Global optimality of Elman-type RNNs in the mean-field regime

Andrea Agazzi [1 2] **Jianfeng Lu** [2 3] **Sayan Mukherjee** [2 4 5 6]

Abstract

We analyze Elman-type Recurrent Reural Networks (RNNs) and their training in the mean-field regime. Specifically, we show convergence of gradient descent training dynamics of the RNN to the corresponding mean-field formulation in the large width limit. We also show that the fixed points of the limiting infinite-width dynamics are globally optimal, under some assumptions on the initialization of the weights. Our results establish optimality for feature-learning with wide RNNs in the mean-field regime.

1. Introduction

During the last decade, artificial intelligence and in particular deep leaning have achieved a significant series of groundbreaking successes, partly due to the unprecedented increase of data and computational power at our disposal. Notably, the range of disciplines that have recently been revolutionized by machine learning is virtually unlimited: from medicine (Rajkomar et al., 2019) to finance (Dixon et al., 2020), from games (Silver et al., 2016; Vinyals et al., 2019) to image analysis (LeCun et al., 1989), to the point where almost no domain has remained unaltered by the emergence of these technologies.

This revolution would have been unthinkable without the advent of deep neural networks. This extremely flexible family of function approximators has outperformed classical methods in almost every domain where it has been applied. A discipline that has been profoundly revolutionized by these models is the analysis of time series and, more generally, the problem of learning dynamical systems. For instance, Recurrent Neural Networks (RNNs) (Jordan, 1997; Rumelhart et al., 1985) and more specifically Long-Short Term Memory (LSTM) RNNs (Hochreiter & Schmidhuber, 1997; Gers et al., 2000) and Gated Recurrent Units (Cho et al., 2014) have dramatically increased the predictive performance of machine learning in this context. These models take as input temporal sequences of data and act iteratively on the elements of such sequences, storing the information about previous timepoints into the hidden state of the network. This structure allows to learn datasets with strong time-correlations using relatively few parameters, and has provided benchmarks for state-of-the-art time-series learning algorithms for over a decade. However, despite the groundbreaking success of these models in practice, the theoretical underpinnings of such success remain elusive to the computer science community. More specifically, many questions about the theoretical reasons for the performance of these models applied far into the overparametrized regime, such as for example explanations for their optimal behavior and their generalization error, remain open.

Only recently, a theory of neural network learning has started to emerge in the context of wide, single-layer neural networks. The two main theoretical frameworks are based on either understanding mean-field training dynamics (Chizat & Bach, 2018a; Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2018; Wojtowytsch, 2020; Sirignano & Spiliopoulos, 2021; Agazzi & Lu, 2020; Chizat & Bach, 2020) or based on linearized dynamics in the overparametrized regime (Jacot et al., 2018; Chizat & Bach, 2018b; Ghorbani et al., 2021). These two frameworks provide contrasting explanations for the success of neural networks. On one hand the linearized dynamics gives strong convergence guarantees for the training process but fails to explain the feature-learning properties of neural networks. On the other hand the the mean-field framework is better at accurately capturing the highly nonlinear dynamics arguably resulting in feature-learning (Ghorbani et al., 2019), but the resulting training process is typically harder to analyze. Consequently, it remains a challenge to extend the mean-field results listed above to more realistic structures such as RNNs.

This paper aims to extend the mean-field framework to Elman-type RNNs. Specifically, we aim to establish optimality of the fixed points of the training dynamics for wide RNNs trained with classical gradient descent. This

[1]Department of Mathematics, University of Pisa, Pisa (IT) [2]Department of Mathematics, Duke University, Durham, NC (USA) [3]Department of Physics and Department of Chemistry, Duke University, Durham, NC (USA) [4]Department of Statistical Science, Department of Computer Science, Department of Biostatistics & Bioinformatics, Duke University, Durham, NC [5]Center for Scalable Data Analytics and Artificial Intelligence, Universität Leipzig, Leipzig (DE) [6]Max Planck Institute for Mathematics in the Sciences, Leipzig, DE. Correspondence to: Andrea Agazzi <andrea.agazzi@unipi.it>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

provides an explanation of the outstanding performance of these models in a certain idealized regime.

1.1. Previous results

The training dynamics of neural networks in the mean-field infinite width limit was pioneered by the series of papers (Mei et al., 2018; Chizat & Bach, 2018a; Rotskoff & Vanden-Eijnden, 2018; Sirignano & Spiliopoulos, 2021). Here, the authors proved that the training dynamics of infinitely wide, single layer neural networks in the mean-field regime can be studied by representing the parametric state of the network as a probability distribution in the space of weights. Using this representation it was possible to prove that the limiting points of the training dynamics are global optimizers of the loss function.

These ideas have been extended to the reinforcement learning setting (Sirignano & Spiliopoulos, 2022; Agazzi & Lu, 2020; 2022), to non-differeniable network nonlinearities such as ReLU (Wojtowytsch, 2020) and to the deep ResNet architecture (Lu et al., 2020). A recent series of papers (Nguyen & Pham, 2020; Pham & Nguyen, 2020) has bypassed the difficulties related to the representation of the network state as a distribution by introducing the *neuronal embedding* framework. This framework allows for the investigation of the mean-field dynamics of deep, purely feedforward neural networks. However, none of these results can be directly applied to the RNN setting: the presence of weight-sharing in the RNN structure and the interaction of the unrolled network with the input violate fundamental assumptions in these analyses.

The performance of RNNs in the infinite width limit was studied in (Alemohammad et al., 2021). Here, the authors explored the performance of the network in the so-called Neural Tangent Kernel (NTK) regime, arising under a particular scaling of the weights at initialization. This scaling linearizes the training dynamics of the network, which behaves essentially like a kernel method. It is therefore widely believed that in this regime feature learning is not possible. A recent paper (Yang, 2020) considers a similar scaling to (Alemohammad et al., 2021) in combination with more general architectures. In contrast to these works, the mean-field scaling we consider retains the nonlinear training dynamics.

Finally, a seemingly related result about mean-field theory for RNNs has been presented in (Chen et al., 2018). That work, however, uses *dynamical* mean-field theory to explain the role of gating in RNN architectures and thus our proof techniques differ greatly from that paper. The scope of the results is also significantly different, as their results aim to explore forward propagation of signal through vanilla RNNs, and do not aim to establish optimality of the fixed points after training.

1.2. Contributions

This paper adapts the neuronal embedding analysis framework developed in (Nguyen & Pham, 2020; Pham & Nguyen, 2020) to unrolled Elman-type RNNs (Elman, 1990). We prove optimality of the fixed points of the training dynamics in the mean-field regime under some assumptions on the expressive power of the network at initialization. Specifically we prove:

1. Convergence of the dynamics of the finite-width RNN to its infinite-width limit. To do so we adapt the coupling formulation presented in (Nguyen & Pham, 2020) in the context of fully-connected feedforward networks to the RNN framework, thereby extending it to networks with weight-sharing.

2. Gradient descent trains these networks to optimal fixed points given infinite training time. This optimality result holds in the feature-learning regime, as opposed to previous results that hold in the NTK regime.

3. To prove the above results, we show universal approximation for deep neural networks with uniformly bounded hidden weights. This result extends classical universal approximation theorems, where weights are critically assumed to be in a vector space and, as such, to be unbounded.

A standard initialization assumption in feedforward neural networks, for example 3-layer networks, with a large number of nodes is to initialize the weights randomly and independently. In this paper, we further observe that feedback in an RNN requires stronger assumptions on the weights of the network at initialization to achieve a comparable level of expressivity as a 3-layer feedforward network. We examine this issue in some detail through our analysis.

The paper is organized as follows: in Section 2 we introduce the notation and the model being investigated, together with its mean-field limit. Then, in Section 3 we outline our main results. The results are exemplified with some numerical experiments in Section 4, and conclusions follow in Section 5. The proofs of our main theorems are given in the appendix.

2. Notation

2.1. Predictors

To put the data-generation process in an abstract framework for dynamical systems, we consider as predictors subsets of a bi-infinite observation sequence $\mathbf{x} \in (\mathbb{R}^d)^{\mathbb{Z}}$. For a given subshift $\mathcal{T} : (\mathbb{R}^d)^{\mathbb{Z}} \to (\mathbb{R}^d)^{\mathbb{Z}}$, we generate the elements of $\mathbf{x}$ as $\mathbf{x}_{k+1} = \mathcal{T}(\mathbf{x})_k$. We make the following assumption on the underlying dynamical system

Assumption 1. *There exists a continuous function $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$ for all $k \in \mathbb{Z}$. We further assume that this map is uniquely ergodic (upon* *possibly restricting it to a forward invariant set $\mathbb{X}$) and that the corresponding invariant measure has finite fourth moments. For the definition of unique ergodicity, see (Katok & Hasselblatt, 1995).*

We denote by $\Pi^0 : (\mathbb{R}^d)^{\mathbb{Z}} \to \mathbb{R}^d$ the projection of a bi-infinite sequence on its 0-th element. We further define $\nu \in \mathcal{M}_+^1(\mathbb{R}^{\mathbb{Z}})$ as the invariant measure of the map $\mathcal{T}$, which exists and is unique by the above assumption. The marginal $\nu_0 \in \mathcal{M}_+^1(\mathbb{R})$ on the 0-th component of ν is the invariant measure of T, and $\Pi^0_{\#}\nu = \nu_0$.

2.2. Loss function

We assume that we have access to an infinite-length sample from the invariant measure ν, from a dynamical system satisfying Assumption 1 to train the RNN. Our objective is to learn a map $F^* : (\mathbb{R}^d)^{\mathbb{N}} \to \mathbb{R}$ from sequences of arbitrary length to reals. We restrict our attention to functions with a fixed, finite memory $L \in \mathbb{N}$.

Assumption 2. *The function F^* only depends on $\{\mathbf{x}_{-L}, \ldots, \mathbf{x}_0\}$, for a fixed $L \in \mathbb{N}$.*

Our objective is to learn an estimate of F^* by minimizing the sample Mean Squared Error (MSE) between the target function F^* and a parametric family of estimators $\{F(\,\cdot\,; W)\}_W$ indexed by the parameter vector W on a sample of length-L trajectories of size m. In other words, we aim to find the minimizer $\hat{F} \in \{F(\,\cdot\,; W)\}_W$ of the empirical risk

$$\mathcal{L}_m(F^*, \hat{F}) := \frac{1}{m}\sum_{k=1}^{m} \frac{1}{2}(F^*(\mathcal{T}^k(\mathbf{x})) - \hat{F}(\mathcal{T}^k(\mathbf{x}))^2.$$

In the large sample limit $M \to \infty$, the above loss function can be rewritten as the population risk

$$\begin{aligned}\mathcal{L}(W) &= \lim_{m\to\infty} \mathcal{L}_m(F^*(\,\cdot\,), \hat{F}(\,\cdot\,; W)) \\ &= \frac{1}{2}\int (F^*(\mathbf{x}) - \hat{F}(\mathbf{x}; W))^2 \nu(\mathrm{d}\mathbf{x})\,,\end{aligned} \tag{2.1}$$

expressed above as a function of the parameters of the estimator. While our analysis extends to more general loss functions, for concreteness and ease of exposition we restrict our discussion to the MSE.

2.3. RNN structure

The family of models we consider are Elman-type Recurrent Neural Networks of hidden width $n \in \mathbb{N}$. Such a neural network can be written as

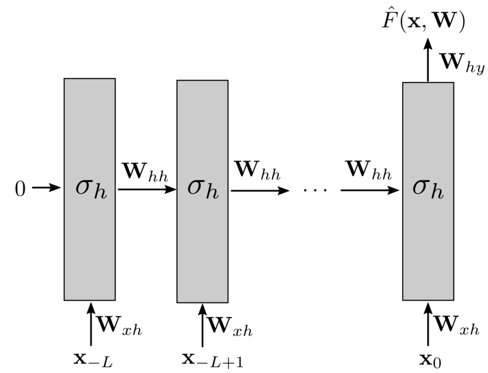


Figure 1: The many-to-one structure of the Elman-type RNN.

$$\begin{aligned}\hat{F}(\mathbf{x}; \mathbf{W}) &= \mathbf{H}_{hy}(\mathbf{x}) \\ \mathbf{H}_{hy}(\mathbf{x}) &= \frac{1}{n}\mathbf{W}_{hy}\sigma_h(\mathbf{H}_{hh}(\mathbf{x}, 0) + \mathbf{H}_{xh}(\mathbf{x}_0)) \\ \mathbf{H}_{hh}(\mathbf{x}, k) &= \frac{1}{n}\mathbf{W}_{hh}\sigma_h(\mathbf{H}_{hh}(\mathbf{x}, k+1) + \mathbf{H}_{xh}(\mathbf{x}_{-(k+1)})) \\ \mathbf{H}_{hh}(\mathbf{x}, L) &= 0 \\ \mathbf{H}_{xh}(\mathbf{x}_k) &= \mathbf{W}_{xh} \cdot \mathbf{x}_k\end{aligned} \tag{2.2}$$

where we assume that $\mathbf{x}_k \in \mathbb{R}^d$, $\mathbf{W}_{xh} \in \mathbb{R}^{n\times d}$, $\mathbf{W}_{hy} \in \mathbb{R}^n$, $\mathbf{W}_{hh} \in \mathbb{R}^{n\times n}$ and the activation function $\sigma_h : \mathbb{R} \to \mathbb{R}$ is applied component-wise. The structure of the network is represented in Fig. 1.

To investigate the convergence properties of RNNs as $n \to \infty$, we will apply the neuronal embedding formalism from (Nguyen & Pham, 2020; Pham & Nguyen, 2020). This formalism lifts the labeling of the neurons of the network to an abstract probability space $(\Omega_h, \mathcal{F}_h, P_h)$, and the neural network weights are interpreted as a function of these abstract indices. This lifting allows for the representation of any network as a specific choice of labelings, and equivalent relabelings of the neural network weights are different realizations of an abstract (random) labeling process. In this formalism, the weight functions can then be written as

$$W_{xh}(\vartheta) \in \mathbb{R}^d \qquad W_{hh}(\vartheta, \vartheta') \in \mathbb{R} \qquad W_{hy}(\vartheta') \in \mathbb{R}$$

for $\vartheta, \vartheta' \in \Omega_h$. A precise definition of ϑ, ϑ' and of the coupling procedure to identify the neuronal embedding with the infinite width limit of the network (2.2) is given in the next section. The mean-field representation is the continuous version of the network introduced above, representing matrix multiplications as integral kernels, and can be written as

$$\begin{aligned}
\hat{F}(\mathbf{x};W) =& H_{hy}(\mathbf{x}) \\
H_{hy}(\mathbf{x}) =& \int W_{hy}(\vartheta)\sigma_h(H_{hh}(\vartheta;\mathbf{x},0) \\
&+ H_{xh}(\vartheta;\mathbf{x}_0))P_h(\mathrm{d}\vartheta) \\
H_{hh}(\vartheta;\mathbf{x},k) =& \int W_{hh}(\vartheta,\vartheta')\sigma_h(H_{hh}(\vartheta';\mathbf{x},k+1) \\
&+ H_{xh}(\vartheta';\mathbf{x}_{-(k+1)}))P_h(\mathrm{d}\vartheta') \\
H_{hh}(\vartheta;\mathbf{x},L) \equiv& 0 \\
H_{xh}(\vartheta;\mathbf{x}_k) =& W_{xh}(\vartheta)\cdot\mathbf{x}_k
\end{aligned} \tag{2.3}$$

As the next example shows, any finite-width RNN $\hat{F}(\mathbf{W};\mathbf{x})$ can be *embedded* into the mean-field representation.

