Scaling Limit of Neural Networks with the Xavier Initialization and Convergence to a Global Minimum

Justin Sirignano* and Konstantinos Spiliopoulos†‡

July 10, 2019

Abstract

We analyze single-layer neural networks with the Xavier initialization in the asymptotic regime of large numbers of hidden units and large numbers of stochastic gradient descent training steps. We prove the neural network converges in distribution to a random ODE with a Gaussian distribution using mean field analysis. The limit is completely different than in the typical mean-field results for neural networks due to the $\frac{1}{\sqrt{N}}$ normalization factor in the Xavier initialization (versus the $\frac{1}{N}$ factor in the typical mean-field framework). Although the pre-limit problem of optimizing a neural network is non-convex (and therefore the neural network may converge to a local minimum), the limit equation minimizes a (quadratic) convex objective function and therefore converges to a global minimum. Furthermore, under reasonable assumptions, the matrix in the limiting quadratic objective function is positive definite and thus the neural network (in the limit) will converge to a global minimum with zero loss on the training set.

1 Introduction

Consider a single-layer neural network with the Xavier initialization:

$$g^N(x;\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} C^i \sigma(W^i x),$$

where $C^i \in \mathbb{R}$, $W^i \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, and $\sigma(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$. The number of hidden units is N and the output is scaled by a factor $\frac{1}{\sqrt{N}}$ (the widely-used Xavier initialization, see [2]).

The objective function is

$$\mathcal{L}^N(\theta) = \mathbb{E}\Big[(Y - g^N(X;\theta))^2\Big],$$

where the data $(X,Y) \sim \pi(dx,dy)$, $Y \in \mathbb{R}$, and the parameters $\theta = (C^1,\ldots,C^N,W^1,\ldots,W^N) \in \mathbb{R}^{N\times(1+d)}$. For notational convenience, we may refer to $g^N(x;\theta)$ as $g^N(x)$ in our analysis below.

The model parameters θ are trained using stochastic gradient descent. The parameter updates are given by:

$$\begin{aligned} C_{k+1}^i &= C_k^i + \frac{\alpha^N}{\sqrt{N}}(y_k - g_k^N(x_k))\sigma(W_k^i x_k), \\ W_{k+1}^i &= W_k^i + \frac{\alpha^N}{\sqrt{N}}(y_k - g_k^N(x_k))C_k^i \sigma'(W_k^i x_k)x_k, \\ g_k^N(x) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} C_k^i \sigma(W_k^i x), \end{aligned} \tag{1.1}$$

*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)

for $k = 0, 1, \ldots, TN$ where $T > 0$. α^N is the learning rate (which may depend upon N). The data samples are (x_k, y_k) are i.i.d. samples from a distribution $\pi(dx, dy)$.

We impose the following assumption.

Assumption 1.1. We have that

- The activation function $\sigma \in C_b^2(\mathbb{R})$, i.e. σ is twice continuously differentiable and bounded.
- The randomly initialized parameters (C_0^i, W_0^i) are i.i.d., mean-zero random variables with a distribution $\mu_0(dc, dw)$.
- The random variable C_0^i has compact support and $\langle \|w\|, \mu_0 \rangle < \infty$.
- The sequence of data samples (x_k, y_k) is i.i.d. from the probability distribution $\pi(dx, dy)$.
- There is a fixed dataset of M data samples $(x^{(i)}, y^{(i)})_{i=1}^M$ and therefore $\pi(dx, dy) = \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} \delta_{(x^{(i)}, y^{(i)})}(dx, dy)$.

Note that the last assumption also implies that $\pi(dx, dy)$ has compact support.

We will study the limiting behavior of the network output $g_k^N(x)$ for $x \in \mathcal{D} = \{x^{(1)}, \ldots, x^{(M)}\}$ as the number of hidden units N and stochastic gradient descent steps TN simultaneously become large. The network output converges in distribution to the solution of a random ODE as $N \to \infty$.

1.1 Main Results

Define the empirical measure

$$\nu_k^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{C_k^i, W_k^i}.$$

Note that the neural network output can be written as the inner-product

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

Due to Assumption 1.1, as $N \to \infty$ and for $x \in \mathcal{D}$,

$$g_0^N(x) \xrightarrow{d} \mathcal{G}(x), \tag{1.2}$$

where $\mathcal{G} \in \mathbb{R}^M$ is a Gaussian random variable. We also of course have that

$$\nu_0^N \xrightarrow{p} \nu_0 \equiv \mu_0.$$

Define the scaled processes

$$\begin{aligned} h_t^N &= g_{\lfloor Nt \rfloor}^N, \\ \mu_t^N &= \nu_{\lfloor Nt \rfloor}^N, \end{aligned}$$

where $g_k^N = \left(g_k^N(x^{(1)}), \ldots, g_k^N(x^{(M)}) \right)$, $h_t^N(x) = g_{\lfloor Nt \rfloor}^N(x)$, and $h_t^N = \left(h_t^N(x^{(1)}), \ldots, h_t^N(x^{(M)}) \right)$.

We will study convergence in distribution of the random process (μ_t^N, h_t^N) as $N \to \infty$ in the space $D_E([0, T])$ where $E = \mathcal{M}(\mathbb{R}^{1+d}) \times \mathbb{R}^M$. $D_E([0, T])$ is the Skorokhod space and $\mathcal{M}(S)$ is the space of probability measures on S.

The main contribution of our work is a rigorous proof that a neural network with the Xavier initialization and trained with stochastic gradient descent converges in distribution to a random ODE as the number of units and training steps become large. In addition, our convergence analysis will also address several interesting questions:

- Our results provide a rigorous convergence guarantee for the Xavier initialization (i.e., the $\frac{1}{\sqrt{N}}$ normalization factor), which is almost universally used in deep learning models. *A priori* it is unclear if the neural network $g_k^N(x)$ will converge as $N \to \infty$ since, for $k > 0$, the $C^i\sigma(W^i x)$ is correlated with $C^j\sigma(W^j x)$ and therefore a limit may not exist. If a limit did not exist, this would imply that the neural network model could have poor numerical behavior for large N. We prove that a limit does exist.

- Although the pre-limit problem of optimizing a neural network is non-convex (and therefore the neural network may converge to a local minimum), the limit equation minimizes a quadratic objective function.

- We show that the matrix in the limiting quadratic objective function is positive definite, and therefore the neural network (in the limit) will converge to a global minimum with zero loss on the training set.

Convergence to a global minimum for a neural network has been recently proven in [3], [4], and [5]. Our work contributes to this growing literature by showing that convergence to a global minimum is a simple consequence of the mean-field limit for neural networks. A detailed discussion of these papers and other related literature is provided in Section 1.2.

Our main results are presented below.

