purposes of this paper, we may choose $r=\sqrt{D}C_o$, $B=3\sqrt{D}C_o$, and $\Theta=(-B,B)^D$. In particular, notice for instance that $b(c,w)c\sigma'(wx)$, and its partial derivatives, vanish on the boundary of Θ.

Going back to (4.9), the aforementioned discussion implies that we can write for example

$$\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f,\Xi^N_s\right\rangle=\left\langle b(c,w)c\sigma'(w\cdot x)x\cdot\nabla_w f,\Xi^N_s\right\rangle$$
$$\left\langle \sigma(w\cdot x)\partial_c f,\Xi^N_s\right\rangle=\left\langle b(c,w)\sigma(w\cdot x)\partial_c f,\Xi^N_s\right\rangle$$

Hence, let us define the operators

$$\begin{aligned}\mathcal{G}_1 f &= b(c,w)c\sigma'(wx)x\cdot\nabla_w f,\\ \mathcal{G}_2 f &= b(c,w)\sigma(wx)\partial_c f.\end{aligned} \tag{4.12}$$

Let $\{f_a\}_{a=1}^\infty$ be a complete orthonormal basis for $W_0^{J_1,2}(\Theta)$. Since $J_1-L>\frac{D}{2}$, the embedding $W_0^{J_1,2}(\Theta)\hookrightarrow W_0^{L,2}(\Theta)$ is of Hilbert-Schmidt type and

$$\sum_a \|f_a\|_L^2<\infty. \tag{4.13}$$

(See Theorem 6.53 of [1] for details.)

Let $f=f_a$ in (4.9) and sum over all $a\geq 1$. Using Parseval's identity, we now have the bound

$$\begin{aligned}\mathbb{E}\Big[\left\|\Xi^N_t\right\|^2_{-J_1}\Big] &\leq C_1\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathbb{E}\Big[\left\|\Xi^N_s\right\|^2_{-J_1}+\left(y-\left\langle c\sigma(w\cdot x),\mu^N_s\right\rangle\right)\left\langle\Xi^N_s,\mathcal{G}_1^*\Xi^N_s\right\rangle_{-J_1}\\ &+\left(y-\left\langle c\sigma(w\cdot x),\mu^N_s\right\rangle\right)\left\langle\Xi^N_s,\mathcal{G}_2^*\Xi^N_s\right\rangle_{-J_1}\Big]\pi(dx,dy)ds+C_2.\end{aligned} \tag{4.14}$$

Since μ^N_t takes values in a compact set and $\mathcal{X}\times\mathcal{Y}$ is compact, we have that:

$$\begin{aligned}\mathbb{E}\Big[\left\|\Xi^N_t\right\|^2_{-J_1}\Big] &\leq C_1\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathbb{E}\Big[\left\|\Xi^N_s\right\|^2_{-J_1}+|\left\langle\Xi^N_s,\mathcal{G}_1^*\Xi^N_s\right\rangle_{-J_1}|\\ &+|\left\langle\Xi^N_s,\mathcal{G}_2^*\Xi^N_s\right\rangle_{-J_1}|\Big]\pi(dx,dy)ds+C_2.\end{aligned} \tag{4.15}$$

The terms $|\left\langle\Xi^N_s,\mathcal{G}_1^*\Xi^N_s\right\rangle_{-J_1}|$ and $|\left\langle\Xi^N_s,\mathcal{G}_2^*\Xi^N_s\right\rangle_{-J_1}|$ must now be analyzed. By the Riesz representation theorem for Hilbert spaces, for $\Xi\in W^{-J_1,2}$ there exists a unique $\Psi=F(\Xi)\in W_0^{J_1,2}$ such that,

$$\langle f,\Xi\rangle = \langle f,\Psi\rangle_{J_1},\ \text{for } f\in W_0^{J_1,2}.$$

Lemma 4.3. *For $\Xi\in W^{-J,2}$ with $J\geq J_1=2\lceil\frac{D}{2}\rceil+4$, we have*

$$\begin{aligned}\langle\Xi,\mathcal{G}_1^*\Xi\rangle_{-J} &\leq K\|\Xi\|^2_{-J},\\ \langle\Xi,\mathcal{G}_2^*\Xi\rangle_{-J} &\leq K\|\Xi\|^2_{-J}.\end{aligned}$$

Proof. Notice that $\{\Xi\in W^{-J,2}:F(\Xi)\in W_0^{J+1,2}\}$ is dense in $W^{-J,2}$. For $\Xi\in W^{-J,2}$ such that $\Psi=F(\Xi)\in W_0^{J+1,2}$ we have by definition

$$\langle\Xi,\mathcal{G}^*\Xi\rangle_{-J}=\langle\Psi,\mathcal{G}^*\Xi\rangle=\langle\mathcal{G}\Psi,\Xi\rangle=\langle\mathcal{G}\Psi,\Psi\rangle_J,$$

since $\mathcal{G}\Psi\in W_0^{J,2}$ for either $\mathcal{G}=\mathcal{G}_1$ or $\mathcal{G}=\mathcal{G}_2$.

By setting $g(c,w)=b(c,w)c\sigma'(wx)x$ in Lemma A.1 and $\mathcal{X}$ being a compact set,

$$\langle\mathcal{G}_1\Psi,\Psi\rangle_J \leq C\|\Psi\|^2_J=C\|\Xi\|^2_{-J}.$$

Similarly, setting $g(c,w)=b(c,w)\sigma(wx)$ in Lemma A.1 and since $\mathcal{X}$ is a compact set,

$$\langle\mathcal{G}_2\Psi,\Psi\rangle_J \leq C\|\Psi\|^2_J=C\|\Xi\|^2_{-J},$$

completing the proof of the lemma. □

Lemma 4.3 and equation (4.15) produce the bound

$$\mathbb{E}\left\|\Xi_t^N\right\|_{-J_1} \quad \leq \quad C_1 \int_0^t \mathbb{E}\bigg(\left\|\Xi_s^N\right\|_{-J_1}^2 \bigg) ds + C_2.$$

Note that we have again used the fact that $\mathcal{X} \times \mathcal{Y}$ is a compact set.

By Gronwall's Lemma,

$$\sup_{N\in\mathbb{N}} \sup_{t\in[0,T]} \mathbb{E}\left\|\Xi_t^N\right\|_{-J_1}^2 < C. \tag{4.16}$$

Recall that

$$\left\langle f, \eta_t^N \right\rangle = \left\langle f, \Xi_t^N \right\rangle + \left\langle f, \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t) \right\rangle.$$

Therefore,

$$\left\langle f, \eta_t^N \right\rangle^2 \leq 2\left\langle f, \Xi_t^N \right\rangle^2 + 2\left\langle f, \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t) \right\rangle^2. \tag{4.17}$$

The second term is a sequence of i.i.d. random variables. That is,

$$\begin{aligned}
\mathbb{E}\left\langle f, \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t) \right\rangle^2 \quad &= \quad \mathbb{E}\bigg[\bigg(\frac{1}{N}\sum_{i=1}^N [f(\tilde{c}_t^i, \tilde{w}_t^i) - \langle f, \bar{\mu}_t \rangle]\bigg)^2\bigg] \\
&\leq \quad \mathbb{E}\bigg[\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N [f(\tilde{c}_t^i, \tilde{w}_t^i) - \langle f, \bar{\mu}_t \rangle][f(\tilde{c}_t^j, \tilde{w}_t^j) - \langle f, \bar{\mu}_t \rangle]\bigg] \\
&= \quad \frac{1}{N}\big(\langle f^2, \bar{\mu}_t \rangle - \langle f, \bar{\mu}_t \rangle^2 \big) \\
&\leq \quad C\left\|f\right\|_L^2,
\end{aligned} \tag{4.18}$$

where the last inequality follows from the compact support of $\bar{\mu}$ and the bound (4.5).

Now, we are in position to complete the proof of Lemma 4.1.