Example 2.1. *(Finite-width RNN) For any choice of parameters $\mathbf{W} = \{\mathbf{W}_{xh}, \mathbf{W}_{hh}, \mathbf{W}_{hy}\}$ for a width-n network for $n < \infty$ and assuming $d = 1$ we can set $\Omega_h := \{1, 2, \dots, n\}$. The measure P_h can be chosen as the uniform measure on $\{1, 2, \dots, n\}$. Then, it is readily seen that, setting $W_{hh}(i,j) := (\mathbf{W}_{hh})_{ij}$, $W_{xh}(i) := (\mathbf{W}_{xh})_i$ and $W_{hy}(j) := (\mathbf{W}_{hy})_j$ for $i, j \in \{1, \dots, n\}$ we have that (2.3) gives the same output as (2.2).*

2.4. Initialization and Coupling procedure

We now introduce the coupling procedure that connects the evolution of finite-width neural networks (2.2) to their mean-field representation (2.3). This coupling procedure is performed *at initialization*, *i.e.*, before training starts. We will respectively denote the weights of the finite-width network and of the mean-field limit at initialization by $\mathbf{W}^0$ and W^0. Instrumental to introducing the coupling procedure between the finite-width and the infinite-width neural network is the notion of *neuronal embedding*. Given a family I of initialization laws indexed by the width n of the hidden layer,

$$I = \{\varrho_n \,:\, \varrho_n \text{ is the law of } \mathbf{W}^0 \text{ for a network of width } n\}$$

we consider the parameters $\mathbf{W}^0$ of the width-n network as samples from the corresponding distribution $\varrho_n \in I$.

We call (Ω_h, P_h, W) a neuronal embedding for the neural network with initialization laws in I if for every $\varrho_n \in I$ there exists a sampling rule $\bar{P}_n$ such that

1. $\bar{P}_n$ is a distribution on Ω_h^n (not necessarily a product distribution) with marginals given by P_h
2. The mean-field weights $W = (W_{xh}, W_{hh}, W_{hy})$ are such that, if $(\vartheta(j))_j \sim \bar{P}_n$, then for every n with $i, j \in \{1, \dots, n\}$:

   $$\mathrm{Law}(W_{xh}(\vartheta(i)), W_{hh}(\vartheta(i), \vartheta(j)), W_{hy}(\vartheta(j))) = \varrho_n.$$

The above definition decomposes the concept of neural network weights to two parts: the first part is a deterministic function of possibly continuous arguments and the second part consists of a random map ϑ transforming the index i to a (random) argument of the weight function W. A finite-width network is then seen as a choice of the map ϑ and weight function W. The evolution of the weights is captured, for a choice of ϑ, by the dependence of W in time (the time evolution will be detailed in the next section). Specifically, we couple $\mathbf{W}^0$ and W^0 as follows:

1. Given a family of initialization laws I, we choose (Ω_h, P_h, W^0) to be a neuronal embedding of I and initialize the dynamical quantities $W^0(\cdot)$.
2. Given $n \in \mathbb{N}$ and the sampling rule $\bar{P}_n$, we sample $(\vartheta(1), \dots, \vartheta(n)) \sim \bar{P}_n$ and set $\mathbf{W}^0_{hh}(i,j) = W^0_{hh}(\vartheta(i), \vartheta(j))$, $\mathbf{W}^0_{xh}(i) = W^0_{xh}(\vartheta(i))$ and $\mathbf{W}^0_{hy}(j) = W^0_{hy}(\vartheta(j))$ for $j \in \{1, \dots, n\}$.

The key property of the neuronal embedding construction is the decomposition of the probability space generating an instance of the neural network into a product space over different layers. This decomposition captures the symmetry of the neural network's output under certain permutations of the indices of the neurons, thereby generalizing the representation as an empirical measure used in (Chizat & Bach, 2018a; Sirignano & Spiliopoulos, 2021; Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2018). The following example helps clarify this analogy.

Example 2.2. *In the case of the finite-width network discussed in Example 2.1, the sampling rule $\vartheta(i) = i + \omega$ with $\omega \in \Omega_h$ common to the whole layer and distributed uniformly on Ω_h satisfies the above conditions. We further notice that $\vartheta(i) = \iota(i)$ for any (random) permutation ι of $\{1, \dots, n\}$ realizes the same neural network, i.e., a neural network with the* same *weights $\mathbf{W}$, up to permutation of the indices of its neurons.*

While the example above illustrates the connection between neuronal embeddings and finite-width neural networks by using finite probability spaces, the same connection can be established more abstractly in the case of IID initializations for arbitrary and infinite-width networks by means of the Kolmogorov extension theorem.

Example 2.3. *In the case of* IID *initialization, the neuronal embedding acquires a more explicit formulation. For a given probability space $(\Lambda, \mathcal{G}, P_0)$ we define $p_{xh}(c), p_{hh}(c, c')$ and $p_{hy}(c)$ which are respectively $\mathbb{R}^d$-valued, $\mathbb{R}$-valued and $\mathbb{R}$-valued random processes indexed by $(c, c') \in [0,1] \times [0,1]$. For any n and any collection of indices $\{c^{(i)}, (c')^{(i)} \,:\, i \in \{1, \dots, n\}\}$ let S be the set of indices and R be the set of pairs of indices, we let $\{p_{xh}(c) \,:\, c \in S\}$, $\{p_{hh}(c, c') \,:\, (c, c') \in R\}$, $\{p_{hy}(c) \,:\, c \in S\}$ be independent. Then we let $\mathrm{Law}(p_{xh}(c)) = \varrho_{xh}$, $\mathrm{Law}(p_{hh}(c, c')) = \varrho_{hh}$, $\mathrm{Law}(p_{hy}(c)) = \varrho_{hy}$ for all $c \in S$, $(c, c') \in R$. This space exists by Kolmogorov extension theorem. The desired* *neuronal embedding is obtained by taking* $\Omega_h = \Lambda \times [0,1]$, *equipped with the measure* $P_h = P_0 \times \mathrm{Unif}([0,1])$ *and we define the weight functions as*

$$\begin{aligned} W_{xh}((\lambda_1, c)) &= p_{xh}(c)(\lambda_1) \\ W_{hh}((\lambda_1, c), (\lambda_2, c')) &= p_{hh}(c, c')(\lambda_1, \lambda_2) \\ W_{hy}((\lambda_2, c)) &= p_{hy}(c)(\lambda_2). \end{aligned}$$

In order to state our results we assume that the dependence of $\vartheta(i)$ and $\vartheta(j)$ for $i \neq j$ is sufficiently weak, as stated in Assumption 4 below. While this condition is trivially satisfied by IID initialization introduced above, it makes our analysis applicable to more general – not fully necessarily IID – initialization procedures.

2.5. Training dynamics

A popular algorithm to minimize the MSE (2.1) is given by *gradient descent*: starting from an initial condition $\mathbf{W}(0)$, we update the parameters $\mathbf{W}$ in the direction of steepest descent of the loss function:

$$\mathbf{W}(j+1) := \mathbf{W}(j) - \beta D_{\mathbf{W}} \mathcal{L}(\mathbf{W}) \,, \qquad (2.4)$$

where $D_{\mathbf{W}}$ represents the Fréchet derivative with respect to $\mathbf{W}$, $j \in \mathbb{N}_0$ indexes the timesteps of the algorithm and β denotes the stepsize of the discrete-time update.

In this work, we consider the regime of asymptotically small constant step-sizes, $\beta \to 0$. In this continuum limit, the stochastic component of the dynamics is averaged before the parameters of the model can change significantly. This allows us to consider the parametric update as a deterministic dynamical system emerging from the averaging of the underlying stochastic algorithm corresponding to the limit of infinite sample sizes. This is known as the ODE method (Borkar, 2009) for analyzing stochastic approximations. We focus on the analysis of this deterministic system to highlight the core dynamical properties of our training algorithm.

We denote by

$$\mathbf{W}(t) := \{\mathbf{W}_{xh}(t;\cdot), \mathbf{W}_{hh}(t;\cdot,\cdot), \mathbf{W}_{hy}(t;\cdot)\}$$

the continuous-time, averaged trajectory of the finite-width weights with initial conditions $\mathbf{W}_{xh}(0;\cdot) = \mathbf{W}^0_{xh}(\cdot)$, $\mathbf{W}_{hh}(0;\cdot,\cdot) = \mathbf{W}^0_{hh}(\cdot,\cdot)$, $\mathbf{W}_{hy}(0;\cdot) = \mathbf{W}^0_{hy}(\cdot)$. The gradient descent dynamics for these quantities can be written as the following ODEs

$$\partial_t \mathbf{W}(t) = -D_{\mathbf{W}} \mathcal{L}(\mathbf{W}(t)) \,. \qquad (2.5)$$

While the dynamics of both $\mathbf{W}_{xh}$ and $\mathbf{W}_{hy}$ will be described by the above equation, we truncate the evolution of $\mathbf{W}_{hh}$ in an interval of width $R > 0$ as follows:

$$\partial_t \mathbf{W}_{hh}(t) = -\,\chi_R(\mathbf{W}_{hh}(t)) \odot D_{\mathbf{W}_{hh}} \mathcal{L}(\mathbf{W}(t)) \qquad (2.6)$$

where $\odot$ denotes the Hadamard product and $\chi_R : \mathbb{R} \to \mathbb{R}$ is a smooth indicator function acting component-wise on its argument and such that $\chi_R(w) = w$ if $\|w\| \leq R/2$ and $\chi_R(w) \equiv 0$ if $\|w\| \geq R$. We comment on the reasons for this truncation in Remark 3.2.

Analogously, we denote

$$W(t) := \{W_{xh}(t;\cdot), W_{hh}(t;\cdot,\cdot), W_{hy}(t;\cdot)\} \,,$$

as the continuous-time trajectory of the mean-field weights with initial condition $W_{xh}(0;\cdot) = W^0_{xh}(\cdot)$, $W_{hh}(0;\cdot,\cdot) = W^0_{hh}(\cdot,\cdot)$, $W_{hy}(0;\cdot) = W^0_{hy}(\cdot)$, obeying the set of ODEs

$$\begin{aligned} \partial_t W_{hy}(t;\vartheta) &= -\,\frac{\delta}{\delta W_{hy}} \mathcal{L}(W(t)) \\ \partial_t W_{hh}(t;\vartheta,\vartheta') &= -\,\chi_R(W_{hh}(t;\vartheta,\vartheta')) \frac{\delta}{\delta W_{hh}} \mathcal{L}(W(t)) \\ \partial_t W_{xh}(t;\vartheta) &= -\,\frac{\delta}{\delta W_{xh}} \mathcal{L}(W(t)) \end{aligned} \qquad (2.7)$$

where $\frac{\delta}{\delta W}$ denotes the variational derivative (Fréchet derivative) with respect to W. While the explicit expressions for these dynamics are derived in Appendix A, we give here the update for the last layer of mean-field weights:

$$\begin{aligned} \partial_t W_{hy}(t;\vartheta) = -\int & (\hat{F}(\mathbf{x}; W(t)) - F^*(\mathbf{x})) \qquad (2.8) \\ & \sigma_h\big(H_{hh}(\vartheta; \mathbf{x}, 0) + H_{xh}(\vartheta; \mathbf{x}_0)\big)\nu(\mathrm{d}\mathbf{x}) \,. \end{aligned}$$

In the next section we will leverage the fact that this quantity must be 0 at stationarity to establish the desired optimality result.

3. Convergence and Optimality Results

To state the main results of this paper, denoting by $L^\infty_R(P_h)$ whe set of functions on Ω_h that are essentially bounded by $R > 0$, we formulate the following assumption:

Assumption 3. *Consider a neuronal embedding* (Ω_h, P_h, W) *and consider a mean-field limit associated with the neuronal ensemble* (Ω_h, P_h) *with initialization* $W(0) = W^0$. *We assume that there exists* $K > R$ *such that*

a) Regularity of σ: σ_h *is bounded, differentiable,* $\sigma_h(0) = 0$, $\sigma'_h(0) \neq 0$ *and* $D\sigma_h$ *is* K*-bounded and* K*-Lipschitz.*

b) Universal approximation*: The span of* $\{\sigma_h(W_{xh} \cdot \mathbf{x}_0) : W_{xh} \in \mathbb{R}^d\}$ *is dense in* $L^2(\nu_0)$.

c) Diversity at initialization*: The support of the weight functions* W^0_{hh}, W^0_{xh} *at initialization satisfies*

$$\begin{aligned} \mathit{supp}(W^0_{xh}(\vartheta), & W^0_{hh}(\cdot,\vartheta), W^0_{hh}(\vartheta,\cdot)) \\ & = \mathbb{R}^d \times L^\infty_R(P_h) \times L^\infty_R(P_h) \,. \end{aligned}$$

Throughout the paper we denote by $W_{hh}(\cdot,\vartheta)$ *the random (in* ϑ*) mapping* $\vartheta' \mapsto W_{hh}(\vartheta',\vartheta)$.

d) Regularity at initialization*: The weight functions $W_{hy}^0, W_{hh}^0, W_{xh}^0$ at initialization satisfy $\sup_{\vartheta,\vartheta'} |W_{hh}^0(\vartheta,\vartheta')| \leq R$ and given $E_1(m) = \mathbb{E}(|W_{xh}^0(\vartheta)^m|)^{1/m}$ and $E_2(m) = \mathbb{E}(|W_{hy}^0(\vartheta)^m|)^{1/m}$ then*

$$\sup_{m\geq 1} \frac{1}{\sqrt{m}} \left[E_1(m) \vee E_2(m)\right] < K\,.$$

Most of the assumptions made above are standard in the literature on mean-field limits of neural networks, and were first formulated in similar terms in (Chizat & Bach, 2018a) and (Nguyen & Pham, 2020). Assumption 3a) gives technical conditions on the regularity of the nonlinearities, ensuring that the training dynamics are well-behaved. The condition on the nonvanishing derivative at the preimage of 0, which without loss of generality is assumed to be at 0 itself, is required to preserve expressivity of the network while allowing for uniform in time boundedness of the hidden weights. Assumption 3b) demands sufficient expressivity of the activation function, required to approximate any function of a finite list of inputs $\{\mathbf{x}_{-L}, \ldots, \mathbf{x}_0\}$. This condition replaces the convexity assumption from (Chizat & Bach, 2018a), and is satisfied by any nonlinearity for which the universal approximation theorem holds (Cybenko, 1989; Barron, 1993), *e.g.*, $\tanh$. Assumption 3c) guarantees that the initial condition is such that the expressivity from b) can actually be exploited. This property, which as we shall show is preserved by the network throughout training, ensures that the argument of the nonlinearity at each layer is sufficiently varied, and was first introduced in (Nguyen & Pham, 2020). Combining this with Assumption 3b) ensures, by induction, that there is no information bottleneck throughout the depth of the unrolled network and that the model is highly expressive throughout training. Finally, Assumption 3d) is a technical assumption on the data and on the weights guaranteeing the well-posedness of the training dynamics. We note that our results can also be obtained relaxing the assumption on the boundedness of the W_{hh} weights at initialization, *i.e.*, allowing hidden weights to be initialized in regions there their training dynamics is trivial. This assumption is mainly made to simplify the proof and to avoid, in practice, the computation of forward/backward passes for neurons that would not be updated by the dynamics.