Theorem 1.2. The process (μ_t^N, h_t^N) converges in distribution in the space $D_E([0,T])$ as $N \to \infty$ to (μ_t, h_t) which satisfies, for every $f \in C_2^b(\mathbb{R}^{1+d})$, the random ODE

$$\begin{aligned} h_t(x) &= h_0(x) + \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - h_t(x')) \left\langle \sigma(wx)\sigma(wx'), \mu_t \right\rangle \pi(dx', dy)dt \\ &+ \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - h_t(x')) \left\langle c^2\sigma'(wx')\sigma'(wx)xx', \mu_t \right\rangle \pi(dx', dy)dt, \\ h_0(x) &= \mathcal{G}(x), \\ \langle f, \mu_t \rangle &= \langle f, \mu_0 \rangle . \end{aligned} \tag{1.3}$$

Proof. See Sections 3, 4, and 5. □

Recall that $\mathcal{G} \in \mathbb{R}^M$ is a Gaussian random variable; see equation (1.2). In addition, note that $\bar{\mu}_t$ in the limit equation (1.3) is a constant, i.e. $\mu_t = \mu_0$ for $t \in [0,T]$. Therefore, (1.3) reduces to

$$\begin{aligned} h_t(x) &= h_0(x) + \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - h_t(x')) \left\langle \sigma(wx)\sigma(wx'), \mu_0 \right\rangle \pi(dx', dy)dt \\ &+ \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - h_t(x')) \left\langle c^2\sigma'(wx')\sigma'(wx)xx', \mu_0 \right\rangle \pi(dx', dy)dt, \\ h_0(x) &= \mathcal{G}(x). \end{aligned} \tag{1.4}$$

Since (1.4) is a linear equation in $C_{\mathbb{R}^M}([0,T])$, the solution h_t is unique.

To better understand (1.4), define the matrix $A \in \mathbb{R}^{M\times M}$ where

$$A_{x,x'} = \frac{\alpha}{M} \left\langle \sigma(wx)\sigma(wx'), \mu_0 \right\rangle + \frac{\alpha}{M} \left\langle c^2\sigma'(wx')\sigma'(wx)xx', \mu_0 \right\rangle,$$

where $x, x' \in \mathcal{D}$. A is finite-dimensional since we fixed a training set of size M in the beginning.

Then, (1.4) becomes

$$\begin{aligned} dh_t &= A\left(\hat{Y} - h_t\right)dt, \\ h_0 &= \mathcal{G}, \end{aligned}$$

where $\hat{Y} = (y^{(1)}, \ldots, y^{(M)})$.

Therefore, h_t is the solution to a continuous-time gradient descent algorithm which minimizes a quadratic objective function.

$$\begin{aligned} \frac{dh_t}{dt} &= -\frac{1}{2}\nabla_h J(\hat{Y}, h_t), \\ J(y,h) &= \big(y-h\big)^\top A\big(y-h\big)\Big], \\ h_0 &= \mathcal{G}. \end{aligned}$$

Therefore, even though the pre-limit optimization problem is non-convex, the neural network's limit will minimize a quadratic objective function.

An interesting question is whether $h_t \to \hat{Y}$ as $t \to \infty$. That is, in the limit of large numbers of hidden units and many training steps, does the neural network model converge to a global minimum with zero training error. Theorem 1.3 shows that $h_t \to \hat{Y}$ as $t \to \infty$ if A is positive definite. Corollary 1.4 proves that, under reasonable hyperparameter choices and if the data samples are distinct, A will be positive definite.

Theorem 1.3. If A is positive definite, then

$$h_t \to \hat{Y} \quad \text{as} \quad t \to \infty.$$

Proof. Consider the transformation $\tilde{h}_t = h_t - \hat{Y}$. Then,

$$\begin{aligned} d\tilde{h}_t &= -A\tilde{h}_t dt, \\ \tilde{h}_0 &= \mathcal{G} - \hat{Y}. \end{aligned}$$

Then, $\tilde{h}_t \to 0$ (and consequently $h_t \to \hat{Y}$) as $t \to \infty$ if A is positive definite.

□

Corollary 1.4. A sufficient condition for A being positive definite is $\sigma(\cdot)$ is a continuous, monotone increasing function where $\lim_{z\to-\infty}\sigma(z) = -1$ (or $= 0$) and $\lim_{z\to-\infty}\sigma(z) = 1$, $W_0^{i,j}$ and $W_0^{i,j'}$ are independent for $j \neq j'$, $C_0^i = 0$, $W_0^i \sim \mathcal{N}(0,1)$, and the data samples are distinct (i.e., $x^{(i)} \neq x^{(j)}$ for $i \neq j$).

Proof. See Section 6.

□

Examples of activation units $\sigma(\cdot)$ satisfying the conditions in Corollary 1.4 include sigmoid functions and hyperbolic tangent functions. The data samples in the dataset will be distinct with probability 1 if the random variable X has a probability density function. Using a normal distribution for the initialization of the parameters in the neural network is a common choice in practice.

1.2 Literature Review

[11], [12], [13], and [15] study the asymptotics of single-layer neural networks with a $\frac{1}{N}$ normalization; that is, $g^N(x;\theta) = \frac{1}{N}\sum_{i=1}^N C^i\sigma(W^i x)$. [14] studies the asymptotics of deep (i.e., multi-layer) neural networks with a $\frac{1}{N}$ normalization in each hidden layer. In the single layer case, the limit for the neural network satisfies a partial differential equation. As discussed in [11], it is *not* necessarily true that the limiting equation (a PDE in this case) will converge to the global minimum of an objective function with zero training error.

The $\frac{1}{N}$ normalization studied in [11], [12], [13], and [15] is convenient since the single-layer neural network is then in a traditional mean-field framework where it can be described via an empirical measure of the parameters. However, the $\frac{1}{\sqrt{N}}$ normalization that we study in this paper is more widely-used in practice (it is referred to as the Xavier initialization and was first introduced in [2]). The $\frac{1}{\sqrt{N}}$ normalization requires different analysis than the standard mean-field analysis $\frac{1}{N}$, and it produces a completely different limit. Importantly, under reasonable conditions, the limit equation converges to a global minimum with zero training error. In addition, for the limit to hold, we show that the $\frac{1}{\sqrt{N}}$ normalization requires the effective learning rate for the parameters to be of the order $N^{-3/2}$.

Convergence to a global minimum for a neural network has been recently proven in [3], [4], and [5]. Although it has been long understood that neural networks have universal approximation properties (see [8], [9], and [10]), it has until recently been commonly believed that training algorithms for neural networks (e.g., gradient descent) may converge to a local minimum (and not a global minimum) since neural networks are non-convex. [3], [4], and [5] showed that neural networks (under suitable conditions) will converge to a global minimum during training. This result is quite remarkable considering the optimization problem is non-convex, and it provides an important mathematical guarantee for the field of deep learning.

[3], [4], and [5] do not study the mean-field limit of a neural network with the Xavier initialization, which is the focus of our paper. Once the mean field limit is established, we show that convergence to a global minimum is a simple consequence of the limit equation. There are also some differences between our assumptions and the assumptions required for the theorems of [3], [4], and [5]. [3] and [4] study gradient descent while our paper studies stochastic gradient descent, which introduces additional technical challenges due to the stochastic dynamics. [5] studies stochastic gradient descent for a framework where the neural network's output layer parameters are not trained. In their paper, the C^i parameters are randomly generated and then frozen (i.e., they do not change during training). In practice, all of the parameters in the neural network, including the output layer parameters, are trained with stochastic gradient descent and therefore it is worthwhile to consider the more general case. [5] also imposes an assumption that the loss function vanishes at infinity. [3], [4], and [5] all require that every data sample has the same magnitude, i.e. $\left\|x^{(i)}\right\| = 1$ for every $i = 1, \ldots, M$. We do not require this assumption.