Proof of Lemma 4.1. Let $f = f_a$ in (4.17) and sum over all $a \geq 1$. The lemma follows from Parseval's identity, (4.13), (4.16), and (4.18). ☐

4.2 Regularity of η^N

Let $\{f_a\}_{a=1}^\infty$ be a complete orthonormal basis for $W_0^{J_2,2}(\Theta)$ for $J_2 = 3\lceil \frac{D}{2} \rceil + 6$. Recall that η_t^N can be written via the decomposition

$$\eta_t^N = \sqrt{N}(\mu_t^N - \tilde{\mu}_t^N) + \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t). \tag{4.19}$$

Let $\Xi_t^N = \sqrt{N}(\mu_t^N - \tilde{\mu}_t^N)$ and $Z_t^N = \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t)$. For $0 \leq r < t < T$,

$$\big(\left\langle f_a, \eta_t^N \right\rangle - \left\langle f_a, \eta_r^N \right\rangle \big)^2 \leq 2\big(\left\langle f_a, \Xi_t^N \right\rangle - \left\langle f_a, \Xi_r^N \right\rangle \big)^2 + 2\big(\left\langle f_a, Z_t^N \right\rangle - \left\langle f_a, Z_r^N \right\rangle \big)^2.$$

Using Young's inequality, the Cauchy-Schwarz inequality, and the fact that $[0,T] \times \mathcal{X} \times \mathcal{Y}$ is a compact set,

$$
\begin{aligned}
\big(\left\langle f_a,\Xi_t^N\right\rangle-\left\langle f_a,\Xi_r^N\right\rangle\big)^2 \le C\Bigg[&\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(y\left\langle\sigma(w\cdot x)\partial_c f_a,\Xi_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(y\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f_a,\Xi_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\left\langle c\sigma(w\cdot x),\mu_s^N\right\rangle\left\langle\sigma(w\cdot x)\partial_c f_a,\Xi_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\left\langle c\sigma(w\cdot x),\Xi_s^N\right\rangle\left\langle\sigma(w\cdot x)\partial_c f_a,\tilde{\mu}_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\Big\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_s^N-\bar{\mu}_s)\Big\rangle\left\langle\sigma(w\cdot x)\partial_c f_a,\tilde{\mu}_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\left\langle c\sigma(w\cdot x),\mu_s^N\right\rangle\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f_a,\Xi_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\left\langle c\sigma(w\cdot x),\Xi_s^N\right\rangle\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f_a,\tilde{\mu}_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\Big\langle c\sigma(w\cdot x),\sqrt{N}(\tilde{\mu}_s^N-\bar{\mu}_s)\Big\rangle\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f_a,\tilde{\mu}_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&+\big(\left\langle f_a,M_t^N\right\rangle-\left\langle f_a,M_r\right\rangle\big)^2+(R_t^{1,N}-R_r^{1,N})^2+(R_t^{2,N}-R_r^{2,N})^2\Bigg].
\end{aligned}
\tag{4.20}
$$

Recall that $J_1=2\lceil\frac{D}{2}\rceil+4$. Since $\pi(dx,dy)$ has compact support, we have

$$
\begin{aligned}
\Big|\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f_a,\Xi_s^N\right\rangle\Big| &\le \left\|c\sigma'(w\cdot x)x\cdot\nabla_w f_a\right\|_{J_1}\left\|\Xi_s^N\right\|_{-J_1}\\
&\le C\left\|f_a\right\|_{J_1+1}\left\|\Xi_s^N\right\|_{-J_1},
\end{aligned}
\tag{4.21}
$$

and, we have by (4.16) that $\sup_{n\in\mathbb{N}}\sup_{t\in[0,T]}\mathbb{E}\left\|\Xi_t^N\right\|_{-J_1}^2<C$. Recall also that $R_t^{1,N}=N^{-3/2}\sum_{k=0}^{\lfloor Nt\rfloor-1}G_k^N$ where $|G_k^N|\le\|f\|_L$ where $L=\lceil\frac{D}{2}\rceil+3$. Therefore,

$$
\begin{aligned}
(R_t^{1,N}-R_r^{1,N})^2 &= \left(N^{-3/2}\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}G_k^N\right)^2\\
&\le \frac{1}{N}(t-r)^2\left\|f_a\right\|_L^2+\frac{1}{N^3}\left\|f_a\right\|_L^2.
\end{aligned}
\tag{4.22}
$$

In addition,

$$
\begin{aligned}
(R_t^{2,N}-R_r^{2,N})^2 &\le 2\big(\sqrt{N}V_t^N\big)^2+2\big(\sqrt{N}V_r^N\big)^2\\
&\le \frac{C}{N}\left\|f_a\right\|_L^2.
\end{aligned}
\tag{4.23}
$$

Therefore, using (4.20), (4.21), (4.22), (4.23), (4.18), the fact that $\mathcal{X}\times\mathcal{Y}$ is a compact set, and the compact containment of μ_t^N and $\tilde{\mu}_t^N$,

$$
\begin{aligned}
\mathbb{E}\Big[\big(\left\langle f_a,\Xi_t^N\right\rangle-\left\langle f_a,\Xi_r^N\right\rangle\big)^2\Big] &\le C\Bigg[\int_r^t\Big(\left\|f_a\right\|_L^2\mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2\Big]+\left\|f_a\right\|_{J_1+1}^2\mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2\Big]\Big)ds\\
&+ \left\|f_a\right\|_L^2(t-r)+\frac{1}{N}(t-r)^2\left\|f_a\right\|_L^2\Bigg]+C_2\big(\frac{1}{N}+\frac{1}{N^3}\big)\left\|f_a\right\|_L^2.
\end{aligned}
\tag{4.24}
$$

Since we have chosen $J_2 = 3\lceil \frac{D}{2} \rceil + 6$, we certainly have that $J_2 > J_1 + 1 + \frac{D}{2}$, which then implies that $\sum_{a\geq 1} \|f_a\|_{J_1+1}^2 < \infty$ and $\sum_{a\geq 1} \|f_a\|_L^2 < \infty$. Hence, using the uniform bound (4.16) and Parseval's identity, we obtain

$$\mathbb{E}\bigg[\|\Xi_t^N - \Xi_r^N\|_{-J_2}^2 \bigg] \leq C_1(t-r) + C_2 \frac{1}{N}.$$

Using a similar approach, we can show that

$$\mathbb{E}\bigg[\|Z_t^N - Z_r^N\|_{-J_2}^2 \bigg] \leq C(t-r).$$

Lemma 4.4. *Let $J_2 = 3\lceil \frac{D}{2} \rceil + 6$. If $(t-r) < \delta$, then there are unimportant constants $C_1, C_2 < \infty$ such that*

$$\mathbb{E}\bigg[\|\eta_t^N - \eta_r^N\|_{-J_2}^2 \bigg] \leq C_1 \delta + C_2 \frac{1}{N}. \tag{4.25}$$

Then, the regularity condition B of Theorem 8.6 of Chapter 3 of [10] is satisfied. (See also Remark 8.7 B of Chapter of [10] regarding replacing $\sup_N$ *with* $\lim_N$*.) In particular,*

$$\lim_{\delta\to 0} \lim_{N\to\infty} \left(C_1 \delta + C_2 \frac{1}{N} \right) = 0.$$

4.3 Compact containment of $\sqrt{N} M^N$

Let $\mathfrak{F}_t$ be the $\sigma-$algebra generated by μ_s^N and M_s^N for $s \leq t$ and note that $\left\langle f, \sqrt{N} M_t^N \right\rangle$ is a martingale since

$$\begin{aligned}
\mathbb{E}\bigg[\left\langle f, \sqrt{N} M_t^N \right\rangle \bigg| \mathfrak{F}_r \bigg] &= \mathbb{E}\bigg[\left\langle f, \sqrt{N} M_t^N - \sqrt{N} M_r^N \right\rangle \bigg| \mathfrak{F}_r \bigg] + \mathbb{E}\bigg[\left\langle f, \sqrt{N} M_r^N \right\rangle \bigg| \mathfrak{F}_r \bigg] \\
&= \sum_{k=\lfloor Nr \rfloor}^{\lfloor Nt \rfloor - 1} \mathbb{E}\bigg[\bigg(\left\langle f, M_k^{1,N} \right\rangle + \left\langle f, M_k^{2,N} \right\rangle \bigg| \mathcal{F}_{\lfloor Nr \rfloor}^N \bigg] + \left\langle f, \sqrt{N} M_r^N \right\rangle \\
&= \left\langle f, \sqrt{N} M_r^N \right\rangle.
\end{aligned}$$

By Doob's martingale inequality,

$$\mathbb{E}\bigg[\sup_{t\in[0,T]} \left\langle f, \sqrt{N} M_t^N \right\rangle^2 \bigg] \leq C \mathbb{E}\bigg[\left\langle f, \sqrt{N} M_T^N \right\rangle^2 \bigg].$$

This of course implies that

$$\mathbb{E}\bigg[\sup_{t\in[0,T]} \left\langle f, \sqrt{N} M_t^N \right\rangle^2 \bigg] \quad \leq \quad C \|f\|_L^2, \tag{4.26}$$

where the last inequality is proven using the same approach as in equation (4.8) using also Lemma 3.1 of [25]. Here, we recall that $L = \lceil \frac{D}{2} \rceil + 3$.