Remark 3.1. *We note that Assumption 3c) is significantly stronger than the analogous "sufficient support" assumption from (Chizat & Bach, 2018a). In particular, this assumption is* not *satisfied if the weights of each layer are sampled* IID *from* any *initialization law μ. As we comment in the proof of our results, relaxing this assumption to include* IID *initialization would significantly reduce the expressivity of the untrained infinite-width network with respect to predictors $\mathbf{x}_k$ at timesteps $k < 0$. More specifically, an* IID *initialization of the weights combined with the infinite width limit we are considering results in a highly degenerate hidden state of the network. Because of the intrinsic depth of RNN structures, this generates in turn a bottleneck effect preventing information from values of the predictors in the distant past to propagate through the network.*

Remark 3.2. *The truncation of the dynamics of the hidden layer weights (2.6) (2.7) was introduced in order to guarantee existence and uniqueness of the solution to both the finite-width and the mean-field equations. Indeed, in the absence of this cutoff, weight-sharing in this class of RNNs would result in a non-Lipschitz RHS for the dynamical equation (2.5), as shown explicitly in Appendix A. Given this lack of regularity, existence of the solution cannot be guaranteed by standard analytical tools. However, in practice the weights are stored using a floating-point representation which is intrinsically bounded, and we argue that in this sense the truncation of their trajectories is a relatively natural assumption.*

We now proceed to present the main results of the paper, which we divide into two parts:

3.1. Convergence

The main result in this section is the convergence of the finite-width network trajectories to the mean-field limit, analogously to Thm. 18 in (Nguyen & Pham, 2020). More specifically, for a given neuronal ensemble (Ω, P) and sample $\mathbf{W}$ from P we define the following distance or error metric $\mathcal{D}_\tau(W, \mathbf{W})$ for any $\tau > 0$ as

$$\begin{aligned}\mathcal{D}_\tau(W,\mathbf{W}) &:= \\ &\sup_{t\in(0,\tau)} \left(\frac{1}{n^2}\|W_{hh}(t;\vartheta(i),\vartheta(j)) - \mathbf{W}_{hh}(t;i,j)\|_2\right.\\ &\vee \frac{1}{n}\|W_{xh}(t;\vartheta(j)) - \mathbf{W}_{xh}(t;j)\|_2\\ &\left.\vee \frac{1}{n}\|W_{hy}(t;\vartheta(j)) - \mathbf{W}_{hy}(t;j)\|_2\right)\end{aligned}$$

where $\|\cdot\|_2$ denotes, depending on its argument, the Frobenius norm or the classical ℓ_2 norm.

Theorem 3.3. *For any $R > 0$, let Assumptions 1, 2, 3 and 4 hold. There exist constants $c, c' > 0$ such that, for any $\delta > 0$, any $L \in \mathbb{N}$ and $\tau > 0$, there exists $n^* \in \mathbb{N}$ such that for any $n > n^*$ with probability at least $1 - \delta - \bar{K} n \exp(-\bar{K} n^{c'})$ we have*

$$\mathcal{D}_\tau(W,\mathbf{W}) \leq \bar{K} n^{-c} \sqrt{\log\left(n^2/\delta + e\right)}$$

where $\bar{K}$ is a constant that depends on L and R.

The proof of the above result mimics the one in (Nguyen & Pham, 2020) and is provided in the appendix for completeness. The main argument of the proof is similar to classical propagation of chaos results (Sznitman, 1991). The first step of the argument establishes sufficient regularity of the gradient dynamics and guarantees existence and uniqueness of the solution to (2.7). Then, one bounds the difference in differential updates for the particle system and the mean-field dynamics as a function of the distance $\mathcal{D}_t(W, \mathbf{W})$. The proof is concluded by an application of Grönwall's inequality.

3.2. Optimality

The main optimality result is presented in the following theorem.

Theorem 3.4. *For any $R > 0$ let Assumption 1, 2 and 3 hold and assume that the trajectory $W(t)$ solving* (2.7) *converges to $\bar{W}$ in the following sense: for all $i \in \{1, \dots, L\}$ the following quantities vanish in the limit $t \to \infty$,*

- $\text{ess-sup}_{\vartheta \in \text{supp}(P_h)} |\partial_t W_{hy}(t; \vartheta)|$
- $\int |\bar{W}_{hy}(\vartheta) - W_{hy}(t; \vartheta)|^2 P_h(d\vartheta)$
- $\int \bar{W}_{hy}(\vartheta^{(0)})^2 \left(\prod_{j=1}^{i-1} \bar{W}_{hh}(\vartheta^{(j-1)}, \vartheta^{(j)}) \right)^2 \left(\bar{W}_{hh}(\vartheta^{(i-1)}, \vartheta^{(i)}) - W_{hh}(t; \vartheta^{(i-1)}, \vartheta^{(i)}) \right)^2 P_h^{\otimes i+1}(\mathrm{d}\boldsymbol{\vartheta})$
- $\int \bar{W}_{hy}(\vartheta^{(0)})^2 \left(\prod_{j=1}^{i-1} \bar{W}_{hh}(\vartheta^{(j-1)}, \vartheta^{(j)}) \right)^2 \left(\bar{W}_{xh}(\vartheta^{(i-1)}) - W_{xh}(t; \vartheta^{(i-1)}) \right)^2 P_h^{\otimes i}(\mathrm{d}\boldsymbol{\vartheta})$

Then $\lim_{t\to\infty} \mathcal{L}(W(t)) = 0$.

This result asserts that if the gradient descent dynamics (2.7) converges to a stationary point $\bar{W}$, that point must be a global minimizer of the population risk, *i.e.*, it must approximate the underlying function to arbitrary accuracy. We prove the above result in three steps. First, we show in Proposition D.1 that if the weights at initialization are sufficiently varied (Assumption 3c)) then the network enjoys a high level of expressivity, inherited from the properties of σ_h Assumption 3a) and b). Such expressivity in turn implies that the mean-field vector fields evaluated at a suboptimal fixed point of the dynamics (2.7) cannot vanish everywhere in neuronal embedding space. In other words, a network whose weights have sufficient support cannot correspond to a suboptimal stationary point of the gradient dynamics.

We then show in Lemma D.4 that such sufficient notion of support (Assumption 3c)) is preserved by the gradient descent dynamics (2.7) throughout training. For any finite time, this is true by topological arguments: the full support property cannot be altered by a continuous vector field such as (2.7).

Finally, we show that the gradient descent dynamics cannot converge to a spurious fixed point by combining the two partial results above. In particular, we show that by the preserved expressivity of the network throughout the dynamics proven above, the fact that the time derivative of $W_{hy}(t)$ (2.8) must vanish almost-everywhere as $t \to \infty$ implies that the difference between the approximator and the target function F^* must also vanish almost everywhere in the limit. In other words, combining the assumption on convergence of $W_{hy}(t)$ with the nondegeneracy of the W_{hy}-Jacobian of the network (following from expressivity) imples that the limiting point must be optimal.

There are multiple technical challenges that need to be addressed in this proof with respect to the proof techniques used in previous results. The most important one stems from the fact that the input structure of the (unrolled) network is different from a standard feedforward network or ResNet. The additive combination of the input with the hidden state of the previous "layer", together with weight sharing, results in possible degeneracies of the dynamics that need to be taken into account in the proof. By studying the risk minimization problem in equation (2.1) and considering exclusively the dynamics of W_{hy} (2.8), we bypass these degenercy problems by leveraging the expressivity Assumption 3c), as we now explain. As it can be observed from the explicit expressions for the ODE RHS derived in Appendix A, the time differentials of W_{hh} and W_{xh} consist of a sum of L terms. This is a direct consequence of weight sharing and does not appear in the feedforward analysis. Because of this sum, the RHS of the ODE might vanish a) if all the terms of the sum are 0 or b) if those terms do not vanish but cancel additively with each other, leading to a potentially suboptimal fixed point. The study of the locus in parameter space where b) occurs, which is necessary if one wants to characterize the set of fixed points of the ODEs, is extremely challenging. We bypass this problem by considering the differential of the W_{hy} terms where, because of the absence of the sum structure mentioned above, the nature of fixed points is of type a). This allows to establish our results without considering the potentially degenerate points of the W_{hh} and W_{xh} dynamics resulting from weight sharing.

Finally, we note that the boundedness of W_{hh} resulting from the truncation discussed in Remark 3.2, prevents us from using any of the classical expressivity results leveraging the vector space structure of the space of admissible weights. In adapting our proof to bypass the issues resulting from such boundedness, we leverage the fact that, by Assumption 3a), the image under σ_h of a function whose supremum is close to 0 is close to the identity. Combining this with the possibility of choosing arbitrarily small hidden weights results in the network being able to propagate information throughout its layers at arbitrary levels of accuracy. Finally, the unboundedness of W_{hy} allows to recover this information and therefore to realize the expressive potential of the network.

4. Numerical Experiments

In this section we numerically validate our theoretical optimality and convergence results (respectively Theorem 3.3 and 3.4) with some simulations. More specifically, the first experiment aims to exemplify our optimality result Theorem 3.4 by training a wide network model - intended as an approximation of the mean-field ODE - to show that the empirical risk always converges, given sufficient training time, to 0. The second experiment aims instead to exemplify the convergence result Theorem 3.3 by training a sequence of RNN of increasing width and showing that the evolution of a specific observable – the empirical risk – approaches a limiting curve as $n \to \infty$, reflecting the fact that the dynamics of the network converge to a limiting object (the mean-field ODEs) over finite time intervals.

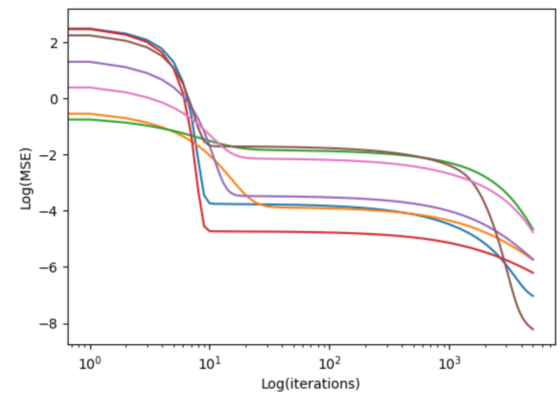
Network architecture and model The network model we consider is a wide a RNN in the mean-field regime in a teacher-student scenario. For the optimality results we set the width to be $n_s = 1000$ and for the convergence results we train the student RNN with increasing size ($n_s \in \{20, 60, 100, 140, \ldots, 300\}$).

We train a so-called *student* many-to-one RNN $\hat{F}$ with $d = 1$ and hidden layer width n_s to learn the output of a *teacher* many-to-one RNN F^*, with the same input and output size and hidden width $n_t = 15$. Both neural networks have hidden activation $\sigma_h(\cdot) = \tanh(\cdot)$ and their weights are initialized IID as follows:

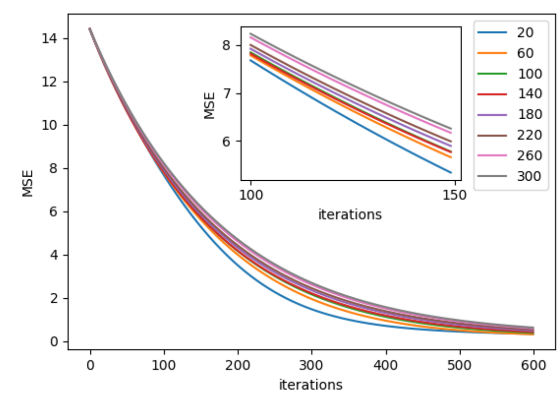
$$\text{teacher: } \begin{cases} \mathbf{W}_{xh} \sim \mathcal{N}(1,1) \\ \mathbf{W}_{hh} \sim \mathcal{N}(0,1) \\ \mathbf{W}_{hy} \sim \mathcal{N}(0,1) \end{cases} \qquad \text{student: } \begin{cases} \mathbf{W}_{xh} \sim \mathcal{N}(0,5) \\ \mathbf{W}_{hh} \sim \mathcal{N}(0,10) \\ \mathbf{W}_{hy} \sim \mathcal{N}(0,10). \end{cases}$$

Simulated data The predictors for our simulation are generated as samples of length $L = 10$ from the stationary trajectories of the shift map $T(x) = x + 1$ acting on the sphere $\mathbb{X} = S^1 = [0, 2\pi)$. To do so, we sample the initial point $\mathbf{x}^{(j)}_{-L}$ IID from the invariant measure $\nu_0(\mathrm{d}x) = \frac{1}{2\pi}\mathrm{d}x$ of T supported on $\mathbb{X}$ and generate the corresponding input sequence as $\mathbf{x}^{(j)}_{-k+1} := T(\mathbf{x}^{(j)}_{-k})$ for $k \in (1, \ldots, L)$.

Training specifications The training of the student RNN is performed using the `nn` package in pytorch (Paszke et al., 2019). We train the parameters $\hat{\mathbf{W}}$ to minimize the empirical Mean Squared Error $\tilde{\mathcal{L}}_m(\hat{\mathbf{W}}) := \frac{1}{m}\sum_{j=1}^{m}(F^*(\mathbf{x}^{(j)}) - \hat{F}(\mathbf{x}^{(j)}, \mathbf{W}))^2$ where $m = 2^{13} \approx 10^4$ denotes the size of the database. Combining this with our sampling of $\mathbf{x}^{(j)}$ results in IID samples from the invariant measure $\nu(\mathbf{x})$ of T, and therefore in the finite-sample equivalent of the population risk (2.1). The optimization is performed using stochastic gradient descent (`pytorch.optim.SGD`), which is called with a stepsize $\gamma = 3 \cdot 10^{-3}$ and batch size of m, the full database size, thereby resulting in full-fledged gradient descent. The results of the simulation are shown in Fig. 2. Code is available at (anonymous).



(a)



(b)

Figure 2: Results of numerical experiments. In Fig. 2a we plot evolution of the MSE $\tilde{\mathcal{L}}(\hat{\mathbf{W}})$ on a log-log scale as a function of training steps for predictors generated by a deterministic dynamical system. Here, different curves correspond to different (random) initializations of the student network. For each experiment, the data, the weights of the teacher and the initialization weights of the student are generated anew independently from previous experiments. In Fig. 2b, we plot the evolution of the MSE for a network of growing size (the legend indicates the size of the student network). The inset zooms on the same plots for timesteps $k \in [100, 150]$.

Initialization for the convergence results For the convergence results we need to initialize the family of student networks in a consistent way. Our procedure for enforcing consistency draws the weights without replacement from a reference student network of width 300.

5. Conclusions

This work shows that, despite the increased complexity, RNNs share common optimality properties with simpler single-layer neural networks (Chizat & Bach, 2018a). Specifically we show that, under some conditions on the expressivity of the network at initialization, the fixed points of wide Elman-type RNNs with gradient descent training dynamics in the mean-field regime are globally optimal, *i.e.*, that the neural network will perfectly learn the given function of the dynamical system's trajectories. In this sense, while extending previous results on the optimality properties of shallow and deep neural networks to novel architectures, this work contributes to the understanding of deep learning applied to dynamical systems data. The proof is carried out by unrolling the RNN structure and showing that the fixed points of the training dynamics, which preserve a certain notion of support in parameter space, can only be globally optimal.

We note that the adaptation of the neural embedding framework from (Nguyen & Pham, 2020) to the RNN setting requires a number of technical innovations, mainly resulting from the fact that weight-sharing in RNNs requires the truncation of the dynamics to prevent its solutions to blow up in finite time as noted in Remark 3.2. This has deep repercussions in our analysis. On one hand, this requires the development of brand new expressivity results that relax the unboundedness of the space of network weights that is pivotal in the classical proofs of universal approximation theorems for neural network. Furthermore, our proof relaxes the requirement from (Nguyen & Pham, 2020) in the case of MSE loss of applying a bounded nonlinearity to the output of the network, allowing to fully exploit the expressivity discussed in the previous point. This in turn is reflected in the weaker L^2 bounds (as opposed to the L^∞ bounds established in (Nguyen & Pham, 2020)) that are required in our proof.

Possible future developments include relaxing the truncation of the hidden weights' dynamics in (2.7) by proving existence and uniqueness of the full-fledged gradient descent ODEs and relaxing the assumption about the support of the weights at initialization Assumption 3c) to a condition that is simpler to realize in practice. Drawing an analogy with autoregressive processes, a promising insight towards solving the latter problem consists in injecting, at each iteration of the RNN, new directions in the function space spanned by the model by means of, *e.g.*, the network biases, so as to reduce the hidden layer's null space. Another possible avenue of future research consists of relaxing the adiabaticity assumption, *i.e.*, considering the stochastic approximation problem resulting from the finite number of samples and the finite gradient stepsize. We note that, because of this assumption, our analysis is immune to the exploding gradients problem (Pascanu et al., 2013). To prevent this problem to affect a finite timestep analysis, another important extension of the present work is to establish similar results for different RNN architectures, such as the LSTM, which given its extensive use in practice is of great interest.