[6] proved a limit for neural networks with a Xavier initialization when they are trained with continuous-time gradient descent. Our paper proves a limit for neural networks trained with the (standard) discrete-time stochastic gradient descent algorithm which is used in practice. Our method of proof is also different than the approach of [6]. Whereas [6] begins their analysis in continuous time (due to their framework being continuous-time gradient descent), our paper rigorously passes from discrete time (where the stochastic gradient descent updates evolve) to continuous time through weak convergence analysis of appropriate stochastic processes and measure-valued processes. In [6], the authors directly study the evolution of the derivatives of the output with respect to the parameters, while we address the limiting behavior of the underlying associated stochastic processes and measure-valued processes.

1.3 Organization of Paper

Section 2 derives equations describing the evolution of the pre-limit process (μ^N, h^N). Relative compactness of the family of processes (μ^N, h^N) is proven in Section 3. Section 4 proves that any limit point of the process must satisfy the equation (1.3). These results are collected together in Section 5 to prove that (μ^N, h^N) converges in distribution to the solution of equation (1.3). Corollary 1.4 is proven in Section 6.

2 Evolution of the Pre-limit Process

We begin by analyzing evolution of the network output $g_k^N(x)$. Using a Taylor expansion,

$$\begin{aligned}
g_{k+1}^N(x) &= g_k^N(x) + \frac{1}{\sqrt{N}}\sum_{i=1}^N C_{k+1}^i \sigma(W_{k+1}^i x) - \frac{1}{\sqrt{N}}\sum_{i=1}^N C_k^i \sigma(W_k^i x) \\
&= g_k^N(x) + \frac{1}{\sqrt{N}}\sum_{i=1}^N \left(C_{k+1}^i \sigma(W_{k+1}^i x) - C_k^i \sigma(W_k^i x) \right) \\
&= g_k^N(x) + \frac{1}{\sqrt{N}}\sum_{i=1}^N \left((C_{k+1}^i - C_k^i)\sigma(W_{k+1}^i x) + (\sigma(W_{k+1}^i x) - \sigma(W_k^i x))C_k^i \right) \\
&= g_k^N(x) + \frac{1}{\sqrt{N}}\sum_{i=1}^N \left((C_{k+1}^i - C_k^i)\left[\sigma(W_k^i x) + \sigma'(W_k^{i,*} x_k) x (W_{k+1}^i - W_k^i) \right] \right. \\
&+ \left. \left[\sigma'(W_k^i x)(W_{k+1}^i - W_k^i)x + \frac{1}{2}\sigma''(W_{k+1}^{i,**} x)\left((W_{k+1}^i - W_k^i)x\right)^2 \right] C_k^i \right), \qquad (2.1)
\end{aligned}$$

for points $W_k^{i,*}$ and $W_k^{i,*,*}$ in the line segment connecting the points W_k^i and W_{k+1}^i. Let $\alpha^N = \frac{\alpha}{N}$. Substituting (1.1) into (2.1) yields

$$\begin{aligned} g_{k+1}^N(x) &= g_k^N(x) + \frac{\alpha}{N^2}\sum_{i=1}^N (y_k - g_k^N(x_k))\sigma(W_k^i x_k)\sigma(W_k^i x) \\ &+ \frac{\alpha}{N^2}\sum_{i=1}^N \sigma'(W_k^i x)(y_k - g_k^N(x_k))\sigma'(W_k^i x_k)x_k x\big(C_k^i\big)^2 + \mathcal{O}(N^{-3/2}). \end{aligned} \tag{2.2}$$

We can re-write the evolution of $g_k^N(x)$ in terms of the empirical measure ν_k^N.

$$\begin{aligned} g_{k+1}^N(x) &= g_k^N(x) + \frac{\alpha}{N}(y_k - g_k^N(x_k))\left\langle \sigma(wx_k)\sigma(wx), \nu_k^N\right\rangle \\ &+ \frac{\alpha}{N}(y_k - g_k^N(x_k))x_k x\left\langle \sigma'(wx)\sigma'(wx_k)c^2, \nu_k^N\right\rangle + \mathcal{O}(N^{-3/2}). \end{aligned} \tag{2.3}$$

Using (2.3), we can write the evolution of h_t^N for $t \in [0,T]$ as

$$\begin{aligned} h_t^N &= h_0^N + \sum_{k=0}^{\lfloor Nt\rfloor -1}(g_{k+1}^N - g_k) \\ &= h_0^N + \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}(y_k - g_k^N(x_k))\left\langle \sigma(wx_k)\sigma(wx), \nu_k^N\right\rangle \\ &+ +\frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}(y_k - g_k^N(x_k))x_k x\left\langle \sigma'(wx)\sigma'(wx_k)c^2, \nu_k^N\right\rangle \\ &+ \mathcal{O}(N^{-1/2}) \end{aligned}$$

Next, we decompose the summations into a drift and martingale component.

$$\begin{aligned} h_t^N &= h_0^N + \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}\int_{\mathcal{X}\times\mathcal{Y}}(y - g_k^N(x'))\left\langle \sigma(wx')\sigma(wx), \nu_k^N\right\rangle \pi(dx', dy) \\ &+ \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}\int_{\mathcal{X}\times\mathcal{Y}}(y - g_k^N(x'))x' x\left\langle \sigma'(wx)\sigma'(wx')c^2, \nu_k^N\right\rangle \pi(dx', dy) \\ &+ \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}\left((y_k - g_k^N(x_k))\left\langle \sigma(wx_k)\sigma(wx), \nu_k^N\right\rangle - \int_{\mathcal{X}\times\mathcal{Y}}(y - g_k^N(x'))\left\langle \sigma(wx')\sigma(wx), \nu_k^N\right\rangle \pi(dx', dy)\right) \\ &+ \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor -1}\left((y_k - g_k^N(x_k))x_k x\left\langle \sigma'(wx)\sigma'(wx_k)c^2, \nu_k^N\right\rangle - \int_{\mathcal{X}\times\mathcal{Y}}(y - g_k^N(x'))x' x\left\langle \sigma'(wx)\sigma'(wx')c^2, \nu_k^N\right\rangle \pi(dx', dy)\right) \\ &+ \mathcal{O}(N^{-1/2}) \end{aligned}$$

For convenience, we define the martingale terms (the third and fourth terms in the equation above) as $M_t^{N,1}$ and $M_t^{N,2}$, respectively. The equation for h_t^N can be re-written in terms of a Riemann integral and the scaled measure μ_t^N, yielding

$$\begin{aligned} h_t^N &= h_0^N + \alpha\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}(y - h_s^N(x'))\left\langle \sigma(wx')\sigma(wx), \mu_s^N\right\rangle \pi(dx', dy)ds \\ &+ \alpha\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}(y - h_s^N(x'))x' x\left\langle \sigma'(wx)\sigma'(wx')c^2, \mu_s^N\right\rangle \pi(dx', dy)ds \\ &+ M_t^{N,1} + M_t^{N,2} + \mathcal{O}(N^{-1/2}). \end{aligned} \tag{2.4}$$

In addition, using conditional independence of the terms in the series for $M_t^{N,1}$ and $M_t^{N,2}$ as well as the bounds from Lemmas 3.2 and 3.1, we have that

$$\begin{aligned}\mathbb{E}\left[\left(M_t^{N,1}\right)^2\right] &\leq \frac{K}{N},\\ \mathbb{E}\left[\left(M_t^{N,2}\right)^2\right] &\leq \frac{K}{N}.\end{aligned}$$