Lemma 4.5. *If $J_1 = 2\lceil \frac{D}{2} \rceil + 4$, then there is a constant $C < \infty$ such that*

$$\sup_{N\in\mathbb{N}} \mathbb{E}\bigg[\sup_{t\in[0,T]} \left\| \sqrt{N} M_t^N \right\|_{-J_1}^2 \bigg] \leq C. \tag{4.27}$$

Proof. Let $\{f_a\}_{a\geq 1}$ be a complete orthonormal basis for $W_0^{J_1,2}(\Theta)$. Using equation (4.26), Parseval's identity and (4.13) we get

$$\begin{aligned}
\sup_{N\in\mathbb{N}} \mathbb{E}\Bigg[\sup_{t\in[0,T]} \left\|\sqrt{N}M_t^N\right\|_{-J_1}^2\Bigg] &= \sup_{N\in\mathbb{N}} \mathbb{E}\Bigg[\sup_{t\in[0,T]} \sum_{a\geq 1}\left\langle f_a, \sqrt{N}M_t^N\right\rangle^2\Bigg] \\
&\leq \sup_{N\in\mathbb{N}} \sum_{a\geq 1} \|f_a\|_L^2 < \infty,
\end{aligned}$$

completing the proof of the lemma. □

4.4 Regularity of $\sqrt{N}M^N$

Let $\{f_a\}_{a=1}^{\infty}$ be a complete orthonormal basis for $W_0^{J_1,2}$ with $J_1 = 2\lceil\frac{D}{2}\rceil + 4$. For $0 \leq r < t < T$. The equation below is the sum of jump terms at discrete times $\frac{1}{N}, \frac{2}{N}, \ldots, \frac{T}{N}$. By Lemma 3.1 of [25] (see also Theorem 3.2 of [2]) we get

$$\begin{aligned}
&\mathbb{E}\Bigg[\Big(\left\langle f_a, \sqrt{N}M_t^N\right\rangle - \left\langle f_a, \sqrt{N}M_r^N\right\rangle\Big)^2\Bigg|\mathfrak{F}_r\Bigg] \\
= \;& N\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbb{E}\Bigg[\Big(\left\langle f_a, M_k^{1,N}\right\rangle + \left\langle f_a, M_k^{2,N}\right\rangle\Big)^2\Bigg|\mathcal{F}_{\lfloor Nr\rfloor}^N\Bigg] \\
\leq \;& \frac{C}{N}\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \|f_a\|_L^2 \\
\leq \;& C_1\|f_a\|_L^2(t-s) + \frac{C_2}{N}\|f_a\|_L^2,
\end{aligned}$$

where $L = \lceil\frac{D}{2}\rceil + 3$ and the third line is derived using the same approach as in equation (4.8). By Parseval's identity and (4.13), for any $0 < p < 1$,

$$\begin{aligned}
\mathbb{E}\Bigg[\left\|\sqrt{N}M_t^N - \sqrt{N}M_r^N\right\|_{-J_1}^2\Bigg|\mathfrak{F}_r\Bigg] &\leq C_1(t-r) + \frac{C_2}{N}. \\
&\leq C_1(t-r)^p\mathbf{1}_{t-r<1} + C_1(t-r)^pT^{1/p}\mathbf{1}_{t-r\geq 1} + \frac{C_2}{N} \\
&\leq C_1(t-r)^p + \frac{C_2}{N}. \qquad (4.28)
\end{aligned}$$

Lemma 4.6. *Let $J_1 = 2\lceil\frac{D}{2}\rceil + 4$. If $(t-r) < \delta$, then there are unimportant constants $C_1, C_2 < \infty$ such that for any $0 < p < 1$,*

$$\mathbb{E}\Bigg[\left\|\sqrt{N}M_t^N - \sqrt{N}M_r^N\right\|_{-J_1}^2\Bigg|\mathfrak{F}_r\Bigg] \leq C_1\delta^p + \frac{C_2}{N}.$$

Then, the regularity condition B of Theorem 8.6 of Chapter 3 of [10] is satisfied. (See also Remark 8.7 B of Chapter 3 of [10] regarding replacing $\sup_N$ with $\lim_N$.) In particular,

$$\lim_{\delta\to 0}\lim_{N\to\infty}\left(C_1\delta^p + \frac{C_2}{N}\right) = 0.$$

4.5 Compact containment of the fluctuations process η^N

The main result of this section is Lemma 4.7 below.

Lemma 4.7. *If $J_2 = 3\lceil \frac{D}{2} \rceil + 6$, then there is a constant $C < \infty$ such that*

$$\sup_{N\in\mathbb{N}} \mathbb{E} \sup_{t\in[0,T]} \left\|\eta_t^N\right\|_{-J_2}^2 < C. \tag{4.29}$$

In particular, the process $\{\eta_\cdot^N\}_{N\in\mathbb{N}}$ satisfies the compact containment condition in $W^{-J,2}(\Theta)$ with $J \geq J_2 + 1 = 3\lceil \frac{D}{2} \rceil + 7$.

Proof. The proof of this statement follows by the representation (4.17) together with the a-priori bounds of Lemma 4.1 and 4.5.

Let $\{f_a\}_{a=1}^{\infty}$ be a complete orthonormal basis for $W_0^{J_2,2}$ with $J_2 = 3\lceil \frac{D}{2} \rceil + 6$. Equation (4.17) with $f = f_a$ gives

$$\mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left\langle f_a, \eta_t^N \right\rangle^2 \quad \leq \quad 2\mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left\langle f_a, \Xi_t^N \right\rangle^2 + 2\mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left\langle f_a, \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t) \right\rangle^2.$$

Following the arguments in equations (4.20)-(4.24) with $r = 0$ and using Lemma 4.5 gives

$$\begin{aligned}
\mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left[\left(\left\langle f_a, \Xi_t^N \right\rangle - \left\langle f_a, \Xi_0^N \right\rangle \right)^2 \right] &\leq C\bigg[\int_0^T \bigg(\sum_{a\geq 1} \|f_a\|_L^2 \, \mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2 \Big] + \sum_{a\geq 1} \|f_a\|_{J_1+1}^2 \, \mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2 \Big] \bigg) ds \\
&\quad + \sum_{a\geq 1} \|f_a\|_L^2 + N^{-1} \sum_{a\geq 1} \|f_a\|_L^2 \bigg] \\
&\leq C\bigg[\sup_{s\in[0,T]} \mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2 \Big] \sum_{a\geq 1} \|f_a\|_L^2 + \sup_{s\in[0,T]} \mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2 \Big] \sum_{a\geq 1} \|f_a\|_{J_1+1}^2 \\
&\quad + \sum_{a\geq 1} \|f_a\|_L^2 + N^{-1} \sum_{a\geq 1} \|f_a\|_L^2 \bigg]
\end{aligned}$$

Similarly, using now (4.18), we have

$$\mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left\langle f_a, \sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t) \right\rangle^2 \leq C \sum_{a\geq 1} \|f_a\|_L^2$$

Putting the last displays together we obtain

$$\begin{aligned}
\mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left\langle f_a, \eta_t^N \right\rangle^2 &\leq C\bigg[\sup_{s\in[0,T]} \mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2 \Big] \sum_{a\geq 1} \|f_a\|_L^2 + \sup_{s\in[0,T]} \mathbb{E}\Big[\left\|\Xi_s^N\right\|_{-J_1}^2 \Big] \sum_{a\geq 1} \|f_a\|_{J_1+1}^2 \\
&\quad + \sum_{a\geq 1} \|f_a\|_L^2 + N^{-1} \sum_{a\geq 1} \|f_a\|_L^2 + \mathbb{E}\Big[\left\|\Xi_0^N\right\|_{-J_2}^2 \Big] \bigg].
\end{aligned}$$

By Lemma 4.1 we have that $\sup_{N\in\mathbb{N}} \sup_{t\in[0,T]} \mathbb{E} \left\|\eta_t^N\right\|_{-J_1}^2 \leq C$. Since, $J_2 > J_1 + 1 + \frac{D}{2} > L + \frac{D}{2}$, we also obtain (by Sobolev embedding as before) that $\sum_{a\geq 1} \|f_a\|_{J_1+1}^2 < \infty$ and $\sum_{a\geq 1} \|f_a\|_L^2 < \infty$. In addition, since $J_2 > J_1$ we have that $\|\cdot\|_{-J_2} \leq C \|\cdot\|_{-J_1}$ which then, due to (4.16), leads to $\sup_{N\in\mathbb{N}} \mathbb{E}\Big[\left\|\Xi_0^N\right\|_{-J_2}^2 \Big] < \infty$. Hence, we indeed have that

$$\sup_{N\in\mathbb{N}} \mathbb{E}\sum_{a\geq 1} \sup_{t\in[0,T]} \left\langle f_a, \eta_t^N \right\rangle^2 \leq C.$$