From the theoretical standpoint, the most important open question concerns establishing quantitative convergence of mean-field dynamics of neural networks: even in the single-layer, supervised setting, despite recent results in specific settings (Chizat, 2022), these guarantees still elude the community's research efforts.

ACKNOWLEDGMENTS.

We thank the anonymous referee for pointing out a gap in our previous proof of Theorem B.1 and for their many insightful suggestions during the reviewing process. All authors acknowledge partial support of the TRIPODS NSF grant CCF-1934964. AA acknlwledges partial support of project PRA 2022_85, funded by the University of Pisa, as well as PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. JL acknowledges the partial support of the NSF grant DMS-2012286. SM acknowledges partial support of HFSP RGP005, NSF DMS 17-13012, NSF BCS 1552848, NSF DBI 1661386, NSF IIS 15-46331, NSF DMS 16- 13261, as well as high-performance computing partially supported by grant 2016-IDG-1013 from the North Carolina Biotechnology Center. AA and SM thank Katerina Papagiannouli and Andrea Aveni for insightful discussions and acknowledge the hospitality of the Max Planck Institute for Mathematics in the Sciences and of the ScaDS institute of the University of Leipzig and Technical University of Dresden during the final part of this project.

References

Agazzi, A. and Lu, J. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. In *International Conference on Learning Representations*, 2020.

Agazzi, A. and Lu, J. Temporal-difference learning with nonlinear function approximation: lazy training and mean field regimes. In Bruna, J., Hesthaven, J., and Zdeborova, L. (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 37–74. PMLR, 16–19 Aug 2022.

Alemohammad, S., Wang, Z., Balestriero, R., and Baraniuk, R. The recurrent neural tangent kernel. In *International Conference on Learning Representations*, 2021.

anonymous. Code for numerical simulations. url-https://github.com/anonymous229321329857123/rnn.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Chen, M., Pennington, J., and Schoenholz, S. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning*, pp. 873–882. PMLR, 2018.

Chizat, L. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1-2):487–532, 2022.

Chizat, L. and Bach, F. On the Global Convergence of Gradient Descent for Over-parameterized Models Using Optimal Transport. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 3040–3050, USA, 2018a. Curran Associates Inc.

Chizat, L. and Bach, F. On the Global Convergence of Gradient Descent for Over-parameterized Models Using Optimal Transport. In *NIPS 31*, 2018b.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, January 2014.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.

Dixon, M. F., Halperin, I., and Bilokon, P. *Machine learning in Finance*, volume 1170. Springer, 2020.

Elman, J. Finding Structure in Time. *Cognitive Science*, 14: 179–211, 1990.

Gers, F. A., Schmidhuber, J., and Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of Lazy Training of Two-layers Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 9108–9118, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021. doi: 10.1214/20-AOS1990. URL `https://doi.org/10.1214/20-AOS1990`.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pp. 8571–8580, 2018.

Jordan, M. I. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pp. 471–495. Elsevier, 1997.

Katok, A. and Hasselblatt, B. *Introduction to the Modern Theory of Dynamical Systems*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1995. doi: 10.1017/CBO9780511809187.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Lu, Y., Ma, C., Lu, Y., Lu, J., and Ying, L. A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pp. 6426–6436. PMLR, 2020.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Nguyen, P.-M. and Pham, H. T. A Rigorous Framework for the Mean Field Limit of Multilayer Neural Networks. *arXiv e-prints*, art. arXiv:2001.11443, January 2020.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318. PMLR, 2013.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pham, H. T. and Nguyen, P.-M. Global Convergence of Three-layer Neural Networks in the Mean Field Regime. In *International Conference on Learning Representations*, 2020.

Rajkomar, A., Dean, J., and Kohane, I. Machine learning in medicine. *New England Journal of Medicine*, 380(14): 1347–1358, 2019.

Rotskoff, G. and Vanden-Eijnden, E. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7146–7155. Curran Associates, Inc., 2018.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 2021.

Sirignano, J. and Spiliopoulos, K. Asymptotics of reinforcement learning with neural networks. *Stochastic Systems*, 12(1):2–29, 2022.

Sznitman, A.-S. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pp. 165–251. Springer, 1991.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wojtowytsch, S. On the Convergence of Gradient Descent Training for Two-layer ReLU-networks in the Mean Field Regime. *arXiv preprint arXiv:2005.13530*, 2020.

Yang, G. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

A. Computation of mean-field ODEs

In this section we explicitly compute the RHS of the mean-field ODEs (2.7). Recall the definitions of the population risk $\mathcal{L}(W)$ from (2.1) and of the mean-field approximator

$$\begin{aligned}
\hat{F}(\mathbf{x}; W) &= H_{hy}(\mathbf{x}) \\
H_{hy}(\mathbf{x}) &= \int W_{hy}(\vartheta)\sigma_h(H_{hh}(\vartheta; \mathbf{x}, 0) + H_{xh}(\vartheta; \mathbf{x}_0))P_h(\mathrm{d}\vartheta) \\
H_{hh}(\vartheta; \mathbf{x}, k) &= \int W_{hh}(\vartheta, \vartheta')\sigma_h(H_{hh}(\vartheta'; \mathbf{x}, k+1) + H_{xh}(\vartheta'; \mathbf{x}_{-(k+1)}))P_h(\mathrm{d}\vartheta') \\
H_{xh}(\vartheta; \mathbf{x}_{-k}) &= W_{xh}(\vartheta) \cdot \mathbf{x}_{-k}
\end{aligned} \tag{A.1}$$

Further, for notational convenience, we define throughout the argument of the nonlinearity as

$$H_\sigma(\vartheta; \mathbf{x}, k) := H_{hh}(\vartheta; \mathbf{x}, k) + H_{xh}(\vartheta; \mathbf{x}_{-k}) \tag{A.2}$$

and, when necessary, we will slightly abuse notation and explicitly write the set of weights generating the hidden state H_σ in its argument as $H_\sigma[W](\vartheta; \mathbf{x}, k)$. Furthermore, we define

$$\Delta F(W, \mathbf{x}) := \hat{F}(\mathbf{x}; W(t)) - F^*(\mathbf{x}) \tag{A.3}$$

so that we can write

$$\frac{\delta}{\delta W_{hy}}\mathcal{L}[W](\vartheta) = \int \Delta F(W, \mathbf{x})\sigma_h\big(H_\sigma[W](\vartheta; \mathbf{x}, 0)\big)\nu(\mathrm{d}\mathbf{x})$$

We proceed to compute the derivative WRT W_{xh}:

$$\begin{aligned}
\frac{\delta}{\delta W_{xh}}\mathcal{L}[W](\vartheta) &= \int \Delta F(W, \mathbf{x}) \left[\frac{\delta}{\delta W_{xh}} \int W_{hy}(\vartheta')\sigma_h\big(H_{hh}(\vartheta'; \mathbf{x}, 0) + H_{xh}(\vartheta'; \mathbf{x}_0)\big)P_h(\mathrm{d}\vartheta')\right](\vartheta)\nu(\mathrm{d}\mathbf{x}) \\
&= \int \Delta F(W, \mathbf{x}) \int W_{hy}(\vartheta')\Xi_0(\vartheta; \vartheta', \mathbf{x})P_h(\mathrm{d}\vartheta')\nu(\mathrm{d}\mathbf{x})
\end{aligned}$$

where, denoting here and throughout by $\delta(\vartheta)$ the Dirac delta distribution, we define recursively

$$\begin{aligned}
\Xi_i[W](\vartheta; \vartheta', \mathbf{x}) &:= \frac{\delta}{\delta W_{xh}}\sigma_h\big(H_{hh}(\vartheta'; \mathbf{x}, i) + H_{xh}(\vartheta'; \mathbf{x}_{-i})\big)(\vartheta) \\
&= \sigma_h'\big(H_\sigma(\vartheta'; \mathbf{x}, i)\big)\left(\left[\frac{\delta}{\delta W_{xh}}H_{hh}(\vartheta'; \mathbf{x}, i)\right](\vartheta) + \mathbf{x}_i\delta(\vartheta' - \vartheta)\right) \\
&= \sigma_h'\big(H_\sigma(\vartheta'; \mathbf{x}, i)\big)\left(\int W_{hh}(\vartheta', \vartheta'')\Xi_{i-1}(\vartheta; \vartheta'', \mathbf{x})P_h(\mathrm{d}\vartheta'') + \mathbf{x}_i\delta(\vartheta' - \vartheta)\right)
\end{aligned}$$

and throughout we slightly abuse notation by suppressing the dependency of Ξ_i on W when clear from the context. Therefore, we obtain

$$\frac{\delta}{\delta W_{xh}}\mathcal{L}[W](\vartheta) = \int \Delta F(W, \mathbf{x})\left(\sum_{i=0}^{L}\Gamma_i(W, \vartheta, \mathbf{x})\mathbf{x}_{-i}\right)\nu(\mathrm{d}\mathbf{x}) \tag{A.4}$$

where for $i \in \{0, 1, \ldots, L\}$ we define

$$\begin{aligned}
\Gamma_i(W, \vartheta, \mathbf{x}) = \int W_{hy}(\vartheta_0)\sigma_h'(H_\sigma(\vartheta_0; \mathbf{x}, 0)) \int W_{hh}(\vartheta_0, \vartheta_1)\sigma_h'(H_\sigma(\vartheta_1; \mathbf{x}, 1)) \\
\cdots \int W_{hh}(\vartheta_i, \vartheta)\sigma_h'(H_\sigma(\vartheta; \mathbf{x}, i))\, P_h^{\otimes i+1}(\vartheta_0, \ldots, \vartheta_i)
\end{aligned} \tag{A.5}$$

Analogously, we proceed to compute the derivative WRT W_{hh}:

$$\frac{\delta}{\delta W_{hh}}\mathcal{L}[W](\vartheta, \vartheta') = \int \Delta F(W, \mathbf{x})\left[\frac{\delta}{\delta W_{hh}}\int W_{hy}(\vartheta_0)\sigma_h\big(H_\sigma[W](\vartheta_0; \mathbf{x}, 0)\big)P_h(\mathrm{d}\vartheta_0)\right](\vartheta, \vartheta')\nu(\mathrm{d}\mathbf{x})$$

$$= \int \Delta F(W, \mathbf{x}) \int W_{hy}(\vartheta_0) \Xi_0'[W](\vartheta, \vartheta'; \vartheta_0, \mathbf{x}) P_h(\mathrm{d}\vartheta_0) \nu(\mathrm{d}\mathbf{x})$$

where we define recursively

$$\begin{aligned}
\Xi_i'[W](\vartheta, \vartheta'; \vartheta_i, \mathbf{x}) &:= \frac{\delta}{\delta W_{hh}} \sigma_h\big(H_{hh}[W](\vartheta_i; \mathbf{x}, i) + H_{xh}[W](\vartheta_i; \mathbf{x}_{-i})\big)(\vartheta, \vartheta') \\
&= \sigma_h'\big(H_\sigma[W](\vartheta_i; \mathbf{x}, i)\big) \left[\frac{\delta}{\delta W_{hh}} H_{hh}[W](\vartheta_i; \mathbf{x}, i)\right](\vartheta, \vartheta') \\
&= \sigma_h'\big(H_\sigma[W](\vartheta_i; \mathbf{x}, i)\big) \Bigg(\int W_{hh}(\vartheta_i, \vartheta_{i+1}) \Xi_{i+1}'[W](\vartheta, \vartheta'; \vartheta_{i+1}, \mathbf{x}) P_h(\mathrm{d}\vartheta_{i+1}) \\
&\qquad + \int \sigma_h\big(H_\sigma[W](\vartheta_{i+1}; \mathbf{x}, i+1)\big) \delta(\vartheta_i - \vartheta) \delta(\vartheta_{i+1} - \vartheta') P_h(\mathrm{d}\vartheta_{i+1}) \Bigg)
\end{aligned}$$

and throughout we slightly abuse notation by suppressing the dependency of Ξ_i' on W when clear from the context. Therefore, we obtain

$$\frac{\delta}{\delta W_{hh}} \mathcal{L}[W](\vartheta, \vartheta') = \int \Delta F(W, \mathbf{x}) \left(\sum_{i=0}^{L} \Gamma_i(W, \vartheta, \mathbf{x}) \sigma_h(H_\sigma(\vartheta'; \mathbf{x}, i+1)) \right) \nu(\mathrm{d}\mathbf{x}) \tag{A.6}$$

for Γ_i defined in (C.8).

B. Existence and uniqueness of solutions to ODEs

We now proceed to sketch the proof of existence and uniqueness of the solutions to the mean-field ODEs, stated below. To this aim, fixing throughout a value of the cutoff $R > 0$ for (2.7), we define the sub-Gaussian norm

$$\begin{aligned}
[\![W_{hh}]\!]_{\psi,t} &:= \sqrt{50} \sup_{m \geq 1} \frac{1}{\sqrt{m}} \left(\int \sup_{s<t} |W_{hh}(s, \vartheta, \vartheta')|^m P_h^{\otimes 2}(d\vartheta, d\vartheta') \right)^{1/m} \\
[\![W_{hy}]\!]_{\psi,t} &:= \sqrt{50} \sup_{m \geq 1} \frac{1}{\sqrt{m}} \left(\int \sup_{s<t} |W_{hy}(s, \vartheta)|^m P_h(d\vartheta) \right)^{1/m} \\
[\![W_{xh}]\!]_{\psi,t} &:= \sqrt{50} \sup_{m \geq 1} \frac{1}{\sqrt{m}} \left(\int \sup_{s<t} |W_{xh}(s, \vartheta)|^m P_h(d\vartheta) \right)^{1/m}
\end{aligned}$$

inducing the norm on the weights W

$$[\![W]\!]_{\psi,t} := \max\left([\![W_{hh}]\!]_{\psi,t}, [\![W_{xh}]\!]_{\psi,t}, [\![W_{hy}]\!]_{\psi,t}\right) .$$

From these definitions we have that $[\![W_{hh}]\!]_{\psi,t} \geq \|W_{hh}\|_t$, $[\![W_{hy}]\!]_{\psi,t} \geq \|W_{hy}\|_t$, $[\![W_{xh}]\!]_{\psi,t} \geq \|W_{xh}\|_t$ where

$$\begin{aligned}
\|W_{hh}\|_t &= \left(\int \sup_{s \leq t} |W_{hh}(s, \vartheta, \vartheta')|^{50} P_h^{\otimes 2}(\mathrm{d}\vartheta, \mathrm{d}\vartheta') \right)^{1/50} \\
\|W_{hy}\|_t &= \left(\int \sup_{s<t} |W_{xh}(s, \vartheta)|^{50} P_h(\mathrm{d}\vartheta) \right)^{1/50} \\
\|W_{xh}\|_t &= \left(\int \sup_{s<t} |W_{hy}(s, \vartheta)|^{50} P_h(\mathrm{d}\vartheta) \right)^{1/50}
\end{aligned} \tag{B.1}$$

so that

$$[\![W]\!]_{\psi,t} \geq \|W\|_t \tag{B.2}$$

for

$$\|W\|_t := \|W_{hh}\|_t \vee \|W_{xh}\|_t \vee \|W_{hy}\|_t .$$

Note that by Assumption 3 we have $\|W_{hh}\|_t \leq R < K$ uniformly in $t \geq 0$.