We can also analyze the evolution of the empirical measure ν_k^N in terms of test functions $f \in C_b^2(\mathbb{R}^{1+d})$. Using a Taylor expansion, we find that

$$\begin{aligned}\left\langle f, \nu_{k+1}^N\right\rangle - \left\langle f, \nu_k^N\right\rangle &= \frac{1}{N}\sum_{i=1}^N \Big(f(C_{k+1}^i, W_{k+1}^i) - f(C_k^i, W_k^i)\Big)\\ &= \frac{1}{N}\sum_{i=1}^N \partial_c f(C_k^i, W_k^i)(C_{k+1}^i - C_k^i) + \frac{1}{N}\sum_{i=1}^N \nabla_w f(C_k^i, W_k^i)(W_{k+1}^i - W_k^i)\\ &+ \frac{1}{N}\sum_{i=1}^N \partial_c^2 f(\bar{C}_k^i, \bar{W}_k^i)(C_{k+1}^i - C_k^i)^2 + \frac{1}{N}\sum_{i=1}^N (C_{k+1}^i - C_k^i)\nabla_{cw} f(\bar{C}_k^i, \bar{W}_k^i)(W_{k+1}^i - W_k^i)\\ &+ \frac{1}{N}\sum_{i=1}^N (W_{k+1}^i - W_k^i)^\top \nabla_w^2 f(\bar{C}_k^i, \bar{W}_k^i)(W_{k+1}^i - W_k^i), \qquad (2.5)\end{aligned}$$

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.

Substituting (1.1) into (2.5) yields

$$\begin{aligned}\left\langle f, \nu_{k+1}^N\right\rangle - \left\langle f, \nu_k^N\right\rangle &= N^{-5/2}\sum_{i=1}^N \partial_c f(C_k^i, W_k^i)\alpha(y_k - g_k^N(x_k))\sigma(W_k^i x_k)\\ &+ N^{-5/2}\sum_{i=1}^N \alpha(y_k - g_k^N(x_k))C_k^i\sigma'(W_k^i x_k)\nabla_w f(C_k^i, W_k^i)\cdot x_k + O_p\left(N^{-2}\right)\\ &= N^{-3/2}\alpha(y_k - g_k^N(x_k))\left\langle \partial_c f(c,w)\sigma(wx_k), \nu_k^N\right\rangle\\ &+ N^{-3/2}\alpha(y_k - g_k^N(x_k))\left\langle c\sigma'(wx_k)\nabla_w f(c,w)\cdot x_k, \nu_k^N\right\rangle + O_p\left(N^{-2}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\left\langle f, \mu_t^N\right\rangle &= \left\langle f, \mu_0^N\right\rangle + \sum_{k=0}^{\lfloor Nt\rfloor - 1}\Big(\left\langle f, \nu_{k+1}^N\right\rangle - \left\langle f, \nu_k^N\right\rangle\Big)\\ &= \left\langle f, \mu_0^N\right\rangle + N^{-3/2}\sum_{k=0}^{\lfloor Nt\rfloor - 1}\alpha(y_k - g_k^N(x_k))\left\langle \partial_c f(c,w)\sigma(wx_k), \nu_k^N\right\rangle\\ &+ N^{-3/2}\sum_{k=0}^{\lfloor Nt\rfloor - 1}\alpha(y_k - g_k^N(x_k))\left\langle c\sigma'(wx_k)\nabla_w f(c,w)\cdot x_k, \nu_k^N\right\rangle + O_p\left(N^{-1}\right). \qquad (2.6)\end{aligned}$$

3 Relative Compactness

In this section we prove that the family of processes $\{\mu^N, h^N\}_N$ is relatively compact. Section 3.1 proves compact containment. Section 3.2 proves regularity. Section 3.3 combines these results to prove relative compactness.

3.1 Compact Containment

We first establish a priori bounds for the parameters (C_k^i, W_k^i).

Lemma 3.1. For all $i \in \mathbb{N}$ and all k such that $k/N \leq T$,

$$\begin{aligned} |C_k^i| &< C < \infty \\ \mathbb{E}\left\|W_k^i\right\| &< C < \infty. \end{aligned}$$

Proof. The unimportant finite constant $C < \infty$ may change from line to line. We first observe that

$$\begin{aligned} |C_{k+1}^i| &\leq |C_k^i| + \alpha N^{-3/2} \left|y_k - g_k^N(x_k)\right| |\sigma(W_k^i x_k)| \\ &\leq |C_k^i| + \frac{C|y_k|}{N^{3/2}} + \frac{C}{N^2} \sum_{i=1}^N |C_k^i|, \end{aligned}$$

where the last inequality follows from the definition of $g_k^N(x)$ and the uniform boundedness assumption on $\sigma(\cdot)$.

Then, we subsequently obtain that

$$\begin{aligned} |C_k^i| &= |C_0^i| + \sum_{j=1}^k \left[|C_j^i| - |C_{j-1}^i| \right] \\ &\leq |C_0^i| + \sum_{j=1}^k \frac{C}{N^{3/2}} + \frac{C}{N^2} \sum_{j=1}^k \sum_{i=1}^N |C_{j-1}^i| \\ &\leq |C_0^i| + \frac{C}{\sqrt{N}} + \frac{C}{N^2} \sum_{j=1}^k \sum_{i=1}^N |C_{j-1}^i|. \end{aligned}$$

This implies that

$$\frac{1}{N} \sum_{i=1}^N |C_k^i| \leq \frac{1}{N} \sum_{i=1}^N |C_0^i| + \frac{C}{\sqrt{N}} + \frac{C}{N^2} \sum_{j=1}^k \sum_{i=1}^N |C_{j-1}^i|,$$

Let us now define $m_k^N = \frac{1}{N} \sum_{i=1}^N |C_k^i|$. Since the random variables C_0^i take values in a compact set, we have that $\frac{1}{N} \sum_{i=1}^N |C_0^i| + \frac{C}{\sqrt{N}} < C < \infty$. Then,

$$m_k^N \leq C + \frac{C}{N} \sum_{j=1}^k m_{j-1}^N.$$

By the discrete Gronwall lemma and using $k/N \leq T$,

$$m_k^N \leq C \exp\left(\frac{Ck}{N}\right) \leq C. \tag{3.1}$$

Note that the constants may depend on T.

We can now combine the bounds (3.1) and (3.1) to yield, for any $0 \leq k \leq TN$,

$$\begin{aligned} |C_k^i| &\leq |C_0^i| + \frac{C}{\sqrt{N}} + \frac{C}{N^2}\sum_{j=1}^{k} m_{j-1}^N \\ &\leq |C_0^i| + \frac{C}{\sqrt{N}} + \frac{C}{N^2}\sum_{j=1}^{k} C_2 \\ &\leq |C_0^i| + \frac{C}{\sqrt{N}} + \frac{C}{N} \\ &\leq C, \end{aligned} \tag{3.2}$$

where the last inequality follows from the random variables C_0^i taking values in a compact set.