Hence, by Parseval's identity we obtain

$$\sup_{N\in\mathbb{N}} \mathbb{E} \sup_{t\in[0,T]} \left\|\eta_t^N\right\|_{-J_2}^2 = \sup_{N\in\mathbb{N}} \mathbb{E} \sup_{t\in[0,T]} \sum_{a\geq 1} \left\langle f_a, \eta_t^N \right\rangle^2 \leq C.$$

Now, due to the bound in the last display, we obtain that for every $\epsilon > 0$, there is a constant C_ϵ such that

$$\sup_{N\in\mathbb{N}} \mathbb{P}\left\{ \sup_{t\in[0,T]} \left\|\eta_t^N\right\|_{-J_2}^2 > C_\epsilon \right\} \le \epsilon,$$

and, due to the fact that the set $\left\{\phi \in W^{-(J_2+1),2} : \|\phi\|_{-J_2} \le C_\epsilon\right\}$ is a compact subset of $W^{-(J_2+1),2}$, we obtain the validity of the compact containment condition for $\{\eta_\cdot^N\}_{N\in\mathbb{N}}$ in $W^{-J,2}$ with $J \ge J_2 + 1$, as desired. ◻

4.6 Relative Compactness of η^N and $\sqrt{N}M^N$

Lemma 4.8. *Let $T > 0$ and $J \ge 3\lceil \frac{D}{2}\rceil + 7$. Then, the sequences $\{\mu_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$, $\{\eta_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$ and $\{\sqrt{N}M_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$ are relatively compact in $D_{\mathcal{M}(\mathbb{R}^{1+d})}[0,T]$, $D_{W^{-J,2}}([0,T])$ and $D_{W^{-J,2}}([0,T])$ respectively.*

Proof. Relative compactness of μ^N was proven in [25]. Lemmas 4.4, 4.7 for η^N and Lemmas 4.5, 4.6 for $\sqrt{N}M^N$ combined with Theorem 8.6 of Chapter 3 of [10] (and using Remark 8.7 B of [10]) prove the result. ◻

5 Continuity properties and identification of the limiting equation

Lemma 5.1. *Let $J \ge 3\lceil \frac{D}{2}\rceil + 7$. Any limit point of $\{\eta_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$ is continuous, i.e., it takes values in $C_{W^{-J,2}}([0,T])$.*

Proof. In order to prove that any limit point of $\{\eta_t^N, t \in [0,T]\}_{N\in\mathbb{N}}$ takes values in $C_{W^{-J,2}}([0,T])$, it is sufficient to show that

$$\lim_{N\to\infty} \mathbb{E}\left[\sup_{t\le T} \left\|\eta_t^N - \eta_{t-}^N\right\|_{-J}^2\right] = 0.$$

We again use the decomposition (4.19),

$$\sup_{t\le T} \left\|\eta_t^N - \eta_{t-}^N\right\|_{-J}^2 \le 2\sup_{t\le T} \left\|\Xi_t^N - \Xi_{t-}^N\right\|_{-J}^2 + 2\sup_{t\le T} \left\|\sqrt{N}(\tilde{\mu}_t^N - \bar{\mu}_t) - \sqrt{N}(\tilde{\mu}_{t-}^N - \bar{\mu}_{t-})\right\|_{-J}^2.$$

Since both $\tilde{\mu}_t^N$ and $\bar{\mu}_t$ are continuous, $\left\|(\tilde{\mu}_t^N - \bar{\mu}_t) - (\tilde{\mu}_{t-}^N - \bar{\mu}_{t-})\right\|_{-J} = 0$.

Next, let $\{f_a\}_{a=1}^\infty$ be a complete orthonormal basis for $W_0^{J,2}$. As it follows by (4.3) the discontinuities of $\left\langle f_a, \Xi_t^N\right\rangle$ are those of $\sqrt{N}\left\langle f_a, M_t^N\right\rangle$ and $R_t^{1,N} + R_t^{2,N}$. Hence, we shall have,

$$\left\langle f_a, \Xi_t^N\right\rangle - \left\langle f_a, \Xi_{t-}^N\right\rangle = \sqrt{N}\left\langle f_a, M_t^N\right\rangle - \sqrt{N}\left\langle f_a, M_{t-}^N\right\rangle + R_t^N - R_{t-}^N,$$

where $R_t^N = R_t^{1,N} + R_t^{2,N}$.

Note that $\left\langle f_a, M_t^N\right\rangle$ is a pure jump process where the size of the k-th jump is bounded by

$$\begin{aligned}
&\left|\frac{1}{N}\alpha\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle \sigma(w\cdot x_k)\partial_c f_a, \nu_k^N\right\rangle - D_k^{1,N}\right| \\
+ &\left|\frac{1}{N}\alpha\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle c\sigma'(w\cdot x_k)x\cdot\nabla_w f_a, \nu_k^N\right\rangle - D_k^{2,N}\right|.
\end{aligned} \tag{5.1}$$

Therefore, for $0 \le t \le T$,

$$\begin{aligned}
&\left(\sqrt{N}\left\langle f_a, M_t^N\right\rangle - \sqrt{N}\left\langle f_a, M_{t-}^N\right\rangle\right)^2 \le \\
&\quad\le 2N \sup_{0\le k\le\lfloor Nt\rfloor-1} \left(\frac{1}{N}\alpha\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle \sigma(w\cdot x_k)\partial_c f_a, \nu_k^N\right\rangle - D_k^{1,N}\right)^2 \\
&\quad+ 2N \sup_{0\le k\le\lfloor Nt\rfloor-1} \left(\frac{1}{N}\alpha\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle c\sigma'(w\cdot x_k)x\cdot\nabla_w f_a, \nu_k^N\right\rangle - D_k^{2,N}\right)^2.
\end{aligned}$$

Due to the uniform bound (3.2), the bound (4.5), and $\pi(dx,dy)$ having compact support,

$$
\begin{aligned}
\left| \left\langle f_a, \sqrt{N} M_t^N \right\rangle - \left\langle f_a, \sqrt{N} M_{t^-}^N \right\rangle \right|^2 &\leq \frac{C}{N}\Bigg(\sum_{|\alpha|=1} \sup_{(c,w)\in K} |D^\alpha f_a(c,w)| \Bigg)^2 \\
&\leq \frac{C}{N} \left\| f_a \right\|_L^2 .
\end{aligned}
$$

Similarly,

$$
\left(R_t^N - R_{t^-}^N \right)^2 \leq \frac{C}{N} \left\| f_a \right\|_L^2 .
$$

Therefore, for $0 \leq t \leq T$,

$$
\left\langle f_a, \Xi_t^N - \Xi_{t^-}^N \right\rangle^2 \leq \frac{C}{N} \left\| f_a \right\|_L^2 .
$$

Since $J - L > D/2$, the embedding $W_0^{J,2}(\Theta) \hookrightarrow W_0^L(\Theta)$ is of Hilbert-Schmidt type (Theorem 6.53 of [1]) and we have the bound $\sum_a \left\| f_a \right\|_L^2 < \infty$. Hence, we obtain

$$
\mathbb{E}\left[\sup_{t\leq T} \left\| \Xi_t^N - \Xi_{t^-}^N \right\|_{-J}^2 \right] \leq \frac{C}{N}.
$$

Consequently, $\lim_{N\to\infty} \mathbb{E}\left[\sup_{t\leq T} \left\| \eta_t^N - \eta_{t^-}^N \right\|_{-J}^2 \right] = 0$, concluding the proof of the lemma. ☐

Lemma 5.2. *Let $J_1 = 2\lceil \frac{D}{2} \rceil + 4$ and for $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $\mu \in \mathcal{M}(\mathbb{R}^{1+d})$ and $h \in \mathcal{C}_0^1(\mathbb{R}^{1+d})$ define the operator*

$$
\mathcal{R}_{x,y,\mu}[h] = \left(y - \langle c\sigma(w\cdot x), \mu \rangle \right) \langle \nabla (c\sigma(w\cdot x)) \cdot \nabla h, \mu \rangle .
$$