Furthermore, for a pair of mean-field weights W, W', we define the analogous norm on differences (note the different exponent):

$$\|W - W'\|_t := \|W_{hh} - W'_{hh}\|_t \vee \|W_{xh} - W'_{xh}\|_t \vee \|W_{hy} - W'_{hy}\|_t$$

for

$$\begin{aligned}
\|W_{hh} - W'_{hh}\|_t &:= \left(\int \sup_{s\le t} |W_{hh}(s,\vartheta,\vartheta') - W'_{hh}(s,\vartheta,\vartheta')|^2 P_h^{\otimes 2}(\mathrm{d}\vartheta,\mathrm{d}\vartheta')\right)^{1/2} \\
\|W_{hy} - W'_{hy}\|_t &:= \left(\int \sup_{s< t} |W_{xh}(s,\vartheta) - W'_{xh}(s,\vartheta)|^2 P_h(\mathrm{d}\vartheta)\right)^{1/2} \\
\|W_{xh} - W'_{xh}\|_t &:= \left(\int \sup_{s< t} |W_{hy}(s,\vartheta) - W'_{hy}(s,\vartheta)|^2 P_h(\mathrm{d}\vartheta)\right)^{1/2}
\end{aligned} \tag{B.3}$$

Throughout this section we fix an initialization W^0 for the mean-field weights of the network.

Theorem B.1. *Assume that the initialization of the MF ODEs satisfies $[\![W^0]\!]_{\psi,0} < K$. Then under Assumption 3 there exists a unique solution to the MF ODEs (2.7).*

Analogously to (Nguyen & Pham, 2020), the proof pivots on the use of Picard's iteration. In order to apply this strategy, we define the trajectory of the weights where the RHS of the MF ODEs is obtained by "plugging in" the evolution of the weights at the previous iteration with initial condition $W(0)$:

$$\begin{aligned}
F_{xh}[W'](t,\vartheta) &:= W_{xh}(0,\vartheta) - \int_0^t \int_X \Delta F(W'(s),\mathbf{x}) \int_{\Omega_h} W'_{hy}(\vartheta')\Xi_0[W'(s)](\vartheta;\vartheta',\mathbf{x}) P_h(\mathrm{d}\vartheta')\nu(\mathrm{d}\mathbf{x})ds \\
F_{hh}[W'](t,\vartheta,\vartheta') &:= W_{hh}(0,\vartheta,\vartheta') - \int_0^t \int_X \Delta F(W'(s),\mathbf{x}) \int_{\Omega_h} W'_{hy}(\vartheta_0)\Xi'_0[W'(s)](\vartheta,\vartheta';\vartheta_0,\mathbf{x}) P_h(\mathrm{d}\vartheta_0)\nu(\mathrm{d}\mathbf{x})ds \\
F_{hy}[W'](t,\vartheta) &:= W_{hy}(0,\vartheta') - \int_0^t \int_X \Delta F(W'(s),\mathbf{x})\sigma_h\big(H_\sigma(\vartheta;\mathbf{x},0)\big)\nu(\mathrm{d}\mathbf{x})ds
\end{aligned}$$

We now present a preparatory lemma, estimating the growth of

$$\|F[W'] - F[W'']\|_t := \|F_{xh}[W'] - F_{xh}[W'']\|_t \vee \|F_{hh}[W'] - F_{hh}[W'']\|_t \ \vee \|F_{hy}[W'] - F_{hy}[W'']\|_t$$

in terms of $\|W' - W''\|_t$ in order to prove contraction of the map F. This result holds provided that the growth of the weight trajectories W', W'' is bounded in an appropriate sense. To state these necessary growth bounds, we introduce the key functional

$$K_0(t) := K^{2L+5}(1+t^2)(1+[\![W^0]\!]_{\psi,0}) \tag{B.4}$$

that depends on a large constant $K > 0$ to be chosen later. For any $T > 0$, we also define the maximal operator

$$\max_T(W) := \sup_{s\le T} |W_{hh}(s;\vartheta,\vartheta')| \vee |W_{hy}(s;\vartheta)| \vee |W_{hx}(s;\vartheta)|$$

Lemma B.2. *Let Assumption 3 hold and $[\![W^0]\!]_{0.\psi} < \infty$. For any $T > 0$ and any $B > 0$, consider two collections of mean-field parameters $W' = \{W'(t)\}_{t\le T}, W'' = \{W''(t)\}_{t\le T}$, assume that $\|W'\|_T \vee \|W''\|_T < K_0(T)$ and*

$$\mathbb{P}(\max_T(W') > K_0(T)B) \vee \mathbb{P}(\max_T(W'') > K_0(T)B) \le 2Le^{1-K_1B^2}$$

for a choice of $K, K_1 > 0$. Then we have

$$\|F[W'] - F[W'']\|_t \le k_1(1+B)\int_0^t \|W' - W''\|_s ds + k_2 e^{-k_3B^2}$$

where $k_1 = (KK_0(T))^{3L+3}$, $k_2 = T\sqrt{L}(KK_0(T))^{3L+3}$, $k_3 = K_1/2$.

Based on the definition of $K_0(t)$ from (B.4) we define the spaces $\mathcal{W}_T, \mathcal{W}_T^0$ as the set of mean-field weight trajectories W' satisfying that there exists $K > 0$ such that, respectively,

$$\|W'\|_T \leq K_0(T)$$

and

$$W'(0) = W^0, \qquad [\![W']\!]_{T,\psi} \leq K_0(T)\,, \qquad \mathbb{P}(\max_T(W') > K_0(T)B) \leq 2Le^{1-K_1B^2} \quad \forall B > 0$$

so that $\mathcal{W}_T^0 \subseteq \mathcal{W}_T$ by (B.2).

.

Proof of Theorem B.1. Fix an arbitrary finite time $T > 0$. By the fact that F is an endomorphism in $\mathcal{W}_T^0$ (Lemma 8 in (Nguyen & Pham, 2020)), we can apply Lemma B.2 (for every B with K, K_1 fixed) and iterating the above estimate to obtain

$$\begin{aligned}
\|F^{(m)}[W'] - F^{(m)}[W'']\|_T &\leq k_1(1+B)\int_0^T \|F^{(m-1)}[W'] - F^{(m-1)}[W'']\|_{t_2} dt_2 + k_2 e^{-k_3B^2} \\
&\leq k_1^2(1+B)^2 \int_0^T \int_0^{t_2} \|F^{(m-2)}[W'] - F^{(m-2)}[W'']\|_{t_3} dt_3 dt_2 \\
&\qquad + k_2 \sum_{\ell=1}^{2} \frac{(Tk_1k_2(1+B))^{\ell-1}}{\ell!} e^{-k_3B^2} \\
&\dots \\
&\leq k_1^m(1+B)^m \int_0^T \int_0^{t_2} \cdots \int_0^{t_m} \|W' - W''\|_{t_{m+1}} dt_{m+1} \dots dt_2 \\
&\qquad + k_2 \sum_{\ell=1}^{m} \frac{(Tk_1k_2(1+B))^{\ell-1}}{\ell!} e^{-k_3B^2} \\
&\leq k_1^m(1+B)^m T^m \frac{1}{m!}\|W' - W''\|_T + k_2 e^{(Tk_1k_2(1+B)) - k_3B^2}
\end{aligned}$$

Setting $B = \sqrt{m}$ and choosing $W'' = F[W']$, from the above estimate we obtain

$$\sum_{m=1}^{\infty} \|F^{(m+1)}[W'] - F^{(m)}[W']\|_T = \sum_{m=1}^{\infty} \|F^{(m)}[W''] - F^{(m)}[W']\|_T < \infty$$

showing that $F^{(m)}[W']$ is a Cauchy sequence and hence converges (the proof of completeness of the spaces $\mathcal{W}_T, \mathcal{W}_T^0$ is similar to (Nguyen & Pham, 2020) and is omitted). The uniqueness of the limit point is obtained by contradiction: Assume that W', W'' with $\|W' - W''\|_T > 0$ are fixed points of F. Then, again choosing $B = \sqrt{m}$, for every $m > 0$ we have

$$\begin{aligned}
\|W' - W''\|_T &= \|F^{(m)}[W'] - F^{(m)}[W'']\|_T \\
&\leq k_1^m(1+\sqrt{m})^m T^m \frac{1}{m!}\|W' - W''\|_T + k_2 e^{(Tk_1k_2(1+\sqrt{m})) - k_3 m}
\end{aligned}$$

which vanishes as $m \to \infty$ contradicting the assumption. Since the above argument goes through for every $T > 0$ we have existence and uniqueness for every $T > 0$. $\square$

We now introduce some more compact notation for the time differential of the mean-field weight trajectories:

$$\Delta^{xh}(\mathbf{x}, \vartheta, W'(s)) := \Delta F(W'(s), \mathbf{x}) \int W'_{hy}(\vartheta')\Xi_0[W'(s)](\vartheta; \vartheta', \mathbf{x}) P_h(\mathrm{d}\vartheta')$$

$$\Delta^{hh}(\mathbf{x}, \vartheta, \vartheta', W'(s)) := \Delta F(W'(s), \mathbf{x}) \int W'_{hy}(\vartheta_0)\Xi'_0[W'(s)](\vartheta, \vartheta'; \vartheta_0, \mathbf{x}) P_h(\mathrm{d}\vartheta_0)$$

$$= \Delta F(W, \mathbf{x}) \left(\sum_{i=0}^{L} \Gamma_i(W, \vartheta, \mathbf{x}) \sigma_h(H_\sigma(\vartheta'; \mathbf{x}, i+1)) \right)$$
$$\Delta^{hy}(\mathbf{x}, \vartheta, W'(s)) := \Delta F(W'(s), \mathbf{x}) \sigma_h\big(H_\sigma(\vartheta; \mathbf{x}, 0)\big)$$

and defining

$$\Delta_i^H(\mathbf{x}, \vartheta, W) := \Delta F(W, \mathbf{x}) \Gamma_i(W, \vartheta, \mathbf{x}) \tag{B.5}$$

we can write

$$\Delta^{hh}(\mathbf{x}, \vartheta, \vartheta', W'(s)) = \sum_{i=0}^{L} \Delta_i^H(\mathbf{x}, \vartheta, W'(s)) \sigma_h(H_\sigma(\vartheta'; \mathbf{x}, i+1)) \tag{B.6}$$

We proceed to establish the necessary a-priori growth and Lipschitz estimates to obtain the above result, defining throughout $\mathbb{E}_X[\cdot] := \int_X \cdot \nu(d\mathbf{x})$.

Lemma B.3. *Under Assumption 3, given an initialization $W(0)$, a solution W to the MF ODEs (2.7) must satisfy that for any $t > 0$*

$$\|W\|_t \vee \max_i \left(\int_{\Omega_h} \sup_{s \le t} \mathbb{E}_X[|\Delta_i^H(\mathbf{x}, \vartheta; W(s))|]^{50} \right)^{1/50} \le K^{2L+5}(1+t^2)(1+\|W\|_0)$$

for a constant $K > 0$ large enough. Similarly for $[\![W]\!]_{\psi,t}$, there exists $K > 0$ large enough such that $[\![W]\!]_{\psi,t} < K_0(t)$ for all $t > 0$. Furthermore, for any $B > 0$

$$\mathbb{P}(\max_t(W) \ge K_0(t)B) \le 2Le^{1-K_1B^2}$$

for a universal constant K_1.

Lemma B.4. *Consider two collections of mean-field parameters $W', W'' \in \mathcal{W}_T$. Under Assumption 3 for any $t < T$ and any $1 \le k \le L$ we have*

$$\left(\int \sup_{s \le t} \mathbb{E}_X \left[|H_\sigma[W'(s)](\vartheta; \mathbf{x}, k) - H_\sigma[W''(s)](\vartheta; \mathbf{x}, k)|^2 \right] \right)^{1/2} \le K^{2L} \|W' - W''\|_t$$
$$\sup_{s \le t} \mathbb{E}_X[|\hat{F}(\mathbf{x}; W'(s)) - \hat{F}(\mathbf{x}; W''(s))|] \le K^{2L} K_0(T) \|W' - W''\|_t$$

Lemma B.5. *For a given $B > 0$ consider two collections of mean-field parameters $W', W'' \in \mathcal{W}_T$ such that*

$$\mathbb{P}(\max_T(W') > K_0(T)B) \le e^{1-K_1B^2},$$
$$\mathbb{P}(\max_T(W'') > K_0(T)B) \le e^{1-K_1B^2}$$

Then under Assumption 3, for any $t \in [0, T]$ the following holds:

$$\left(\int \sup_{s \le t} \mathbb{E}_X \left[|\Delta^{hh}(\mathbf{x}, \vartheta, \vartheta', W'(s)) - \Delta^{hh}(\mathbf{x}, \vartheta, \vartheta', W''(s))| \right]^2 P_h^{\otimes 2}(d\vartheta, d\vartheta') \right)^{1/2} \le D(t, W', W'')$$

where

$$D(t, W', W'') := (KK_0(t))^{3L+3} \left((1+B)\|W' - W''\|_t + \sqrt{L} e^{-K_1B^2/2} \right)$$

Proof of Lemma B.2. The proof of this lemma is performed as in Lemma 9 in (Nguyen & Pham, 2020) combining Lemma B.4 and Lemma B.5, corresponding respectively to Lemma 10 in (Nguyen & Pham, 2020) and Lemma 11 in (Nguyen & Pham, 2020). ☐

Proof of Lemma B.3. The proof of this lemma is analogous to the one of Lemma 6 in (Nguyen & Pham, 2020). In the following we highlight the main differences with Lemma 6 in (Nguyen & Pham, 2020), to which we refer the reader for the details of the proof. We define

$$[\![W_{hh}]\!]_{m,t} := \sqrt{\frac{50}{m}} \left(\int \sup_{s \le t} |W_{hh}(s, \vartheta, \vartheta')|^m P_h^{\otimes 2}(d\vartheta, d\vartheta') \right)^{1/m}$$

and analogously for W_{xh} and W_{hy}.

Starting at the output layer, we have by Cauchy-Schwarz inequality and the fact that the mean-field ODE dynamics decrease the population risk that

$$\sup_{s\leq t}\mathbb{E}_X[\Delta F(W(s),\mathbf{x})^2] \leq \mathbb{E}_X[\Delta F(W(0),\mathbf{x})^2] = \sqrt{\mathcal{L}[W(0)]} < K$$

for a $K>0$ large enough. Consequently, we can bound the RHS of the equation for $\partial_t W_{hy}$ using Cauchy-Schwarz and the boundedness of $\sigma_h < K$ as

$$|\partial_t W_{hy}| = |\int \Delta F(W,\mathbf{x})\sigma_h\big(H_\sigma(\vartheta;\mathbf{x},0)\big)\nu(\mathrm{d}\mathbf{x})| \leq K^2$$

so that

$$[\![W_{hy}]\!]_{m,t} \leq [\![W_{hy}]\!]_{m,0} + K^2 t$$

as desired.

The boundedness result for W_{hh} trivially holds by the truncation introduced by χ_R upon choosing $K>R$ as in Assumption 3, so that $\|W_{hh}(\cdot,\cdot)\|_\infty \leq R < K$ uniformly in t.