Now, we turn to the bound for $\| W_k^i \|$. We start with the bound (using Young's inequality)

$$\begin{aligned} \| W_{k+1}^i \| \leq & \| W_k^i \| + \frac{C}{N^{3/2}} \left(|y_k| + \frac{1}{\sqrt{N}} \sum_{j=1}^{N} |C_k^j| \right) |C_k^i| |\sigma'(W_k^i x_k)| \| x_k \| \\ \leq & \| W_k^i \| + C \left(\frac{1}{N}|y_k|^2 + \frac{1}{N^2}\sum_{j=1}^{N} |C_k^j|^2 + \frac{1}{N}|C_k^i|^2 \| x_k \|^2 \right) \\ \leq & \| W_k^i \| + C \left(\frac{1}{N}|y_k|^2 + \frac{1}{N^2}\sum_{j=1}^{N} |C_k^j|^2 + \frac{1}{N}|C_k^i|^4 + \frac{1}{N} \| x_k \|^4 \right), \end{aligned}$$

for a constant $C < \infty$ that may change from line to line. Taking an expectation, using Assumption 1.1, the bound (3.2), and using the fact that $k/N \leq T$, we obtain

$$\mathbb{E} \| W_k^i \| \leq C < \infty,$$

for all $i \in \mathbb{N}$ and all k such that $k/N \leq T$, concluding the proof of the lemma.

□

Using the bounds from Lemma 3.1, we can now establish a bound for $g_k^N(x)$ for $x \in \mathcal{D}$.

Lemma 3.2. For all $i \in \mathbb{N}$, all k such that $k/N \leq T$, and any $x \in \mathcal{D}$,

$$\mathbb{E}\left[|g_k^N(x)|^2 \right] < C < \infty.$$

Proof. Recall equation (2.2), which describes the evolution of $g_k^N(x)$.

$$\begin{aligned} g_{k+1}^N(x) &= g_k^N(x) + \frac{\alpha}{N^2}\sum_{i=1}^{N} (y_k - g_k^N(x_k))\sigma(W_k^i x_k)\sigma(W_k^i x) \\ &+ \frac{\alpha}{N^2}\sum_{i=1}^{N} \sigma'(W_k^i x)(y_k - g_k^N(x_k))\sigma'(W_k^i x_k) x_k x \left(C_k^i\right)^2 + \frac{C}{N^{3/2}}. \end{aligned}$$

This leads to the bound

$$\begin{aligned} |g_{k+1}^N(x)| &\leq |g_k^N(x)| + \frac{\alpha}{N^2}\sum_{i=1}^{N} |y_k - g_k^N(x_k)| + \frac{\alpha}{N^2}\sum_{i=1}^{N} |y_k - g_k^N(x_k)| \left(C_k^i\right)^2 + \frac{C}{N^{-3/2}} \\ &\leq |g_k^N(x)| + \frac{C}{N}|g_k^N(x_k)| + \frac{C}{N}. \end{aligned}$$

We now square both sides of the above inequality.

$$\begin{aligned}
|g_{k+1}^N(x)|^2 &\leq \big(|g_k^N(x)| + \frac{C}{N}|g_k^N(x_k)| + \frac{C}{N}\big)^2 \\
&\leq |g_k^N(x)|^2 + 2|g_k^N(x)|(\frac{C}{N}|g_k^N(x_k)| + \frac{C}{N}) + (\frac{C}{N}|g_k^N(x_k)| + \frac{C}{N})^2 \\
&\leq |g_k^N(x)|^2 + \frac{C}{N}|g_k^N(x)|^2 + \frac{C}{N},
\end{aligned}$$

where the last line uses Young's inequality.

Therefore,

$$|g_{k+1}^N(x)|^2 - |g_k^N(x)|^2 \leq \frac{C}{N}|g_k^N(x_k)|^2 + \frac{C}{N}.$$

Then, using a telescoping series,

$$\begin{aligned}
|g_k^N(x)|^2 &= |g_0^N(x)|^2 + \sum_{j=1}^{k}\left(|g_j^N(x)|^2 - |g_{j-1}^N(x)|^2\right) \\
&\leq |g_0^N(x)|^2 + \sum_{j=1}^{k}\left(\frac{C}{N}|g_{j-1}^N(x_{j-1})|^2 + \frac{C}{N}\right) \\
&\leq |g_0^N(x)|^2 + C + \frac{C}{N}\sum_{j=1}^{k}|g_{j-1}^N(x_{j-1})|^2.
\end{aligned}$$

Taking expectations,

$$\mathbb{E}\Big[|g_k^N(x)|^2\Big] \leq \mathbb{E}\Big[|g_0^N(x)|^2\Big] + C + \frac{C}{N}\sum_{j=1}^{k}\mathbb{E}\Big[|g_{j-1}^N(x_{j-1})|^2\Big].$$

Taking advantage of the fact that x_j is sampled from a fixed dataset $\mathcal{D}$ of M data samples,

$$\mathbb{E}\Big[|g_k^N(x)|^2\Big] \leq \mathbb{E}\Big[|g_0^N(x)|^2\Big] + C + \frac{C}{N}\sum_{j=1}^{k}\sum_{x'\in\mathcal{D}}\mathbb{E}\Big[|g_{j-1}^N(x')|^2\Big], \tag{3.3}$$

and therefore

$$\begin{aligned}
\sum_{x\in\mathcal{D}}\mathbb{E}\Big[|g_k^N(x)|^2\Big] &\leq \sum_{x\in\mathcal{D}}\mathbb{E}\Big[|g_0^N(x)|^2\Big] + MC + \frac{CM}{N}\sum_{j=1}^{k}\sum_{x'\in\mathcal{D}}\mathbb{E}\Big[|g_{j-1}^N(x')|^2\Big] \\
&\leq \sum_{x\in\mathcal{D}}\mathbb{E}\Big[|g_0^N(x)|^2\Big] + C + \frac{C}{N}\sum_{j=1}^{k}\sum_{x\in\mathcal{D}}\mathbb{E}\Big[|g_{j-1}^N(x)|^2\Big].
\end{aligned} \tag{3.4}$$

Recall that

$$g_0^N(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}C_0^i\sigma(W_0^i x),$$

where (C_0^i, W_0^i) are i.i.d., mean-zero random variables. Then,

$$\begin{aligned}
\mathbb{E}\Big[|g_0^N(x)|^2\Big] &\leq \mathbb{E}\Big[\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}C_0^i\sigma(W_0^i x)\Big)^2\Big] \\
&\leq \frac{C}{N}\sum_{i=1}^{N}\mathbb{E}\Big[(C_0^i)^2\Big] \\
&\leq C.
\end{aligned}$$

Combining this bound with the bound (3.4) and using the discrete Gronwall lemma yields, for any $0 \leq k \leq TN$,

$$\sum_{x \in \mathcal{D}} \mathbb{E}\left[|g_k^N(x)|\right] \leq C.$$

Substituting this bound into equation (3.3) produces the desired bound

$$\mathbb{E}\left[|g_k^N(x)|^2\right] \leq C,$$

for any $0 \leq k \leq TN$.

□

We now prove compact containment for process $\{(\mu_t^N, h_t^N), t \in [0,T]\}_{N \in \mathbb{N}}$. Recall that $(\mu_t^N, h_t^N) \in D_E([0,T])$ where $E = \mathcal{M}(\mathbb{R}^{1+d}) \times \mathbb{R}^M$.