Then, for every $f \in W_0^{J_1,2}(\Theta)$, $\sqrt{N} \langle f, M_t^N \rangle \in D_{\mathbb{R}}([0,T])$ converges in distribution to a distribution valued mean-zero Gaussian martingale $\bar{M}_t$ with variance

$$
Var\Big[\left\langle f, \bar{M}_t \right\rangle \Big] = \alpha^2 \int_0^t \left[\int_{\mathcal{X}\times\mathcal{Y}} \left(\mathcal{R}_{x,y,\bar{\mu}_s}[f] - \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{R}_{x,y,\bar{\mu}_s}[f] \pi(dx,dy) \right)^2 \pi(dx,dy) \right] ds
$$

More generally, for every $f, g \in W_0^{J_1,2}(\Theta)$, $(\sqrt{N}\langle f, M_t^N \rangle, \sqrt{N}\langle g, M_t^N \rangle) \in D_{\mathbb{R}^2}([0,T])$ converges to a distribution valued mean-zero Gaussian martingale with covariance function

$$
\begin{aligned}
Cov\Big[\left\langle f, \bar{M}_t \right\rangle, \left\langle g, \bar{M}_t \right\rangle \Big] &= \alpha^2 \int_0^t \Bigg[\int_{\mathcal{X}\times\mathcal{Y}} \left(\mathcal{R}_{x,y,\bar{\mu}_s}[f] - \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{R}_{x,y,\bar{\mu}_s}[f] \pi(dx,dy) \right) \times \\
&\quad \times \left(\mathcal{R}_{x,y,\bar{\mu}_s}[g] - \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{R}_{x,y,\bar{\mu}_s}[g] \pi(dx,dy) \right) \pi(dx,dy) \Bigg] ds.
\end{aligned}
$$

Proof. Recall that

$$
\begin{aligned}
\sqrt{N} \left\langle f, M_t^N \right\rangle &= N^{1/2} \sum_{k=0}^{\lfloor Nt \rfloor -1} \left(\frac{\alpha}{N} \big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N \right\rangle \big) \left\langle \sigma(w\cdot x_k) \partial_c f, \nu_k^N \right\rangle - D_k^{1,N} \right) \\
&\quad + N^{1/2} \sum_{k=0}^{\lfloor Nt \rfloor -1} \left(\frac{\alpha}{N} \big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N \right\rangle \big) \left\langle c\sigma'(w\cdot x_k) x \cdot \nabla_w f, \nu_k^N \right\rangle - D_k^{2,N} \right) \\
&= \sum_{k=0}^{\lfloor Nt \rfloor -1} X_k^N ,
\end{aligned}
$$

where

$$
\begin{aligned}
X_k^N \quad &:= \quad N^{1/2}\left(\frac{\alpha}{N}\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle \sigma(w\cdot x_k)\partial_c f, \nu_k^N\right\rangle - D_k^{1,N}\right)\\
&+ \quad N^{1/2}\left(\frac{\alpha}{N}\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle c\sigma'(w\cdot x_k)x\cdot\nabla_w f, \nu_k^N\right\rangle - D_k^{2,N}\right) =\\
&= \quad \frac{\alpha}{\sqrt{N}}\Big[\big(y_k - \left\langle c\sigma(w\cdot x_k), \nu_k^N\right\rangle\big)\left\langle \nabla(c\sigma(w\cdot x_k))\cdot\nabla f, \nu_k^N\right\rangle\\
&\quad - \left(\int_{\mathcal{X}\times\mathcal{Y}}(y - \left\langle c\sigma(w\cdot x), \nu_k^N\right\rangle)\left\langle \nabla(c\sigma(w\cdot x))\cdot\nabla f, \nu_k^N\right\rangle\pi(dx,dy)\right)\Big].
\end{aligned}
$$

Due to the compact support of $\pi(dx,dy)$ and the uniform bound $|c^i| + \left\|w^i\right\| < C_o$, $|X_k^N| \leq CN^{-1/2}$. $\sqrt{N}\left\langle f, M_t^N\right\rangle$ is a pure jump process and its quadratic variation is

$$
\begin{aligned}
\left[\sqrt{N}\left\langle f, M_t^N\right\rangle\right] \quad &= \quad \sum_{k=0}^{\lfloor Nt\rfloor-1}(X_k^N)^2\\
&= \quad \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big] + \sum_{k=0}^{\lfloor Nt\rfloor-1}\Big((X_k^N)^2 - \mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big]\Big).
\end{aligned}
\tag{5.2}
$$

The first term on the right hand side of (5.2) becomes:

$$
\begin{aligned}
&\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbb{E}\Big[(X_k^N)^2\big|\mathcal{F}_k^N\Big] = \frac{\alpha^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\Bigg[\int_{\mathcal{X}\times\mathcal{Y}}\Big((y - \left\langle c\sigma(w\cdot x), \nu_k^N\right\rangle)\left\langle \nabla(c\sigma(w\cdot x))\cdot\nabla f, \nu_k^N\right\rangle\Big)^2\pi(dx,dy)\\
&-\left(\int_{\mathcal{X}\times\mathcal{Y}}(y - \left\langle c\sigma(w\cdot x), \nu_k^N\right\rangle)\left\langle \nabla(c\sigma(w\cdot x))\cdot\nabla f, \nu_k^N\right\rangle\pi(dx,dy)\right)^2\Bigg]\\
= \quad &\alpha^2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\Big((y - \left\langle c\sigma(w\cdot x), \mu_s^N\right\rangle)\left\langle \nabla(c\sigma(w\cdot x))\cdot\nabla f, \mu_s^N\right\rangle\Big)^2\pi(dx,dy)ds\\
&-\alpha^2\int_0^t\left(\int_{\mathcal{X}\times\mathcal{Y}}(y - \left\langle c\sigma(w\cdot x), \mu_s^N\right\rangle)\left\langle \nabla(c\sigma(w\cdot x))\cdot\nabla f, \mu_s^N\right\rangle\pi(dx,dy)\right)^2 ds + \mathcal{O}(N^{-1})\\
= \quad &\alpha^2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathcal{R}^2_{x,y,\mu_s^N}[f]\pi(dx,dy)ds - \alpha^2\int_0^t\left(\int_{\mathcal{X}\times\mathcal{Y}}\mathcal{R}_{x,y,\mu_s^N}[f]\pi(dx,dy)\right)^2 ds + \mathcal{O}(N^{-1})\\
= \quad &\alpha^2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\left(\mathcal{R}_{x,y,\mu_s^N}[f] - \int_{\mathcal{X}\times\mathcal{Y}}\mathcal{R}_{x,y,\mu_s^N}[f]\pi(dx,dy)\right)^2\pi(dx,dy)ds + \mathcal{O}(N^{-1})
\end{aligned}
\tag{5.3}
$$

The second term on the right hand side of (5.2) can be bounded as follows:

$$
\begin{aligned}
& \mathbb{E}\Big[\Big(\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[(X_k^N)^2-\mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big]\Big]\Big)^2\Big] \\
= & \sum_{j=0}^{\lfloor Nt\rfloor-1}\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbb{E}\Big[\Big((X_k^N)^2-\mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big]\Big)\Big((X_j^N)^2-\mathbb{E}\big[(X_j^N)^2\big|\mathcal{F}_j^N\big]\Big)\Big] \\
= & \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbb{E}\Big[\Big((X_k^N)^2-\mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big]\Big)^2\Big] \\
& +2\sum_{j=0}^{\lfloor Nt\rfloor-2}\sum_{k=j+1}^{\lfloor Nt\rfloor-1}\mathbb{E}\Big[\mathbb{E}\Big[(X_k^N)^2-\mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big]\Big|\mathcal{F}_k^N\Big]\Big((X_j^N)^2-\mathbb{E}\big[(X_j^N)^2\big|\mathcal{F}_j^N\big]\Big)\Big] \\
= & \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbb{E}\Big[\Big((X_k^N)^2-\mathbb{E}\big[(X_k^N)^2\big|\mathcal{F}_k^N\big]\Big)^2\Big] \\
\leq & \; C\sum_{k=0}^{\lfloor Nt\rfloor-1}N^{-2}\leq\frac{C}{N},
\end{aligned}
$$

where the last inequality uses the bound $|X_k^N|\leq CN^{-1/2}$.

Therefore, since $\mu^N\xrightarrow{p}\bar{\mu}$ in $D_E([0,T])$ and by applying the continuous mapping theorem to (5.3), we have that for each $t\in[0,T]$,

$$
\big[\sqrt{N}\left\langle f,M_t^N\right\rangle\big]\xrightarrow{p}\alpha^2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\Big(\mathcal{R}_{x,y,\bar{\mu}_s}[f]-\int_{\mathcal{X}\times\mathcal{Y}}\mathcal{R}_{x,y,\bar{\mu}_s}[f]\pi(dx,dy)\Big)^2\pi(dx,dy)ds \tag{5.4}
$$

as $N\to\infty$.