Finally, for W_{xh} we have again by Cauchy-Schwarz

$$\begin{aligned}
[\![W_{xh}]\!]_{m,t} &\leq [\![W_{xh}]\!]_{m,0} + \sqrt{\frac{50}{m}}\left(\int_{\Omega_h} t\sup_{s\leq t}|\int \Delta F(W,\mathbf{x})\left(\sum_{i=0}^{L}\Gamma_i(W,\vartheta,\mathbf{x})\mathbf{x}_{-i}\right)\nu(\mathrm{d}\mathbf{x})|^m P_h(\mathrm{d}\vartheta)\right)^{1/m}\\
&\leq [\![W_{xh}]\!]_{m,0} + \sqrt{\frac{50}{m}}\sqrt{\mathcal{L}[W(0)]}\sum_{i=0}^{L}\left(\int|\mathbf{x}_{-i}|^2\nu(\mathrm{d}\mathbf{x})\right)^{1/2}\left(\int_{\Omega_h} t\sup_{s\leq t}\sup_{\mathbf{x}}|\Gamma_i(W,\vartheta,\mathbf{x})|^m P_h(\mathrm{d}\vartheta)\right)^{1/m}\\
&\leq [\![W_{xh}]\!]_{m,0} + \sqrt{\mathcal{L}[W(0)]}L\|\mathbf{x}_0\|_\nu K^{2L} t[\![W_{hy}]\!]_{m,t}\\
&\leq [\![W_{xh}]\!]_{m,0} + LK^{2L+2} t[\![W_{hy}]\!]_{m,t}\\
&\leq [\![W_{xh}]\!]_{m,0} + LK^{2L+2} t([\![W_{hy}]\!]_{m,0} + K^2 t)
\end{aligned}$$

where in the second upper bound we have used that $|\Gamma_i(W,\vartheta,\mathbf{x})| \leq K^{2L}$ uniformly in $i\in\{1,\ldots,L\}$, $\mathbf{x}$ and $\vartheta\in\Theta$ by boundedness of W_{hh} and σ_h,σ_h' from Assumption 3. From this follows that

$$[\![W_{xh}]\!]_{m,t} \leq (1+[\![W]\!]_{m,0})K^{2L+5}(1+t^2)\,.$$

The probability bound follows directly from the fact that $\max_T(W)$ is $K_0(t)$ sub-Gaussian by the bounds established above. □

Proof of Lemma B.4. This proof is analogous to the one of Lemma 10 in (Nguyen & Pham, 2020). Recalling the definition of H_σ from (A.2)

$$H_\sigma(\vartheta;\mathbf{x},k) := H_{hh}(\vartheta;\mathbf{x},k) + H_{xh}(\vartheta;\mathbf{x}_{-k})$$

and slightly abusing that notation by $H_\sigma[W]$ (and similarly for H_{hh}, H_{xh}) to highlight the set of weights with respect to which the hidden state is computed, we define

$$D_k^H(t) := \left(\int_{\Omega_h}\sup_{s\leq t}\mathbb{E}_X\left[|H_\sigma[W'(s)](\vartheta;\mathbf{x},k) - H_\sigma[W''(s)](\vartheta;\mathbf{x},k)|^2\right]P_h(d\vartheta)\right)^{1/2}$$

where we recall that $\mathbb{E}_X[\cdot] = \int_X \cdot\,\nu(d\mathbf{x})$. Proceeding to bound the above for decreasing values of k we have

$$\begin{aligned}
D_L^H(t) &= \left(\int_{\Omega_h}\sup_{s\leq t}\mathbb{E}_X\left[|H_{xh}[W'](\vartheta;\mathbf{x}_{-L}) - H_{xh}[W''](\vartheta;\mathbf{x}_{-L})|^2\right]P_h(d\vartheta)\right)^{1/2}\\
&= \left(\int_{\Omega_h}\sup_{s\leq t}\mathbb{E}_X\left[|W'_{xh}(\vartheta;s)\mathbf{x}_{-L} - W''_{xh}(\vartheta;s)\mathbf{x}_{-L}|^2\right]P_h(d\vartheta)\right)^{1/2}\\
&\leq \mathbb{E}_X[|\mathbf{x}_{-L}|]\left(\int_{\Omega_h}\sup_{s\leq t}|W'_{xh}(\vartheta;s) - W''_{xh}(\vartheta;s)|^2 P_h(d\vartheta)\right)^{1/2}
\end{aligned}$$

$$\leq K d_t(W', W'')$$

where we define

$$d_t(W', W'') := \max \Bigg\{ \left(\int_{\Omega_h^2} \sup_{s \leq t} |W'_{hh}(t; \vartheta, \vartheta') - W''_{hh}(t; \vartheta, \vartheta')|^2 P_h^{\otimes 2}(d\vartheta, d\vartheta') \right)^{1/2}, \left(\int_{\Omega_h} \sup_{s \leq t} |W'_{xh}(t; \vartheta) - W''_{xh}(t; \vartheta)|^2 P_h(d\vartheta) \right)^{1/2}, \left(\int_{\Omega_h} \sup_{s \leq t} |W'_{hy}(t; \vartheta) - W''_{hy}(t; \vartheta)|^2 P_h(d\vartheta) \right)^{1/2} \Bigg\} \tag{B.7}$$

For $i < L$ we have, by triangle inequality and the Lipschitz and boundedness properties on σ_h from Assumption 3

$$\begin{aligned}
D_i^H(t) &= \Bigg(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X \big[|(H_{xh}[W'](\vartheta; \mathbf{x}_{-i}) - H_{xh}[W''](\vartheta; \mathbf{x}_{-i})) \\
&\qquad + (H_{hh}[W'](\vartheta; \mathbf{x}, i) - H_{hh}[W''](\vartheta; \mathbf{x}, i))|^2 \big] P_h(d\vartheta) \Bigg)^{1/2} \\
&\leq \left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X \left[|W'_{xh}(\vartheta; s)\mathbf{x}_{-L} - W''_{xh}(\vartheta; s)\mathbf{x}_{-L}|^2 \right] P_h(d\vartheta) \right)^{1/2} \\
&\qquad + \left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X \left[|H_{hh}[W'(s)](\vartheta; \mathbf{x}, i) - H_{hh}[W''(s)](\vartheta; \mathbf{x}, i))|^2 \right] P_h(d\vartheta) \right)^{1/2} \\
&\leq K d_t(W', W'') + K^2 D_{i+1}^H(t) + K d_t(W', W'') \\
&\leq K^2 (d_t(W', W'') + D_{i+1}^H(t))
\end{aligned}$$

This implies that $\max_{i \in \{0, \dots, L\}} D_i^H(t) \leq K^{2L} d_t(W', W'')$, proving the first claim.
The second claim follows from a similar bound:

$$\begin{aligned}
\sup_{s \leq t} \mathbb{E}_X \left[\left| \hat{F}(\mathbf{x}; W'(s)) - \hat{F}(\mathbf{x}; W''(s)) \right| \right] &\leq K \left(\int_{\Omega_h} \sup_{s \leq t} |W'_{hy}(\vartheta; s) - W''_{hy}(\vartheta; s)|^2 P_h(d\vartheta) \right)^{1/2} + K \|W'_{hy}\|_t D_0(t) \\
&\leq K d_t(W', W'') + K K_0(t) D_0(t)
\end{aligned}$$

yielding the desired estimate. □

Proof of Lemma B.5. Again by similarity with the original reference we simply sketch this proof highlighting the differences with the present framework.
We start the proof establishing the a priori bound

$$\left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X [\Delta_i^H [W'(t)](\vartheta; \mathbf{x})]^{50} P_h(\mathrm{d}\vartheta) \right)^{1/50} \leq K^{2L} K_0(t)$$

which is obtained immediately by the fact that $\left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X [\Delta_L^H [W'(t)](\vartheta; \mathbf{x})]^{50} P_h(d\vartheta) \right)^{1/50} \leq K^2 K_0(T)$ as established above and by the recursion

$$\left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X [\Delta_i^H [W'(t)](\vartheta; \mathbf{x})]^{50} P_h(\mathrm{d}\vartheta) \right)^{1/50} \leq K^2 \left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X [\Delta_{i+1}^H [W'(t)](\vartheta; \mathbf{x})]^{50} P_h(\mathrm{d}\vartheta) \right)^{1/50}.$$

We now consider

$$\tilde{D}_i^H(t) := \left(\int_{\Omega_h} \sup_{s \leq t} \mathbb{E}_X \left[|\Delta_i^H [W'(t)](\vartheta; \mathbf{x}) - \Delta_i^H [W''(t)](\vartheta; \mathbf{x})| \right]^2 P_h(\mathrm{d}\vartheta) \right)^{1/2}.$$

Starting from $i=0$ we have

$$\tilde{D}_0^H(t) \le \tilde{D}_0^{H,1}(t) + \tilde{D}_0^{H,2}(t) + \tilde{D}_0^{H,3}(t)$$

where

$$\tilde{D}_0^{H,1}(t) = K\|W''_{hy}\|_t \sup_{s\le t} \mathbb{E}_X[|\hat{F}(\mathbf{x};W'(s)) - \hat{F}(\mathbf{x};W''(s))|] \le K_0(t)^2 K^{2L+2} d_t(W',W'')$$

$$\tilde{D}_0^{H,2}(t) = K^2 \left(\int_{\Omega_h} \sup_{s\le t} |W'_{hy}(t;\vartheta) - W''_{hy}(t;\vartheta)|^2 P_h(\mathrm{d}\vartheta)\right)^{1/2} \le K^2 d_t(W',W'')$$

$$\begin{aligned}\tilde{D}_0^{H,3}(t) = \Big(\int_{\Omega_h} \sup_{s\le t} \big(&|W'_{hy}(s;\vartheta)|^2 \\ &\mathbb{E}_X\left[|\Delta F(W''(s);\mathbf{x})\left(\sigma_h(H_\sigma[W'(s)](\vartheta;\mathbf{x},0)) - \sigma_h(H_\sigma[W''(s)](\vartheta;\mathbf{x},0))\right)|\right]^2\big) P_h(\mathrm{d}\vartheta)\Big)^{\frac{1}{2}} \\ \le K^{2L+2}K_0(t)(Bd_t(W',W'') + \sqrt{\Xi(B)})&\end{aligned}$$

for any $B>0$, where $\Xi(B) = 2Le^{-K_1 B^2}$ and in the last bound we have separated the expectation in Ω_h using the indicator on the set $\max_t(W) > BK_0(t)$ and its complement. We then proceed estimating $\tilde{D}_i^H(t)$ from $\tilde{D}_{i-1}^H(t)$: using the boundedness of W_{hh}, σ'_h, σ_h and the Lipschitz continuity of σ_h we have

$$\tilde{D}_i^H(t) \le \tilde{D}_i^{H,1}(t) + \tilde{D}_i^{H,2}(t) + \tilde{D}_i^{H,3}(t)$$

where we have,

$$\tilde{D}_i^{H,1}(t) = K^2 \left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X\left[|\Delta_{i-1}^H[W'(t)](\vartheta;\mathbf{x}) - \Delta_{i-1}^H[W''(t)](\vartheta;\mathbf{x})|\right]^2 P_h(\mathrm{d}\vartheta)\right)^{1/2} \le K^2 \tilde{D}_{i-1}^H(t)$$

$$\tilde{D}_i^{H,2}(t) = K^{2L}\left(\int_{\Omega_h^2} \sup_{s\le t} |W'_{hh}(t;\vartheta,\vartheta') - W''_{hh}(t;\vartheta,\vartheta')|^2 P_h^{\otimes 2}(\mathrm{d}\vartheta,\mathrm{d}\vartheta')\right)^{1/2} \le K^{2L} d_t(W',W'')$$

$$\begin{aligned}\tilde{D}_i^{H,3}(t) &= K_0(t)K^{2L+2}\left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X\left[|H_\sigma[W'(s)](\vartheta;\mathbf{x},k) - H_\sigma[W''(s)](\vartheta;\mathbf{x},k)|\right]^2 P_h(\mathrm{d}\vartheta)\right)^{1/2} \\ &\le K_0(t)K^{4L+3}(Bd_t(W',W'') + \sqrt{\Xi(B)})\,.\end{aligned}$$

Combining the above equations results in $\max_{i\in\{0,\ldots,L\}} \tilde{D}_i^H(t) \le K_0(t)^2 K^{6L+2}((1+B)d_t(W',W'') + \sqrt{\Xi(B)})$. This yields, again analogously to (Nguyen & Pham, 2020), estimates on the quantities

$$\tilde{D}_{hh}^w(t) := \left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X[\Delta^{hh}(\mathbf{x},\vartheta,\vartheta',W'(s)) - \Delta^{hh}(\mathbf{x},\vartheta,\vartheta',W''(s))]^2 P_h^{\otimes 2}(\mathrm{d}\vartheta,\mathrm{d}\vartheta')\right)^{1/2}$$

$$\tilde{D}_{xh}^w(t) := \left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X[\Delta^{xh}(\mathbf{x},\vartheta,W'(s)) - \Delta^{xh}(\mathbf{x},\vartheta,W''(s))]^2 P_h(\mathrm{d}\vartheta)\right)^{1/2}$$

$$\tilde{D}_{hy}^w(t) := \left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X[\Delta^{hy}(\mathbf{x},\vartheta,W'(s)) - \Delta^{hy}(\mathbf{x},\vartheta,W''(s))]^2 P_h(\mathrm{d}\vartheta)\right)^{1/2}$$

We only perform these estimates explicitly on the first quantity, as the other ones are analogous. In this case we have, from (B.6) by the Lipschitz continuity of σ_h and the uniform boundedness of Δ_i^H in $L^2(\nu)$,

$$\tilde{D}_{hh}^w(t) \le LK^2(\tilde{D}_{hh}^{w,1}(t) + \tilde{D}_{hh}^{w,2}(t))$$

for

$$\tilde{D}_{hh}^{w,1}(t) = \max_{i\in\{1,\ldots,L\}} \left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X[\Delta_i^H(\mathbf{x},\vartheta,W''(s)) - \Delta_i^H(\mathbf{x},\vartheta,W''(s))]^2 P_h(\mathrm{d}\vartheta)\right)^{1/2},$$

$$\tilde{D}^{w,2}_{hh}(t) = \max_{i\in\{1,\dots,L\}} \left(\int_{\Omega_h} \sup_{s\le t} \mathbb{E}_X [H_\sigma[W'(s)](\vartheta;\mathbf{x},i) - H_\sigma[W''(s)](\vartheta;\mathbf{x},i)]^2 P_h(\mathrm{d}\vartheta) \right)^{1/2}.$$

Having bounded both terms by

$$K_0(t)^2 K^{6L+2}((1+B)d_t(W',W'') + \sqrt{\Xi(B)}) \le K^{3L+2} K_0(T)^{3L+2}((1+B)d_t(W',W'') + \sqrt{\Xi(B)})$$

in the first part of this proof and in Lemma B.4 respectively concludes the argument. □

C. Proof of convergence

To prove finite time convergence for the trajectories of the large-width neural network to the corresponding mean-field limit we bound the distance

$$\mathcal{D}_\tau(W,\mathbf{W}) = \sup_{t\in(0,\tau)} \left(\frac{1}{n^2}\|W_{hh}(t;\vartheta(i),\vartheta(j)) - \mathbf{W}_{hh}(t;i,j)\|_2 \vee \frac{1}{n}\|W_{xh}(t;\vartheta(j)) - \mathbf{W}_{xh}(t;j)\|_2 \vee \frac{1}{n}\|W_{hy}(t;\vartheta(j)) - \mathbf{W}_{hy}(t;j)\|_2 \right) \tag{C.1}$$

This proof, given for completeness, adapts the steps of Proposition 25 in (Nguyen & Pham, 2020) to the present setting, and we therefore give it as a sketch. We require the following additional assumption

Assumption 4. *Let $\eta = n^{0.501}$ and consider a family of initialization laws I. For each $n \in \mathbb{N}$ the sampling rule $\bar{P}_n$ satisfies that $(\vartheta(j))_{j=1}^n \sim \bar{P}_n$ are η-independent, i.e. for all 1-bounded $f : \Omega_h \to \mathcal{H}$ where $\mathcal{H}$ is a separable Hilbert space we have*

$$\|\mathbb{E}[f(\vartheta(j))|\{\vartheta(j') : j' < j\}] - \mathbb{E}[f(\vartheta(j))]\|_{\mathcal{H}} \le \eta \qquad \text{for all } j \in \{1,\dots,n\}$$

We recall the main theorem, stated together with the above assumption

Theorem C.1. *For any $R > 0$, let Assumptions 1, 2, 3 and 4 hold. There exist constants $c, c' > 0$ such that, under Assumption 3, for any $\delta > 0$, any $L \in \mathbb{N}$ and $\tau > 0$, there exists $n^* \in \mathbb{N}$ such that for any $n > n^*$ with probability at least $1 - \delta - \bar{K} n \exp(-\bar{K} n^{c'})$ we have*

$$\mathcal{D}_\tau(W,\mathbf{W}) \le \bar{K} n^{-c} \sqrt{\log(n^2/\delta + e)}$$

where $\bar{K}$ is a constant that depends on L and R.