Lemma 3.3. For each $\eta > 0$, there is a compact subset $\mathcal{K}$ of E such that

$$\sup_{N \in \mathbb{N}, 0 \leq t \leq T} \mathbb{P}[(\mu_t^N, h_t^N) \notin \mathcal{K}] < \eta.$$

Proof. For each $L > 0$, define $K_L = [0, L]^{1+d}$. Then, we have that K_L is a compact subset of $\mathbb{R}^{1+d}$, and for each $t \geq 0$ and $N \in \mathbb{N}$,

$$\mathbb{E}\left[\mu_t^N(\mathbb{R}^{1+d} \setminus K_L)\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\left[|c^i_{\lfloor Nt \rfloor}| + \| w^i_{\lfloor Nt \rfloor} \| \geq L\right] \leq \frac{C}{L}.$$

where we have used Markov's inequality and the bounds from Lemma 3.1. We define the compact subsets of $\mathcal{M}(\mathbb{R}^{1+d})$

$$\hat{K}_L = \overline{\left\{\nu : \nu(\mathbb{R}^{1+d} \setminus K_{(L+j)^2}) < \frac{1}{\sqrt{L+j}} \text{ for all } j \in \mathbb{N}\right\}}$$

and we observe that

$$\begin{aligned}
\mathbb{P}\left\{\mu_t^N \notin \hat{K}_L\right] &\leq \sum_{j=1}^{\infty} \mathbb{P}\left[\mu_t^N(\mathbb{R}^{1+d} \setminus K_{(L+j)^2}) > \frac{1}{\sqrt{L+j}}\right] \leq \sum_{j=1}^{\infty} \frac{\mathbb{E}[\mu_t^N(\mathbb{R}^{1+d} \setminus K_{(L+j)^2})]}{1/\sqrt{L+j}} \\
&\leq \sum_{j=1}^{\infty} \frac{C}{(L+j)^2/\sqrt{L+j}} \leq \sum_{j=1}^{\infty} \frac{C}{(L+j)^{3/2}}.
\end{aligned}$$

Given that $\lim_{L \to \infty} \sum_{j=1}^{\infty} \frac{C}{(L+j)^{3/2}} = 0$, we have that, for each $\eta > 0$, there exists a compact set $\hat{K}_L$ such that

$$\sup_{N \in \mathbb{N}, 0 \leq t \leq T} \mathbb{P}[\mu_t^N \notin \hat{K}_L] < \frac{\eta}{2}.$$

Due to Lemma 3.2 and Markov's inequality, we also know that, for each $\eta > 0$, there exists a compact set $U = [-B, B]^M$ such that

$$\sup_{N \in \mathbb{N}, 0 \leq t \leq T} \mathbb{P}[h_t^N \notin U] < \frac{\eta}{2}.$$

Therefore, for each $\eta > 0$, there exists a compact set $\hat{K}_L \times [-B, B]^M \subset E$ such that

$$\sup_{N \in \mathbb{N}, 0 \leq t \leq T} \mathbb{P}[(\mu_t^N, h_t^N) \notin \hat{K}_L \times [-B, B]^M] < \eta.$$

□

3.2 Regularity

We now establish regularity of the process μ^N in $D_{\mathcal{M}(\mathbb{R}^{1+d})}([0,T])$. Define the function $q(z_1,z_2) = \min\{|z_1-z_2|,1\}$ where $z_1,z_2 \in \mathbb{R}$.

Lemma 3.4. Let $f \in C_b^2(\mathbb{R}^{1+d})$. For any $\delta \in (0,1)$, there is a constant $C<\infty$ such that for $0 \le u \le \delta$, $0 \le v \le \delta \wedge t$, $t \in [0,T]$,

$$\mathbb{E}\left[q(\langle f,\mu^N_{t+u}\rangle, \langle f,\mu^N_t\rangle)q(\langle f,\mu^N_t\rangle, \langle f,\mu^N_{t-v}\rangle)\middle|\mathcal{F}^N_t\right] \le C\delta + \frac{C}{N^{3/2}}.$$

Proof. We start by noticing that a Taylor expansion gives for $0 \le s \le t \le T$

$$\begin{aligned}
|\langle f,\mu^N_t\rangle - \langle f,\mu^N_s\rangle| &= |\left\langle f,\nu^N_{\lfloor Nt\rfloor}\right\rangle - \left\langle f,\nu^N_{\lfloor Ns\rfloor}\right\rangle| \\
&\le \frac{1}{N}\sum_{i=1}^N |f(C^i_{\lfloor Nt\rfloor},W^i_{\lfloor Nt\rfloor}) - f(C^i_{\lfloor Ns\rfloor},W^i_{\lfloor Ns\rfloor})| \\
&\le \frac{1}{N}\sum_{i=1}^N |\partial_c f(\bar C^i_{\lfloor Nt\rfloor},\bar W^i_{\lfloor Nt\rfloor})||C^i_{\lfloor Nt\rfloor} - C^i_{\lfloor Ns\rfloor}| \\
&+ \frac{1}{N}\sum_{i=1}^N \| \nabla_w f(\bar C^i_{\lfloor Nt\rfloor},\bar W^i_{\lfloor Nt\rfloor}) \|\| W^i_{\lfloor Nt\rfloor} - W^i_{\lfloor Ns\rfloor} \|, \qquad (3.5)
\end{aligned}$$

for points $\bar C^i, \bar W^i$ in the segments connecting $C^i_{\lfloor Ns\rfloor}$ with $C^i_{\lfloor Nt\rfloor}$ and $W^i_{\lfloor Ns\rfloor}$ with $W^i_{\lfloor Nt\rfloor}$, respectively.

Let's now establish a bound on $|C^i_{\lfloor Nt\rfloor} - C^i_{\lfloor Ns\rfloor}|$ for $s<t\le T$ with $0<t-s\le\delta<1$.

$$\begin{aligned}
\mathbb{E}\left[|C^i_{\lfloor Nt\rfloor} - C^i_{\lfloor Ns\rfloor}|\middle|\mathcal{F}^N_s\right] &= \mathbb{E}\left[|\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}(C^i_{k+1}-C^i_k)|\middle|\mathcal{F}^N_s\right] \\
&\le \mathbb{E}\left[\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}|\alpha(y_k - g^N_k(x_k))\frac{1}{N^{3/2}}\sigma(W^i_k x_k)|\middle|\mathcal{F}^N_s\right] \\
&\le \frac{1}{N^{3/2}}\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1} C \le \frac{C}{\sqrt{N}}(t-s) + \frac{C}{N^{3/2}} \\
&\le \frac{C}{\sqrt{N}}\delta + \frac{C}{N^{3/2}}, \qquad (3.6)
\end{aligned}$$

where Assumption 1.1 was used as well as the bounds from Lemmas 3.1 and 3.2.