Using the same approach as in Lemma 5.1, we also have that

$$
\lim_{N\to\infty}\mathbb{E}\left[\sup_{t\leq T}\left|\sqrt{N}\left\langle f,M_t^N\right\rangle-\sqrt{N}\left\langle f,M_{t-}^N\right\rangle\right|\right]=0. \tag{5.5}
$$

The first statement of this lemma follows from (5.4), (5.5), and Theorem 7.1.4 of [10]. The convergence of $(\sqrt{N}\left\langle f,M_t^N\right\rangle,\sqrt{N}\left\langle g,M_t^N\right\rangle)$ follows by a similar procedure and the Cramer-Wold theorem.

□

Lemma 5.3. *Let $J\geq 3\lceil\frac{D}{2}\rceil+7$. Any limit point $\bar{\eta}$ must satisfy the stochastic evolution equation*

$$
\begin{aligned}
\langle f,\bar{\eta}_t\rangle = & \langle f,\bar{\eta}_0\rangle+\int_0^t\Big(\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\langle c\sigma(w\cdot x),\bar{\mu}_s\rangle\big)\langle\sigma(w\cdot x)\partial_c f,\bar{\eta}_s\rangle\pi(dx,dy)\Big)ds \\
& -\alpha\int_0^t\Big(\int_{\mathcal{X}\times\mathcal{Y}}\langle c\sigma(w\cdot x),\bar{\eta}_s\rangle\langle\sigma(w\cdot x)\partial_c f,\bar{\mu}_s\rangle\pi(dx,dy)\Big)ds \\
& +\int_0^t\Big(\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\langle c\sigma(w\cdot x),\bar{\mu}_s\rangle\big)\langle c\sigma'(w\cdot x)x\cdot\nabla_w f,\bar{\eta}_s\rangle\pi(dx,dy)\Big)ds \\
& -\alpha\int_0^t\Big(\int_{\mathcal{X}\times\mathcal{Y}}\langle c\sigma(w\cdot x),\bar{\eta}_s\rangle\big)\langle c\sigma'(w\cdot x)x\cdot\nabla_w f,\bar{\mu}_s\rangle\pi(dx,dy)\Big)ds \\
& +\left\langle f,\bar{M}_t\right\rangle,
\end{aligned} \tag{5.6}
$$

for every $f\in W_0^{J,2}(\Theta)$.

Proof. The result can be proven by considering the pre-limit evolution equation (3.5). For each $f \in W_0^{J,2}(\Theta)$, $\sup_{t\in[0,T]} R_t^N \xrightarrow{p} 0$. Due to the uniform bound $\sup_{N\in\mathbb{N}} \sup_{t\in[0,T]} \mathbb{E}[\|\eta_t\|_{-J_1}^2] < C$, it can be shown that $\Gamma_t^{1,N} \xrightarrow{p} 0$ and $\Gamma_t^{2,N} \xrightarrow{p} 0$ uniformly in $t \in [0,T]$. Indeed, recall that

$$
\begin{aligned}
\Gamma_t^{1,N} &= \frac{1}{\sqrt{N}} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} -\alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle \left\langle \sigma(w\cdot x)\partial_c f, \eta_s^N \right\rangle \pi(dx,dy)ds \\
\Gamma_t^{2,N} &= \frac{1}{\sqrt{N}} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} -\alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle \left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \eta_s^N \right\rangle \pi(dx,dy)ds.
\end{aligned}
$$

Recall that $J_1 = 2\lceil \frac{D}{2} \rceil + 4$. Then, using the compactness of $\mathcal{X}\times\mathcal{Y}$, the bound (4.1), and Young's inequality

$$
\begin{aligned}
\mathbb{E}\left[\sup_{t\in[0,T]} \left|\Gamma_t^{1,N}\right|\right] &\leq \mathbb{E}\left[\frac{1}{\sqrt{N}} \int_0^T \int_{\mathcal{X}\times\mathcal{Y}} \|c\sigma(w\cdot x)\|_{J_1} \left\|\eta_s^N\right\|_{-J_1} \|\sigma(w\cdot x)\partial_c f\|_{J_1} \left\|\eta_s^N\right\|_{-J_1} \pi(dx,dy)ds\right] \\
&\leq \frac{C}{\sqrt{N}} \|f\|_{J_1+1}^2 \\
&\leq \frac{C}{\sqrt{N}}.
\end{aligned}
$$

Similarly, $\mathbb{E}\left[\sup_{t\in[0,T]} \left|\Gamma_t^{2,N}\right|\right] \leq \frac{C}{\sqrt{N}}$. Therefore, $\Gamma_t^{1,N} \xrightarrow{p} 0$ and $\Gamma_t^{2,N} \xrightarrow{p} 0$ uniformly in $t \in [0,T]$.

By Lemma 4.8 we have that the sequence $(\mu_t^N, \eta_t^N, \sqrt{N}M_t^N)$ is relatively compact in $D_{\mathcal{M}(\mathbb{R}^{1+d})\times W^{-J,2}\times W^{-J,2}}[0,T]$. Denoting by $(\bar{\mu}_t, \bar{\eta}_t, \bar{M}_t)$ a limiting point of an appropriate subsequence and due to the linearity of the involved operators in (3.5) we obtain by Theorem 5.5 in [11] and Lemma 5.2 that $\bar{\eta}$ satisfies (5.6). □

6 Uniqueness of the stochastic evolution equation

The limiting distribution $\bar{\eta}_t$ satisfies the stochastic evolution equation (5.6). Suppose (5.6) does not have a unique solution. Then, there are at least two solutions $\bar{\eta}^1$ and $\bar{\eta}^2$ which satisfy (5.6). Define $\Phi_t = \bar{\eta}_t^1 - \bar{\eta}_t^2$. Our goal is to show that $\|\Phi_t\|_{-J} = 0$ for all $t \leq T$. Φ_t satisfies the deterministic equation

$$
\begin{aligned}
\langle f, \Phi_t \rangle &= \alpha \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \pi(dx,dy) \Big[\big(y - \langle c\sigma(wx), \bar{\mu}_s\rangle\big) \langle c\sigma'(wx)x\cdot\nabla_w f, \Phi_s\rangle + \big(y - \langle c\sigma(wx), \bar{\mu}_s\rangle\big) \langle \sigma(wx)\partial_c f, \Phi_s\rangle \Big] ds \\
&- \alpha \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \pi(dx,dy) \Big[\langle c\sigma(wx), \Phi_s\rangle \langle c\sigma'(wx)x\cdot\nabla_w f, \bar{\mu}_s\rangle + \langle c\sigma(wx), \Phi_s\rangle \langle \sigma(wx)\partial_c f, \bar{\mu}_s\rangle \Big] ds, \\
\langle f, \Phi_0 \rangle &= 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\langle f, \Phi_t \rangle^2 &= 2\alpha \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \pi(dx,dy) \Big[\big(y - \langle c\sigma(wx), \bar{\mu}_s\rangle\big) \langle c\sigma'(wx)x\cdot\nabla_w f, \Phi_s\rangle \\
&\qquad + \big(y - \langle c\sigma(wx), \bar{\mu}_s\rangle\big) \langle \sigma(wx)\partial_c f, \Phi_s\rangle \Big] \langle f, \Phi_s\rangle \, ds \\
&- 2\alpha \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \pi(dx,dy) \Big[\langle c\sigma(wx), \Phi_s\rangle \langle c\sigma'(wx)x\cdot\nabla_w f, \bar{\mu}_s\rangle + \langle c\sigma(wx), \Phi_s\rangle \langle \sigma(wx)\partial_c f, \bar{\mu}_s\rangle \Big] \langle f, \Phi_s\rangle \, ds.
\end{aligned}
$$

Using Young's inequality, the fact that $\bar{\mu}$ takes values in a compact set, $\pi(dx,dy)$ has compact support, and the bound (4.5),

$$\begin{aligned}\langle f,\Phi_t\rangle^2 &\leq \alpha\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\pi(dx,dy)\Big[\big(y-\langle c\sigma(wx),\bar{\mu}_s\rangle\big)\,\langle c\sigma'(wx)x\nabla_w f,\Phi_s\rangle\\ &\qquad\qquad\qquad+\big(y-\langle c\sigma(wx),\bar{\mu}_s\rangle\big)\,\langle\sigma(wx)\partial_c f,\Phi_s\rangle\Big]\,\langle f,\Phi_s\rangle\,ds\\ &\quad+C\int_0^t\big(\langle f,\Phi_s\rangle^2+\|f\|_L^2\,\|\Phi_s\|_{-J}^2\big)ds,\end{aligned}\tag{6.1}$$

where $L=\lceil\frac{D}{2}\rceil+3$ and $J\geq 3\lceil\frac{D}{2}\rceil+7$.