We will consider the evolution of the truncated version $\underline{W}$ of the initialization W^0, which is obtained by evolving according to (2.7) the initial condition

$$\underline{W_{xh}}(0,\vartheta) := \tilde{\chi}_B(W_{xh}(0,\vartheta))$$
$$\underline{W_{hy}}(0,\vartheta) := \tilde{\chi}_B(W_{hy}(0,\vartheta))$$

and respectively for $\mathbf{W}$

$$\underline{\mathbf{W}_{xh}}(0,i) := \tilde{\chi}_B(\mathbf{W}_{xh}(0,\vartheta(i)))$$
$$\underline{\mathbf{W}_{hy}}(0,i) := \tilde{\chi}_B(\mathbf{W}_{hy}(0,\vartheta(i)))$$

where $\tilde{\chi}_B(u) = u\mathbb{1}(|u| < B) + B\mathrm{sign}(u)\mathbb{1}(|u| \ge B)$ and $\mathbb{1}$ is the indicator function. Note that the W_{hh} weights were not truncated as they are bounded by assumption. Then, analogously to Proposition 27 in (Nguyen & Pham, 2020) one can show that with probability at least $1 - KLn\exp(-Ke^{-KB^2} n^{1/52})$ we have

$$\|\mathbf{W} - \underline{\mathbf{W}}\|_T \vee \|W - \underline{W}\|_T \le K \exp\left(-KB^2 + K^{2L+5}(1+T^2)(1+B)\right) \tag{C.2}$$

We define for any $t > 0$, analogously to (B.3)

$$\|\mathbf{W} - \mathbf{W}'\|_t := \|\mathbf{W}_{hh} - \mathbf{W}'_{hh}\|_t \vee \|\mathbf{W}_{xh} - \mathbf{W}'_{xh}\|_t \vee \|\mathbf{W}_{hy} - \mathbf{W}'_{hy}\|_t$$

for

$$\|\mathbf{W}_{hh} - \mathbf{W}'_{hh}\|_t := \left(\frac{1}{n^2}\sum_{j_1,j_2=1}^{n} \sup_{s\in(0,t)} |\mathbf{W}_{hh}(s,j_1,j_2) - \mathbf{W}'_{hh}(s,j_1,j_2)|^2\right)^{1/2}$$

$$\|\mathbf{W}_{xh} - \mathbf{W}'_{xh}\|_t := \left(\frac{1}{n}\sum_{j_1=1}^{n} \sup_{s\in(0,t)} |\mathbf{W}_{xh}(s,j_1) - \mathbf{W}'_{xh}(s,j_1)|^2\right)^{1/2}$$

$$\|\mathbf{W}_{hy} - \mathbf{W}'_{hy}\|_t := \left(\frac{1}{n}\sum_{j_1=1}^{n} \sup_{s\in(0,t)} |\mathbf{W}_{hy}(s,j_1) - \mathbf{W}'_{hy}(s,j_1)|^2\right)^{1/2}$$

Defining throughout

$$K_t := K^{\kappa}(1+t^{\kappa}) \tag{C.3}$$

for a choice of K, κ that can change from line to line, we proceed to show that with probability at least $1-\delta-KLn\exp(-Kn^{1/52})$

$$\mathcal{D}_t(\underline{W}, \underline{\mathbf{W}}) \le \sqrt{\frac{1}{n}\log\left(\frac{2TLn^2}{\delta}+e\right)}\exp(K_T(1+B)) \tag{C.4}$$

for every choice of $\delta > 0, B > 0$. Combining (C.2) and (C.4) via triangle inequality we obtain that with probability $1-\delta-KLn\exp(-Ke^{-KB^2}n^{1/52})$ we have

$$\begin{aligned}\mathcal{D}_t(W, \mathbf{W}) &\le \mathcal{D}_t(\underline{W}, \underline{\mathbf{W}}) + \|\mathbf{W}-\underline{\mathbf{W}}\|_T + \|W-\underline{W}\|_T \\ &\le \left(\sqrt{\frac{1}{n}\log\left(\frac{2TLn^2}{\delta}+e\right)} + \exp(-KB^2)\right)\exp(K_T(1+B))\end{aligned}$$

and choosing $B = c_0\sqrt{\log n}$ for some suitable constant $c_0 > 0$ yields the claim of Theorem 3.3.
We prove the missing result (C.4). To do so, in the remainder of the section we slightly abuse notation and denote by $W, \mathbf{W}$ the truncated $\underline{W}, \underline{\mathbf{W}}$. Then, for the newly defined $W, \mathbf{W}$, using that

$$|W^0_{hh}(\vartheta,\vartheta')| \le K\,, \tag{C.5}$$

we define the norms

$$\begin{aligned}\|\mathbf{W}_{hh}\|_t &:= \left(\frac{1}{n^2}\sum_{j_1,j_2=1}^{n} \sup_{s\in(0,t)} |\mathbf{W}_{hh}(s,j_1,j_2)|^{50}\right)^{1/50} \\ \|\mathbf{W}_{xh}\|_t &:= \left(\frac{1}{n}\sum_{j_1=1}^{n} \sup_{s\in(0,t)} |\mathbf{W}_{xh}(s,j_1)|^{50}\right)^{1/50} \\ \|\mathbf{W}_{hy}\|_t &:= \left(\frac{1}{n}\sum_{j_1=1}^{n} \sup_{s\in(0,t)} |\mathbf{W}_{hy}(s,j_1)|^{50}\right)^{1/50}\end{aligned} \tag{C.6}$$

and for a given realization of the sampling $\bar{P}_n$,

$$\begin{aligned}\|W\|_{\text{samp},t} = &\left(\frac{1}{n}\sum_{i=1}^{n}\sup_{s<t}|W_{hy}(s,\vartheta(i))|^{50}\right)^{1/50} \vee \left(\frac{1}{n}\sum_{i=1}^{n}\int\sup_{s<t}|W_{hh}(s,\vartheta(i),\vartheta')|^{50}P_h(\mathrm{d}\vartheta')\right)^{1/50} \\ &\vee \left(\frac{1}{n}\sum_{i=1}^{n}\int\sup_{s<t}|W_{hh}(s,\vartheta',\vartheta(i))|^{50}P_h(\mathrm{d}\vartheta')\right)^{1/50}\end{aligned}$$

$$\vee \left(\frac{1}{n^2} \sum_{i,j=1}^{n} \sup_{s<t} |W_{hh}(s, \vartheta(i), \vartheta(j))|^{50} \right)^{1/50} \vee \left(\frac{1}{n} \sum_{i=1}^{n} \sup_{s<t} |W_{xh}(s, \vartheta(i))|^{50} \right)^{1/50}$$

Then, analogously to Lemma B.3 and in turn Lemma 30 in (Nguyen & Pham, 2020) one can show that for each $\|W\|_0$, $\|W\|_{\text{samp},0}$ there exists κ such that, respectively, $\|\mathbf{W}\|_t \le K_t$ and $\|W\|_{\text{samp},t} \le K_t$.
Further, we define $\mathcal{E}$ as the event

$$\mathcal{E} := \{\|\mathbf{W}\|_0 \vee \|W\|_{\text{samp},0} < K\}$$

which holds with a probability of at least $1 - KLn\exp(-Kn^{1/52})$ by Lemma 29 in (Nguyen & Pham, 2020) since by assumption $\|W\|_0 < K$. This directly implies that $\|\mathbf{W}\|_t \vee \|W\|_{\text{samp},t} < K_t$ by Lemma 30 in (Nguyen & Pham, 2020).
We start by decomposing, for any $\xi > 0$,

$$\begin{aligned} \mathcal{D}_t(W, \mathbf{W}) \le K \int_0^t (D^w_{xh}(\lfloor s/\xi \rfloor \xi) + D^w_{hh}(\lfloor s/\xi \rfloor \xi) + D^w_{hy}(\lfloor s/\xi \rfloor \xi))\mathrm{d}s \\ + Kt \sup_{s \in (0, T-\xi)} \sup_{\xi' \in (0,\xi)} \max_{V \in \{W, \mathbf{W}\}} \left(D^\xi_{xh}[V](s, \xi') \vee D^\xi_{hy}[V](s, \xi') \vee D^\xi_{hh}[V](s, \xi') \right) \end{aligned} \tag{C.7}$$

where

$$D^w_{hh}(t) := \left(\frac{1}{n^2} \sum_{j,k=1}^{n} |\partial_t \mathbf{W}_{hh}(t; j, k) - \partial_t W_{hh}(t; \vartheta(j); \vartheta(k))|^2 \right)^{1/2}$$

$$D^w_{xh}(t) := \left(\frac{1}{n} \sum_{j=1}^{n} |\partial_t \mathbf{W}_{xh}(t; j) - \partial_t W_{xh}(t; \vartheta(j))|^2 \right)^{1/2}$$

$$D^w_{hy}(t) := \left(\frac{1}{n} \sum_{j=1}^{n} |\partial_t \mathbf{W}_{hy}(t; j) - \partial_t W_{hy}(t; \vartheta(j))|^2 \right)^{1/2}$$

and

$$D^\xi_{hh}[\mathbf{W}](t, \xi') := \left(\frac{1}{n^2} \sum_{j,k=1}^{n} |\partial_t \mathbf{W}_{hh}(t; j, k) - \partial_t \mathbf{W}_{hh}(t + \xi'; j, k)|^2 \right)^{1/2}$$

$$D^\xi_{xh}[\mathbf{W}](t, \xi') := \left(\frac{1}{n} \sum_{j=1}^{n} |\partial_t \mathbf{W}_{xh}(t; j) - \partial_t \mathbf{W}_{xh}(t + \xi'; j)|^2 \right)^{1/2}$$

$$D^\xi_{hy}[\mathbf{W}](t, \xi') := \left(\frac{1}{n} \sum_{j=1}^{n} |\partial_t \mathbf{W}_{hy}(t; j) - \partial_t \mathbf{W}_{hy}(t + \xi'; j)|^2 \right)^{1/2}$$

$$D^\xi_{hh}[W](t, \xi') := \left(\frac{1}{n^2} \sum_{j,k=1}^{n} |\partial_t W_{hh}(t; \vartheta(j), \vartheta(k)) - \partial_t W_{hh}(t + \xi'; \vartheta(j), \vartheta(k))|^2 \right)^{1/2}$$

$$D^\xi_{xh}[W](t, \xi') := \left(\frac{1}{n} \sum_{j=1}^{n} |\partial_t W_{xh}(t; \vartheta(j)) - \partial_t W_{xh}(t + \xi'; \vartheta(j))|^2 \right)^{1/2}$$

$$D^\xi_{hy}[W](t, \xi') := \left(\frac{1}{n} \sum_{j=1}^{n} |\partial_t W_{hy}(t; \vartheta(j)) - \partial_t W_{hy}(t + \xi'; \vartheta(j))|^2 \right)^{1/2}$$

The following lemma, proven at the end of the section, bounds the error resulting from the time-disctretization in ξ:

Lemma C.2. *For any $\xi \in [0,T]$ we have that almost surely on the event $\mathcal{E}$*

$$\sup_{s\in(0,T-\xi)} \sup_{\xi'\in(0,\xi)} \max_{V\in\{W,\mathbf{W}\}} \left(D^{\xi}_{xh}[V](s,\xi') \vee D^{\xi}_{hy}[V](s,\xi') \vee D^{\xi}_{hh}[V](s,\xi') \right) \le K_T(1+B)\xi$$

We now proceed to bound the terms on the first line of (C.7). To do so we define for $\ell \in \{1,\dots,L\}$

$$G^{\ell}_{hh}(t) := \left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[|\Delta^{\mathbf{H}}_{hh}(\mathbf{x},j,\mathbf{W}(t),\ell) - \Delta^{H}_{hh}(\mathbf{x},j,W(t),\ell)| \right]^2 \right)^{1/2}$$

$$M^{\ell}_{hh}(t) := \left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[|\mathbf{H}_{hh}(\mathbf{x},j,\mathbf{W}(t),\ell) - H_{hh}(\mathbf{x},j,W(t),\ell)| \right]^2 \right)^{1/2}$$

$$M^{\ell}_{xh}(t) := \left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[|\mathbf{H}_{xh}(\mathbf{x},j,\mathbf{W}(t),\ell) - H_{xh}(\mathbf{x},j,W(t),\ell)| \right]^2 \right)^{1/2}$$

$$M^{\ell}_{\sigma}(t) := \left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[|\mathbf{H}_{\sigma}(\mathbf{x},j,\mathbf{W}(t),\ell) - H_{\sigma}(\mathbf{x},j,W(t),\ell)| \right]^2 \right)$$

where

$$\begin{aligned}
H_{xh}(\mathbf{x},j,W,\ell) &:= W_{xh}(\vartheta(j)) \cdot \mathbf{x}_{-\ell} \\
\mathbf{H}_{xh}(\mathbf{x},j,\mathbf{W},\ell) &:= \mathbf{W}_j \cdot \mathbf{x}_{-\ell} \\
H_{hh}(\mathbf{x},j,W,\ell) &:= H_{hh}[W](\vartheta(j),\mathbf{x},\ell) \\
\mathbf{H}_{hh}(\mathbf{x},j,\mathbf{W},\ell) &:= \mathbf{H}_{hh}[\mathbf{W}](\mathbf{x},\ell)_j \\
H_{\sigma}(\mathbf{x},j,W,\ell) &:= H_{hh}(\mathbf{x},\vartheta(j),W,\ell) + H_{xh}(\mathbf{x},\vartheta(j),W,\ell) \\
\mathbf{H}_{\sigma}(\mathbf{x},j,\mathbf{W},\ell) &:= \mathbf{H}_{hh}(\mathbf{x},j,\mathbf{W},\ell) + \mathbf{H}_{xh}(\mathbf{x},j,\mathbf{W},\ell) \\
\Delta^{H}_{hh}(\mathbf{x},j,W,\ell) &:= \Delta^{H}_{\ell}(\mathbf{x},\vartheta(j),W) \\
\Delta^{\mathbf{H}}_{hh}(\mathbf{x},j,\mathbf{W},\ell) &:= \Delta^{\mathbf{H}}_{\ell}(\mathbf{x},j,\mathbf{W})
\end{aligned}$$

for Δ^{H}_{ℓ} defined in (B.5) and

$$\Delta^{\mathbf{H}}_{\ell}(\mathbf{x},j,\mathbf{W}) := \Delta F(\mathbf{W},\mathbf{x}) \mathbf{\Gamma}_{\ell}(\mathbf{W},j,\mathbf{x})$$

for

$$\begin{aligned}
\mathbf{\Gamma}_{\ell}(\mathbf{W},j,\mathbf{x}) := \frac{1}{n}\sum_{j_0=1}^{n} \mathbf{W}_{hy}(j_0)\sigma'_h(\mathbf{H}_{\sigma}(\mathbf{x},j_0,\mathbf{W},0)) \frac{1}{n}\sum_{j_1=1}^{n} \mathbf{W}_{hh}(j_0,j_1)\sigma'_h(\mathbf{H}_{\sigma}(\mathbf{x},j_1,\mathbf{W},1)) \\
\dots \frac{1}{n}\sum_{j_{\ell-1}=1}^{n} \mathbf{W}_{hh}(j_{\ell-1},j)\sigma'_h(\mathbf{H}_{\sigma}(\mathbf{x},j,\mathbf{W},\ell))
\end{aligned} \tag{C.8}$$