Let's now establish a bound on $\| W^i_{\lfloor Nt\rfloor} - W^i_{\lfloor Ns\rfloor} \|$ for $s<t\le T$ with $0<t-s\le\delta<1$. We obtain

$$\begin{aligned}
\mathbb{E}\left[\| W^i_{\lfloor Nt\rfloor} - W^i_{\lfloor Ns\rfloor} \|\middle|\mathcal{F}^N_s\right] &= \mathbb{E}\left[\| \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}(W^i_{k+1}-W^i_k) \|\middle|\mathcal{F}^N_s\right] \\
&\le \mathbb{E}\left[\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\| \alpha(y_k - g^N_k(x_k))\frac{1}{N^{3/2}}C^i_k\sigma'(W^i_k\cdot x_k)x_k \|\middle|\mathcal{F}^N_s\right] \\
&\le \frac{1}{N^{3/2}}\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1} C \\
&\le \frac{C}{\sqrt{N}}(t-s) + \frac{C}{N} \le \frac{C}{\sqrt{N}}\delta + \frac{C}{N^{3/2}},
\end{aligned}$$

where we have again used the bounds from Lemmas 3.1 and 3.2.

Now, we return to equation (3.5). Due to Lemma 3.1, the quantities $(\bar{c}^i_{\lfloor Nt\rfloor}, \bar{w}^i_{\lfloor Nt\rfloor})$ are bounded in expectation for $0 < s < t \le T$. Therefore, for $0 < s < t \le T$ with $0 < t - s \le \delta < 1$

$$\mathbb{E}\left[\left|\left\langle f, \mu^N_t\right\rangle - \left\langle f, \mu^N_s\right\rangle\right| \middle| \mathcal{F}^N_s\right] \le C\delta + \frac{C}{N^{3/2}}.$$

where $C < \infty$ is some unimportant constant. Then, the statement of the Lemma follows. □

We next establish regularity of the process h^N_t in $D_{\mathbb{R}^M}([0,T])$. For the purposes of the following lemma, let the function $q(z_1, z_2) = \min\{\|z_1 - z_2\|, 1\}$ where $z_1, z_2 \in \mathbb{R}^M$ and $\|z\| = |z_1| + \cdots + |z_M|$.

Lemma 3.5. For any $\delta \in (0,1)$, there is a constant $C < \infty$ such that for $0 \le u \le \delta < 1$, $0 \le v \le \delta \wedge t$, $t \in [0,T]$,

$$\mathbb{E}\left[q(h^N_{t+u}, h^N_t) q(h^N_t, h^N_{t-v}) \middle| \mathcal{F}^N_t\right] \le C\delta + \frac{C}{N}.$$

Proof. Recall that

$$g^N_{k+1}(x) = g^N_k(x) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(C^i_{k+1} - C^i_k)\sigma(W^i_{k+1}x) + \sigma'(W^{i,*}_k x)(W^i_{k+1} - W^i_k)xC^i_k \right).$$

Therefore,

$$\begin{aligned} h^N_t(x) - h^N_s(x) &= g_{\lfloor Nt\rfloor}(x) - g_{\lfloor Ns\rfloor}(x) \\ &= \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor} (g^N_{k+1}(x) - g^N_k(x)) \\ &= \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor} \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(C^i_{k+1} - C^i_k)\sigma(W^i_{k+1}x) + \sigma'(W^{i,*}_k x)(W^i_{k+1} - W^i_k)xC^i_k \right). \end{aligned}$$

This yields the bound

$$\begin{aligned} |h^N_t(x) - h^N_s(x)| &\le \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor} |g^N_{k+1}(x) - g^N_k(x)| \\ &\le \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor} \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(|C^i_{k+1} - C^i_k| + \left\|W^i_{k+1} - W^i_k\right\| \right), \end{aligned}$$

where we have used the boundedness of $\sigma'(\cdot)$ (from Assumption 1.1) and the bounds from Lemma 3.1.

Taking expectations,

$$\mathbb{E}\left[|h^N_t(x) - h^N_s(x)| \middle| \mathcal{F}^N_s \right] \le \frac{1}{\sqrt{N}} \sum_{i=1}^N \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor} \mathbb{E}\left[|C^i_{k+1} - C^i_k| + \left\|W^i_{k+1} - W^i_k\right\| \middle| \mathcal{F}^N_s \right].$$

Using the bounds (3.6) and (3.7),

$$\begin{aligned} \mathbb{E}\left[|h^N_t(x) - h^N_s(x)| \middle| \mathcal{F}^N_s \right] &\le \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{C}{\sqrt{N}}(t-s) + \frac{C}{N^{3/2}} \right) \\ &= C(t-s) + \frac{C}{N}. \end{aligned} \tag{3.7}$$

The bound (3.7) holds for each $x \in \mathcal{D}$. Therefore,

$$\mathbb{E}\left[\left\| h_t^N - h_s^N \right\| \middle| \mathcal{F}_s^N \right] \leq C(t-s) + \frac{C}{N}.$$

The statement of the Lemma then follows.

□

3.3 Combining our results to prove relative compactness

Lemma 3.6. The family of processes $\{\mu^N, h^N\}_{N\in\mathbb{N}}$ is relatively compact in $D_E([0,T])$.

Proof. Combining Lemmas 3.3 and 3.4, and Theorem 8.6 of Chapter 3 of [7] proves that $\{\mu^N\}_{N\in\mathbb{N}}$ is relatively compact in $D_{\mathcal{M}(\mathbb{R}^{1+d})}([0,T])$. (See also Remark 8.7 B of Chapter 3 of [7] regarding replacing $\sup_N$ with $\lim_N$ in the regularity condition B of Theorem 8.6.)

Similarly, combining Lemmas 3.3 and 3.5 proves that $\{h^N\}_{N\in\mathbb{N}}$ is relatively compact in $D_{\mathbb{R}^M}([0,T])$.

Since relative compactness is equivalent to tightness, we have that the probability measures of the family of processes $\{\mu^N\}_{N\in\mathbb{N}}$ are tight. Similarly, we have that the probability measures of the family of process $\{h^N\}_{N\in\mathbb{N}}$ are tight. Therefore, $\{\mu^N, h^N\}_{N\in\mathbb{N}}$ is tight. Then, $\{\mu^N, h^N\}_{N\in\mathbb{N}}$ is also relatively compact.

□

4 Identification of the Limit

Let π^N be the probability measure of a convergent subsequence of $\left(\mu^N, h^N\right)_{0\leq t\leq T}$. Each π^N takes values in the set of probability measures $\mathcal{M}\big(D_E([0,T])\big)$. Relative compactness, proven in Section 3, implies that there is a subsequence π^{N_k} which weakly converges. We must prove that any limit point π of a convergent subsequence π^{N_k} will satisfy the evolution equation (1.3).

Lemma 4.1. Let π^{N_k} be a convergent subsequence with a limit point π. Then, π is a Dirac measure concentrated on $(\mu, h) \in D_E([0,T])$ and (μ, h) satisfies equation (1.3).