Lemma 6.1. *For any $f\in W_0^{J,2}(\Theta)$ and every $t\in[0,T]$,*

$$\langle f,\bar{\eta}_t\rangle=\langle bf,\bar{\eta}_t\rangle,$$

where b is the bump function defined in equation (4.11).

Proof. From Lemma 4.2, there exists a bump function $b(c,w)$ such that, for any $f\in W_0^{J,2}(\Theta)$ and every $t\in[0,T]$,

$$\left\langle f,\eta_t^N\right\rangle=\left\langle bf,\eta_t^N\right\rangle.$$

Furthermore, $bf\in C_c^\infty$. Therefore, for all $N\in\mathbb{N}$,

$$\sup_{t\in[0,T]}|\left\langle f,\eta_t^N\right\rangle-\left\langle bf,\eta_t^N\right\rangle|=0.\tag{6.2}$$

Due to relative compactness, there is a sub-sequence

$$\Big(\left\langle f,\eta_\cdot^{N_k}\right\rangle,\left\langle bf,\eta_\cdot^{N_k}\right\rangle,\eta_\cdot^{N_k},\sqrt{N}M_\cdot^{N_k}\Big)\xrightarrow{d}\Big(\langle f,\bar{\eta}_\cdot\rangle,\langle bf,\bar{\eta}_\cdot\rangle,\bar{\eta}_\cdot,\bar{M}_\cdot\Big).$$

in $D_{\mathbb{R}\times\mathbb{R}\times W^{-J,2}\times W^{-J,2}}([0,T])$. Due to (6.2), any limit point must satisfy $\langle f,\bar{\eta}_t\rangle=\langle bf,\bar{\eta}_t\rangle$ for each $t\in[0,T]$. $\square$

Due to Lemma 6.1, we can re-write equation (6.1) as

$$\begin{aligned}\langle f,\Phi_t\rangle^2 &\leq \int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\pi(dx,dy)\Big[\big(y-\langle c\sigma(wx),\bar{\mu}_s\rangle\big)\,\langle b(c,w)c\sigma'(wx)x\cdot\nabla_w f,\Phi_s\rangle\\ &\quad+\big(y-\langle c\sigma(wx),\bar{\mu}_s\rangle\big)\,\langle b(c,w)\sigma(wx)\partial_c f,\Phi_s\rangle\Big]\,\langle f,\Phi_s\rangle\,ds\\ &\quad+C\int_0^t\big(\langle f,\Phi_s\rangle^2+\|f\|_L^2\,\|\Phi_s\|_{-J}^2\big)ds.\end{aligned}\tag{6.3}$$

Let $\{f_a\}_{a=1}^\infty$ be a complete orthonormal basis for $W_0^{J,2}$ where $J\geq 3\lceil\frac{D}{2}\rceil+7$. Let $f=f_a$ in equation (6.3) and then sum (6.1) over all a. By Parseval's identity,

$$\begin{aligned}\|\Phi_t\|_{-J}^2 &\leq \int_0^t\int\pi(dx,dy)\Big[\big(y-\langle c\sigma(wx),\bar{\mu}_s\rangle\big)\,\langle\Phi_s,\mathcal{G}_1^*\Phi_s\rangle_{-J}+\big(y-\langle c\sigma(wx),\bar{\mu}_s\rangle\big)\,\langle\Phi_s,\mathcal{G}_2^*\Phi_s\rangle_{-J}\Big]ds\\ &\quad+C\int_0^t\|\Phi_s\|_{-J}^2\,ds.\end{aligned}$$

The operators $\mathcal{G}^1$ and $\mathcal{G}^2$ are defined in equation (4.12). Since $\bar{\mu}_t$ takes values in a compact set and $\pi(dx,dy)$ has compact support,

$$\begin{aligned}\|\Phi_t\|_{-J}^2 &\leq C_1\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\pi(dx,dy)\Big(|\,\langle\Phi_s,\mathcal{G}_1^*\Phi_s\rangle_{-J}\,|+|\,\langle\Phi_s,\mathcal{G}_2^*\Phi_s\rangle_{-J}\,|\Big)ds\\ &\quad+C_2\int_0^t\|\Phi_s\|_{-J}^2\,ds.\end{aligned}$$

Using Lemma 4.3 and the fact that $\mathcal{X} \times \mathcal{Y}$ is a compact set,

$$\|\Phi_t\|_{-J}^2 \quad \leq \quad C \int_0^t \|\Phi_s\|_{-J}^2 \, ds,$$

which then by Gronwall's inequality gives $\|\Phi_t\|_{-J}^2 = 0$ for $t \in [0,T]$. Thus, we have established the following result.

Theorem 6.2. *Let $J \geq 3\lceil \frac{D}{2} \rceil + 7$ with $D = d+1$. Then, the solution $\bar{\eta}$ to the stochastic evolution equation (5.6) is unique in $W^{-J,2}$.*

7 Proof of the Main Result

We now collect our results and prove Theorem 1.5. By Lemma 4.8 we have that the sequence $(\mu_t^N, \eta_t^N, \sqrt{N} M_t^N)$ is relatively compact in $D_{\mathcal{M}(\mathbb{R}^{1+d}) \times W^{-J,2} \times W^{-J,2}}([0,T])$. Lemma 5.3 establishes that the limit point satisfies the SPDE (1.5) and Theorem 6.2 proves that limit point is unique. Therefore, by Prokhorov's Theorem, $\eta^N \xrightarrow{d} \bar{\eta}$ in $D_{W^{-J,2}}([0,T])$ where $\bar{\eta}$ satisfies the stochastic evolution equation (1.5).

8 Conclusion

Neural networks are nonlinear machine learning models whose parameters are estimated from data using stochastic gradient descent. They have achieved immense practical success over the past decade in a variety of applications in image, speech, and text recognition. However, there is limited mathematical understanding of their properties. This paper studies neural networks with a single hidden layer in the asymptotic regime of large network sizes and large numbers of stochastic gradient descent iterations. We rigorously prove a central limit theorem (CLT) for the empirical distribution of the neural network parameters. The CLT satisfies a stochastic partial differential equation which has a Gaussian distribution.

A Auxiliary lemmas

Lemma A.1. *If $\Psi \in C_0^\infty(\Theta)$, $g \in C_0^\infty(\Theta)$, then, there exists a constant $C < \infty$ such that*

$$\int_\Theta D^k \left[g \frac{\partial \Psi}{\partial w} \right] D^k \Psi dc dw \leq C \|\Psi\|_J^2 . \tag{A.1}$$

Proof of Lemma A.1. We prove the statement for $d = 1$. The algebra for $d > 1$ is similar, albeit more tedious. Let $k = k_1 + k_2$ with $k_1, k_2 \geq 0$ arbitrarily chosen.

$$\begin{aligned} \int_\Theta D^k \left[g \frac{\partial \Psi}{\partial w} \right] D^k \Psi dc dw &= \int_\Theta \frac{\partial^k}{\partial c^{k_1} \partial w^{k_2}} \left[g \frac{\partial \Psi}{\partial w} \right] D^k \Psi dc dw \\ &= \sum_{\substack{\alpha_1 + \alpha_2 = k+1, \alpha_2 \leq k \\ i_1 + i_2 = k_1, \\ j_1 + j_2 = k_2}} \int_\Theta \frac{\partial^{\alpha_1} g}{\partial c^{i_1} \partial w^{j_1}} \frac{\partial^{\alpha_2} \Psi}{\partial c^{i_2} \partial w^{j_2}} D^k \Psi dc dw \\ &+ \int_\Theta g \frac{\partial}{\partial w} \left[\frac{\partial^k \Psi}{\partial c^{k_1} \partial w^{k_2}} \right] D^k \Psi dc dw \end{aligned} \tag{A.2}$$