Proof. We define a map $F(\mu, h) : D_E([0,T]) \to \mathbb{R}_+$ for each $t \in [0,T]$, $f \in C_b^2(\mathbb{R}^{1+d})$, $g_1, \cdots, g_p \in C_b(\mathbb{R}^{1+d})$, $q_1, \cdots, q_p \in C_b(\mathbb{R}^M)$, and $0 \leq s_1 < \cdots < s_p \leq t$.

$$\begin{aligned}
F(\mu,h) &= \left| \left(\langle f, \mu_t\rangle - \langle f, \mu_0\rangle\right) \times \langle g_1, \mu_{s_1}\rangle \times \cdots \times \left\langle g_p, \mu_{s_p}\right\rangle \right| \\
&+ \sum_{x\in\mathcal{D}} \left| \left(h_t(x) - h_0(x) - \alpha \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} (y - h_s(x')) \left\langle \sigma(wx)\sigma(wx'), \mu_s\right\rangle \pi(dx', dy) ds \right.\right. \\
&- \left.\left. \alpha \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} (y - h_s(x')) \left\langle c^2\sigma'(wx')\sigma'(wx)xx', \mu_s\right\rangle \pi(dx, dy) ds \right) \times q_1(h_{s_1}) \times \cdots \times q_p(h_{s_p}) \right|
\end{aligned}$$

Then, using equations (2.4) and (2.6), we obtain

$$\begin{aligned}
\mathbb{E}_{\pi^N}[F(\mu,h)] &= \mathbb{E}[F(\mu^N, h^N)] \\
&= \mathbb{E}\left| \mathcal{O}_p(N^{-1/2}) \times \prod_{i=1}^p \left\langle g_i, \mu_{s_i}^N \right\rangle \right| \\
&+ \mathbb{E}\left| (M_t^{N,1} + M_t^{N,2} + \mathcal{O}_p(N^{-1/2})) \times \prod_{i=1}^p q_i(h_{s_i}^N) \right| \\
&\leq C\left(\mathbb{E}\left[|M^{1,N}(t)|^2 \right]^{\frac{1}{2}} + \mathbb{E}\left[|M^{2,N}(t)|^2 \right]^{\frac{1}{2}} \right) + O(N^{-1/2}) \\
&\leq C\left(\frac{1}{\sqrt{N}} + \frac{1}{N} \right),
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality.

Therefore,

$$\lim_{N\to\infty} \mathbb{E}_{\pi^N}[F(\mu,h)] = 0.$$

Since $F(\cdot)$ is continuous and $F(\mu^N)$ is uniformly bounded (due to the uniform boundedness results of Section 3),

$$\mathbb{E}_{\pi}[F(\mu,h)] = 0.$$

Since this holds for each $t \in [0,T]$, $f \in C_b^2(\mathbb{R}^{1+d})$ and $g_1, \cdots, g_p, q_1, \cdots, q_p \in C_b(\mathbb{R}^{1+d})$, (μ, h) satisfies the evolution equation (1.3). □

5 Proof of Convergence

We now combine the previous results of Sections 3 and 4 to prove Theorem 1.2. Let π^N be the probability measure corresponding to (μ^N, h^N). Each π^N takes values in the set of probability measures $\mathcal{M}\big(D_E([0,T])\big)$. Relative compactness, proven in Section 3, implies that every subsequence π^{N_k} has a further sub-sequence $\pi^{N_{k_m}}$ which weakly converges. Section 4 proves that any limit point π of $\pi^{N_{k_m}}$ will satisfy the evolution equation (1.3). Equation (1.3) is a finite-dimensional, linear equation and therefore has a unique solution. Therefore, by Prokhorov's Theorem, π^N weakly converges to π, where π is the distribution of (μ, h), the unique solution of (1.3). That is, (μ^N, h^N) converges in distribution to (μ, h).

6 Proof of Corollary 1.4

This section proves that under reasonable hyperparameter choices, the matrix A in the limit equation will be positive definite.

Proof. We first show that A is equivalent to the covariance matrix of the random variables $U = \Big(U(x^{(1)}), \ldots, U(x^{(M)})\Big)$, which are defined as

$$\begin{aligned} U(x) &= \sqrt{\frac{\alpha}{M}}\sigma(Wx) + \sqrt{\frac{\alpha}{M}}C\sigma'(Wx)x \\ &= \sqrt{\frac{\alpha}{M}}\sigma(Wx), \end{aligned} \tag{6.1}$$

where $(W, C) \sim \mu_0$ and $x \in \mathcal{D}$. In particular, $W_j \sim \mathcal{N}(0,1)$, W_j is independent of W_k for $k \neq j$, and $C = 0$.

Then,

$$\mathbb{E}\Big[U(x)U(x')\Big] = \mathbb{E}\Big[\frac{\alpha}{M}\sigma(Wx)\sigma(Wx')\Big] = A_{x,x'}. \tag{6.2}$$

To prove that A is positive definite, we need to show that $z^\top Az > 0$ for every non-zero $z \in \mathbb{R}^M$.

$$\begin{aligned} z^\top Az &= z^\top \mathbb{E}\Big[UU^\top\Big]z \\ &= \mathbb{E}\Big[(z^\top U)^2\Big] \\ &= \frac{\alpha}{M}\mathbb{E}\Big[\Big(\sum_{i=1}^M z_i\sigma(x^{(i)}\cdot W)\Big)^2\Big]. \end{aligned} \tag{6.3}$$

The functions $\sigma(x^{(i)}\cdot W)$ are linearly independent since the $x^{(i)}$ are distinct (by Corollary 4.3 of [1]). Therefore, for each non-zero z, there exists a point w^* such that

$$\sum_{i=1}^M z_i\sigma(x^{(i)}\cdot w^*) \neq 0.$$

Consequently, there exists an $\epsilon > 0$ such that

$$\left(\sum_{i=1}^{M} z_i \sigma(x^{(i)} \cdot w^*) \right)^2 > \epsilon.$$

Since $\sigma(\cdot)$ is a continuous function, there exists a set $B = \{w : \|w - w^*\| < \eta\}$ for some $\eta > 0$ such that for $w \in B$

$$\left(\sum_{i=1}^{M} z_i \sigma(x^{(i)} \cdot w) \right)^2 > \frac{\epsilon}{2}.$$

Then,

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{i=1}^{M} z_i \sigma(x^{(i)} \cdot W) \right)^2 \right] &\geq \mathbb{E}\left[\left(\sum_{i=1}^{M} z_i \sigma(x^{(i)} \cdot W) \right)^2 \mathbf{1}_{W \in B} \right] \\ &\geq \mathbb{E}\left[\frac{\epsilon}{2} \mathbf{1}_{W \in B} \right] \\ &= \frac{C\epsilon}{2}, \end{aligned}$$

where $C > 0$.

Therefore, for every non-zero $z \in \mathbb{R}^M$,

$$z^\top A z > 0,$$

and A is positive definite, concluding the proof of the Corollary. □

References

[1] Yoshifusa Ito. Nonlinearity creates linear independence. Advances in Computational Mathematics, 5: 189-203, 1996.

[2] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249-256. 2010.

[3] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[4] S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient Descent Provably Optimizes Over-Parameterized Neural Networks. ICLR, 2019.

[5] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks. arXiv: 1811.08888, 2018.

[6] A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.

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

[8] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359-366, 1989.

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

[10] C. Kuan and K. Hornik. Convergence of learning algorithms with constant learning rates. IEEE Transactions on Neural Networks, 2(5), 484-489, 1991.

[11] S. Mei, A. Montanari, and P. Nguyen. A mean field view of the landscape of two-layer neural networks Proceedings of the National Academy of Sciences, 115 (33) E7665-E767, 2018.

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

[13] J. Sirignano and K. Spiliopoulos. Mean Field Analysis of Neural Networks: A Central Limit Theorem. Stochastic Processes and their Applications, 2019.

[14] J. Sirignano and K. Spiliopoulos. Mean Field Analysis of Deep Neural Networks. arXiv:1903.04440, 2019.

[15] 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. arXiv:1805.00915, 2018.