Since $g \in C_0^\infty(\bar{\Theta})$ and using Young's inequality,

$$\begin{aligned}
\sum_{\substack{\alpha_1+\alpha_2=k+1,\alpha_2\le k\\ i_1+i_2=k_1,\\ j_1+j_2=k_2}} \int_\Theta \frac{\partial^{\alpha_1} g}{\partial c^{i_1}\partial w^{j_1}} \frac{\partial^{\alpha_2}\Psi}{\partial c^{i_2}\partial w^{j_2}} D^k\Psi dcdw &\le C \sum_{\substack{\alpha_1+\alpha_2=k+1,\alpha_2\le k\\ i_1+i_2=k_1,\\ j_1+j_2=k_2}} \int_\Theta \left|\frac{\partial^{\alpha_2}\Psi}{\partial c^{i_2}\partial w^{j_2}}\right| \left|D^k\Psi\right| dcdw \\
&\le C \sum_{\substack{\alpha_1+\alpha_2=k+1,\alpha_2\le k\\ i_1+i_2=k_1,\\ j_1+j_2=k_2}} \int_\Theta \left(\left|\frac{\partial^{\alpha_2}\Psi}{\partial c^{i_2}\partial w^{j_2}}\right|^2 + \left|D^k\Psi\right|^2\right) dcdw \\
&\le C\left\|\Psi\right\|_J^2 . \qquad \text{(A.3)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\int_\Theta D^k\left[g\frac{\partial\Psi}{\partial w}\right] D^k\Psi dcdw &\le C_1\left\|\Psi\right\|_J^2 + \int_\Theta g\frac{\partial}{\partial w}\left[\frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}}\right]\frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}} dcdw \\
&= C_2\left\|\Psi\right\|_J^2 - \int_\Theta \frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}}\frac{\partial}{\partial w}\left[g\frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}}\right] dcdw.
\end{aligned}$$

The inequality on line 2 follows from the bound (A.3). The third line follows from integration by parts and the fact that $g \in C_0^\infty(\Theta)$.

We next consider the other term

$$\begin{aligned}
\frac{\partial}{\partial w}\left[g\frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}}\right] &= gD^k\left[\frac{\partial\Psi}{\partial w}\right] + \frac{\partial g}{\partial w}(c,w)D^k\Psi \\
&= D^k\left[g\frac{\partial\Psi}{\partial w}\right] + \frac{\partial g}{\partial w}(c,w)D^k\Psi - \sum_{\substack{\alpha_1+\alpha_2=k+1,\alpha_2\le k\\ i_1+i_2=k_1,\\ j_1+j_2=k_2}} \frac{\partial^{\alpha_1} g}{\partial c^{i_1}\partial w^{j_1}} \frac{\partial^{\alpha_2}\Psi}{\partial c^{i_2}\partial w^{j_2}},
\end{aligned}$$

where the last term is from (A.2). Now, by applying the same approach as in (A.3), i.e. using Young's inequality and $g \in C_0^\infty(\bar{\Theta})$, we have the bound

$$\begin{aligned}
\int_\Theta D^k\left[g\frac{\partial\Psi}{\partial w}\right] D^k\Psi dcdw &\le C_2\left\|\Psi\right\|_J^2 - \int_\Theta \frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}}\frac{\partial}{\partial w}\left[g\frac{\partial^k\Psi}{\partial c^{k_1}\partial w^{k_2}}\right] dcdw \\
&\le C\left\|\Psi\right\|_J^2 - \int_\Theta D^k\left[g\frac{\partial\Psi}{\partial w}\right] D^k\Psi dcdw.
\end{aligned}$$

Rearranging, we have that there is a constant $C < \infty$ (different than above)

$$\int_\Theta D^k\left[g(c,w)\frac{\partial\Psi}{\partial w}\right] D^k\Psi dcdw \le C\left\|\Psi\right\|_J^2 .$$

□

References

[1] R. Adams. Sobolev Spaces. Academic Press, New York, 1978.

[2] D. L. Burkholder, Distribution function inequalities for martingales, *Annals of Probability*, 1, 19-42, 1973.

[3] P. Bartlett, D. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. Advances in Neural Information Processing Systems, 6241-6250, 2017.

[4] L. Bo and A. Capponi. Systemic risk in interbanking networks. SIAM Journal on Financial Mathematics. 6(1),386-424, 2015.

[5] P. Dai Pra, W. Runggaldier, E. Sartori, and M. Tolotti. Large portfolio losses: A dynamic contagion model. The Annals of Applied Probability. 19(1), 347-394, 2009.

[6] P. Dai Pra and F. Hollander. McKean-Vlasov limit for interacting random processes in random media. Journal of Statistical Physics. 84(3-4), 735-772, 1996.

[7] P. Dai Pra and M. Tolotti. Heterogeneous credit portfolios and the dynamics of the aggregate losses. Stochastic Processes and their Applications. 119(9), 2913-2944, 2009.

[8] E. Del Barrio, P. Deheuvels, and S. Van De Geer. Lectures on Empirical Processes: Theory and Statistical Applications. European Mathematical Society Publishing House, Zurich, 2007.

[9] F. Delarue, J. Inglis, S. Rubenthaler, and E. Tanre Particle systems with a singular mean-field self-excitation. Application to neuronal networks. Stochastic Processes and their Applications, 125(6), 2451-2492, 2015.

[10] S. Ethier and T. Kurtz. Markov Processes: Characterization and Convergence. 1986, Wiley, New York, MR0838085.

[11] S. Ethier and P. Protter. Weak convergence of stochastic integrals and differential equations, II. Infinite-dimensional case, in: Probabilistic Models for Nonlinear Partial Differential Equations, in: Lecture Notes in Mathematics. Springer-Verlag, 197-285, 1996.

[12] B. Fernandez and S. Meleard. A Hilbertian approach for fluctuations on the McKean-Vlasov model. Stochastic Processes and their Applications, 71, 33-53, 1997.

[13] R. Fry and S. McManus. Smooth bump functions and the geometry of Banach spaces: a brief survey. Expositiones Mathematicae, 20(2):143-83, 2002.

[14] K. Giesecke, K. Spiliopoulos, and R. Sowers. Default clustering in large portfolios: Typical events. The Annals of Applied Probability. 23(1), 2013, 348-385.

[15] K. Giesecke, K. Spiliopoulos, R. Sowers, and J. Sirignano. Large portfolio asymptotics for loss from default. Mathematical Finance. 25(1), 77-114, 2015.

[16] A.D. Gottlieb. Markov transitions and the propagation of chaos. In: ProQuest LLC, Ann Arbor, MI. PhD Thesis, University of California, Berkeley. 1998.

[17] B. Hambly and S. Ledger. A stochastic McKean-Vlasov equation for absorbing diffusions on the half-line. The Annals of Applied Probability. 27(5),2698-2752, 2017.

[18] K. Hornik. Approximation capabilities of multilayer feedforward networks. Neural Networks, 4(2), 251-257, 1991.

[19] J. Inglis and D. Talay. Mean-field limit of a stochastic particle system smoothly interacting through threshold hitting-times and applications to neural networks with dendritic component. SIAM Journal on Mathematical Analysis, 47(5), 3884-3916, 2015.

[20] V.N. Kolokoltsov. Nonlinear Markov processes and kinetic equations Vol. 182, Cambridge University Press, 2010.

[21] T. Kurtz and J. Xiong. A stochastic evolution equation arising from the fluctuations of a class of interacting particle systems. Communications in Mathematical Sciences, 2(3), 325–358, 2004.

[22] S. Mei, A. Montanari, and P. Nguyen A mean field view of the landscape of two-layer neural networks 2018, arXiv: 1804.06561.

[23] O. Moynot and M. Samuelides. Large deviations and mean-field theory for asymmetric random recurrent neural networks. Probability Theory and Related Fields, 123(1), 41-75, 2002.

[24] G. M. Rotskoff and E. Vanden-Eijnden Neural Networks as Interacting Particle Systems: Asymptotic Convexity of the Loss Landscape and Universal Scaling of the Approximation Error. 2018, arXiv:1805.00915.

[25] J. Sirignano and K. Spiliopoulos. Mean Field Analysis of Neural Networks 2018, arXiv:1805.01053.

[26] H. Sompolinsky, A. Crisanti, and H. Sommers. Chaos in random neural networks. Physical Review Letters, 61(3), 259, 1988.

[27] K. Spiliopoulos, J. Sirignano, and K. Giesecke. Fluctuation Analysis for the Loss from Default. Stochastic Processes and their Applications. 124, 2322-2362, 2014.

[28] A-S. Sznitman. Topics in propagation of chaos. in Ecole d'Eté de Probabilitiés de Saint-Flour XIX - 1989. series, Lecture Notes in Mathematics, P.-L. Hennequin, Ed. Springer, Berlin Heidelberg. 1464, 165-251, 1991.

[29] J. Touboul. Propagation of chaos in neural fields. The Annals of Applied Probability, 24(3), 1298-1328, 2014.

[30] C. Wang, J. Mattingly, and Y. Lu Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA. 2017, arXiv:1712.04332.