Further defining

$$\begin{aligned}
D^{w,\ell}_{\sigma}(t) := &\left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[1 + |\Delta^{\mathbf{H}}_{hh}(\mathbf{x},j,\mathbf{W}(t),\ell)|^2 + |\Delta^{H}_{hh}(\mathbf{x},j,W(t),\ell)|^2 \right] \right)^{1/2} \\
&\cdot \left[\left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[|\mathbf{H}_{hh}(\mathbf{x},j,\mathbf{W}(t),\ell) - H_{hh}(\mathbf{x},j,W(t),\ell)| \right]^2 \right)^{1/2} \right. \\
&\left. + \left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \left[|\mathbf{H}_{xh}(\mathbf{x},j,\mathbf{W}(t),\ell) - H_{xh}(\mathbf{x},j,W(t),\ell)| \right]^2 \right)^{1/2} \right]
\end{aligned}$$

we have that on the event $\mathcal{E}$ by Lemma 30 in (Nguyen & Pham, 2020) $D^{w,\ell}_{\sigma}(t) \leq K_T M^{\ell}_{\sigma}(t)$, so that on the same event we have

$$D^w_{hh}(t) \leq K_T \sum_{\ell=0}^{L-1} \left(M^{\ell+1}_{\sigma}(t) + G^{\ell+1}_{hh}(t) \right)$$

and analogously

$$D^w_{hy}(t) \leq K_T \left(M^0_{\sigma}(t) + G^0_{hh}(t) \right)$$
$$D^w_{xh}(t) \leq K_T \sum_{\ell=0}^{L} G^{\ell}_{hh}(t)$$

Combining these bounds with Lemma C.2 we obtain that on the event $\mathcal{E}$

$$\mathcal{D}_t(W, \mathbf{W}) \leq K_T \left(\int_0^t \sum_{\ell=0}^{L} \left(M^{\ell}_{\sigma}(s) + 2G^{\ell}_{hh}(s) \right) + (1+B)\xi \, \mathrm{d}s \right) \tag{C.9}$$

We further proceed to bound $M^{\ell}_{\sigma}(s) + 2G^{\ell}_{hh}(s)$ in terms of $\mathcal{D}_t(W, \mathbf{W})$ with high probability as follows:

Lemma C.3. *For any sequence $\{\gamma_j\}_{j=0}^{L}$ with $\gamma_j > 0$, for all $k \in \{0, \ldots, L\}$ and $t \in (0, T)$ the event $\mathcal{E}^{\mathbf{H}}_{t,k}$ where*

$$M^{\ell}_{\sigma}(t) \leq K_T^{L-\ell+1} \left(\mathcal{D}_t(W, \mathbf{W}) + (1+B) \sum_{j=\ell}^{L-1} \gamma_j \right) \qquad \textit{holds for all } \ell \in \{k, k+1, \ldots, L\}$$

has probability

$$\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,k} | \mathcal{E}) \geq 1 - \sum_{j=k}^{L} \frac{n}{\gamma_j} \exp(-n\gamma_j^2 / K_T)\,.$$

Lemma C.4. *For any sequence $\{\beta_j\}_{j=1}^{L}$ with $\beta_j > 0$, for all $k \in \{0, \ldots, L\}$ and $t \in (0, T)$ the event $\mathcal{E}^{\Delta}_{t,k}$ where*

$$G^{\ell}_{hh}(t) \leq K_T^{L+\ell+1} \left((1+B)\mathcal{D}_t(W, \mathbf{W}) + (1+B^2) \left(\sum_{j=0}^{L-1} \gamma_j + \sum_{j=1}^{\ell} \beta_j \right) \right) \qquad \textit{holds for all } \ell \in \{0, 2, \ldots, k\}$$

satisfies

$$\mathbb{P}(\mathcal{E}^{\Delta}_{t,k} \cap \mathcal{E}^{\mathbf{H}}_{t,0} | \mathcal{E}) \geq \mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,0} | \mathcal{E}) - \sum_{j=1}^{k} \frac{n}{\beta_j} \exp(-n\beta_j^2 / K_T)\,.$$

Combining the above lemmas with (C.9) and Lemma C.2 yields that for every $B > 0$ we have

$$\mathcal{D}_t(W, \mathbf{W}) \leq K_T^{2L} \int_0^t \left((1+B)\mathcal{D}_s(W, \mathbf{W}) + (1+B^2) \left(\sum_{j=1}^{L} \gamma_{j+1} + \sum_{j=1}^{L} \beta_{j+1} \right) + (1+B)\xi \right) \mathrm{d}s \tag{C.10}$$

with probability at least

$$1 - \frac{T}{\xi} \left(\sum_{j=1}^{L-1} \frac{n}{\gamma_j} \exp(-n\gamma_j^2 / K_T) - \sum_{j=2}^{L} \frac{n}{\beta_j} \exp(-n\beta_j^2 / K_T) \right) - KLn \exp(-Kn^{1/52})$$

The proof is concluded applying Gronwall's lemma with

$$\gamma_j = \beta_j := \sqrt{\frac{1}{K_T n} \log \left(\frac{2TLn^2}{\delta} + e \right)} \qquad \text{and} \qquad \xi = \frac{1}{\sqrt{n}}$$

which gives, for all $t < T$ and for all $\delta > 0$, $B > 0$

$$\begin{aligned}
\mathcal{D}_t(W, \mathbf{W}) &\leq K_T \left((1+B^2) \left(\sum_{j=1}^{L} \gamma_{j+1} + \sum_{j=1}^{L} \beta_{j+1} \right) + (1+B)\xi \right) \exp(K_T(1+B)T) \\
&\leq K_T \left(\sum_{j=1}^{L} \gamma_{j+1} + \sum_{j=1}^{L} \beta_{j+1} + \xi \right) \exp(K_T(1+B)) \\
&\leq K_T 2L \sqrt{\frac{1}{K_T n} \log\left(\frac{2TLn^2}{\delta} + e \right)} \exp(K_T(1+B))
\end{aligned}$$

with probability

$$\begin{aligned}
\mathbb{P}(\mathcal{E} \cap \mathcal{E}^{\mathbf{H}}_{T,0} \cap \mathcal{E}^{\triangle}_{T,L}) &= \mathbb{P}(\mathcal{E}^{\mathbf{H}}_{T,0} \cap \mathcal{E}^{\triangle}_{T,L} | \mathcal{E}) \mathbb{P}(\mathcal{E}) \\
&> 1 - 2L\sqrt{n}T \left(\frac{n}{\gamma_1} \exp(-n\gamma_1^2 / K_T) \right) - KLn \exp(-Kn^{1/52}) \\
&> 1 - \delta - KLn \exp(-Kn^{1/52})
\end{aligned}$$

thereby proving (C.4), as desired. □

We now proceed with the verification of the claims that led to this conclusion. We limit ourselves to checking Lemma C.2 and Lemma C.3 as the proof of Lemma C.4 is analogous.

Proof of Lemma C.3. We show the claim by induction on the depth of the unrolled network. Starting from M_σ^L we have that with probability 1

$$\begin{aligned}
&\left(\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_X [|\mathbf{H}_\sigma(\mathbf{x}, j, \mathbf{W}(t), L) - H_\sigma(\mathbf{x}, \vartheta(j), W(t), L)|]^2 \right)^{1/2} \\
&\qquad = \left(\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_X [|\mathbf{W}_{xh}(t, j)\mathbf{x}_{-L} - W_{xh}(t, \vartheta(j))\mathbf{x}_{-L}|]^2 \right)^{1/2} \leq K \mathcal{D}_t(W, \mathbf{W}) \qquad \text{(C.11)}
\end{aligned}$$

In other words, the base case holds with probability $\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,1}) = 1$.

We now assume that the claim holds for $M_\sigma^{\ell+1}$ and prove it for M_σ^ℓ. The proof for M_{hh}^ℓ, M_{xh}^ℓ is analogous. To do so we decompose M_σ^ℓ in two parts: the first measures the distance between a randomly sampled, finite set of weights evolving according to $W(t)$ and $\mathbf{W}(t)$, while the second compares the approximation obtained by taking a finite sample from $W(t)$ and the expectation WRT P_h on $W(t)$. More specifically we decompose

$$\begin{aligned}
&|\mathbf{H}_{hh}(x, i, \mathbf{W}(t), \ell) - H_{hh}(x, \vartheta(i), W_{hh}(t), \ell)| = \\
&\qquad = \Big| \frac{1}{n} \sum_{j=1}^{n} \mathbf{W}(t, i, j) \sigma_h(\mathbf{H}_{hh}(\mathbf{x}, j, \mathbf{W}(t), \ell+1) + \mathbf{H}_{xh}(j, \mathbf{x}, \mathbf{W}(t))) \\
&\qquad\quad - \int W(t, \vartheta(i), \vartheta') \sigma_h(H_{hh}(\mathbf{x}, \vartheta', W(t), \ell+1) + H_{xh}[W(t)](\vartheta'; \mathbf{x})) P_h(d\vartheta') \Big| \\
&\qquad = Q_{1,\ell}(t; i) + Q_{2,\ell}(t; i)
\end{aligned}$$

where

$$\begin{aligned}
Q_{1,\ell}(t; i) = \frac{1}{n} \sum_{j=1}^{n} \Big| &\mathbf{W}_{hh}(t, i, j) \sigma_h(\mathbf{H}_{hh}(\mathbf{x}, j, \mathbf{W}(t), \ell+1) + \mathbf{H}_{xh}(j, \mathbf{x}, \mathbf{W}(t))) \\
&- W_{hh}(t, \vartheta(i), \vartheta(j)) \sigma_h(H_{hh}(\mathbf{x}, \vartheta(j), W(t), \ell+1) + H_{xh}[W(t)](\vartheta(j), \mathbf{x})) \Big|
\end{aligned}$$

$$Q_{2,\ell}(t;i) = \left| \frac{1}{n}\sum_{j=1}^{n} W_{hh}(t,\vartheta(i),\vartheta(j))\sigma_h(H_{hh}(\mathbf{x},\vartheta(j),W(t),\ell+1) + H_{xh}[W(t)](\vartheta(j),\mathbf{x})) \right.$$
$$\left. - \int W_{hh}(t,\vartheta(i),\vartheta')\sigma_h(H_{hh}(\mathbf{x},\vartheta',W(t),\ell+1) + H_{xh}[W(t)](\vartheta',\mathbf{x}))P_h(d\vartheta') \right|$$

and we can bound

$$M_\sigma^\ell(t) \le \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_X[|Q_{1,\ell}(t;i)| + |Q_{2,\ell}(t;i)|]^2\right)^{1/2} + \left(\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_X[|\mathbf{W}_{xh}(t,j)\mathbf{x}_{-\ell} - W_{xh}(t,\vartheta(j))\mathbf{x}_{-\ell}|]^2\right)^{1/2}$$

The first term is then bounded by

$$\mathbb{E}_X\left(|Q_{1,\ell}(t;i)|\right)^2 \le \frac{K}{n}\sum_{j=1}^{n}\left(1 + |\mathbf{W}_{hh}(t,j,i)|^2 + |W_{hh}(t,\vartheta(j),\vartheta(i))|^2\right)$$
$$\cdot \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_X\left(|\mathbf{H}_\sigma(\mathbf{x},j,\mathbf{W}(t),\ell+1) - H_\sigma(\mathbf{x},\vartheta(j),W(t),\ell+1)|\right)^2$$
$$+ \frac{K}{n}\sum_{j=1}^{n}|\mathbf{W}_{hh}(t,i,j) - W_{hh}(t,\vartheta(i),\vartheta(j))|^2$$

and therefore, under the event $\mathcal{E}_{t,\ell+1}^{\mathbf{H}}$ and $\mathcal{E}$ we have

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_X\left[|Q_{1,\ell}(t;i)|\right]^2\right)^{1/2} \le K_T M_\sigma^{\ell+1}(t) + K\mathcal{D}_t(W,\mathbf{W})$$

We proceed to bound $Q_{2,\ell}(t)$. Defining

$$Z_\ell^H(t,\vartheta,\vartheta') = W_{hh}(t,\vartheta,\vartheta')\sigma_h(H_{hh}(\mathbf{x},\vartheta',\ell+1) + W_{xh}(t,\vartheta)\mathbf{x}_{-(\ell+1)})$$

Using independence of ϑ, ϑ', we have that the conditional expectation WRT P_h is trivial

$$\mathbb{E}_{P_h}[Z_\ell^H(t,\vartheta(i),\vartheta(j))|\vartheta(i)] = \mathbb{E}_{P_h}[Z_\ell^H(t,\vartheta(j),\vartheta')]$$

and we have that, for almost every $\mathbf{x}$ almost surely by assumption

$$Z_\ell^H(t,\vartheta(i),\vartheta(j)) \le K_T(1+B)$$

Then, by Lemma 28 in (Nguyen & Pham, 2020), since $\gamma_\ell \ge 0$ we have that

$$\mathbb{P}\left(\mathbb{E}_X[Q_{2,\ell}(t)] \ge K_T(1+B)\gamma_\ell\right) \le \frac{1}{\gamma_\ell}\exp(-n\gamma_\ell^2/K_T)\,.$$

The proof is concluded by combinging the bound on H_{hh} with the one on H_{xh} to yield an analogous one on H_σ and taking an union bound over $i \in \{1,\dots,n\}$, resulting in the fact that on the events $\mathcal{E}$ and $\mathcal{E}_{t,\ell+1}^{\mathbf{H}}$

$$M_\sigma^\ell(t) \ge K_T M_\sigma^{\ell+1}(t) + 2K\mathcal{D}_t(W,\mathbf{W}) + K_T(1+B)\gamma_\ell \ge K_T^{L-\ell+1}\left(\mathcal{D}_t(W,\mathbf{W}) + (1+B)\sum_{k=\ell}^{L-1}\gamma_k\right)$$

with probability at most $(n/\gamma_\ell)\exp(-n\gamma_\ell^2/K_T)$. Therefore we get by union bound

$$\mathbb{P}((\mathcal{E}_{t,\ell}^{\mathbf{H}})^c|\mathcal{E}) \le \mathbb{P}((\mathcal{E}_{t,\ell+1}^{\mathbf{H}})^c|\mathcal{E}) + (n/\gamma_\ell)\exp(-n\gamma_\ell^2/K_T) \le \sum_{k=\ell}^{L-1}\frac{n}{\gamma_{k+1}}\exp(-n\gamma_{k+1}^2/K_T)$$

proving the desired claim. □

Proof of Lemma C.2. We again only sketch this proof for the term $D^{\xi}_{hh}[W](t,\xi')$ as the other cases follow analogously. We see that since $\|W\|_0, \|W\|_{\text{samp},t} \le K$ on the event $\mathcal{E}$, we have

$$\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\sup_{s\in(0,t)}|\partial_t W_{hh}(s,\vartheta(i),\vartheta(j))|^{50}\right)^{1/50} \le K + K\left(\frac{1}{n}\sum_{j=1}^{n}\sup_{s\in(0,t)}\mathbb{E}_x[|\Delta_i^H(\mathbf{x},\vartheta(j),W(s))|]^{50}\right)^{1/50} \le K_T$$

for any $t \le T$. Consequently we have

$$\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\sup_{s\in(0,T-\xi)}\sup_{\xi'\in(0,\xi)}\left(W_{hh}(s+\xi',\vartheta(i),\vartheta(j)) - W_{hh}(s,\vartheta(i),\vartheta(j))\right)^2\right)^{1/2} \le K_T\xi$$

The desired bound results from the application of an adapted version of Lemma B.5 to the paths $W'(t) := W_{hh}(t,\vartheta(i),\vartheta(j))$, $W''(t) := W_{hh}(t+\xi,\vartheta(i),\vartheta(j))$ replacing $e^{-K_1B^2} \to 0$ by the assumed trunctaion of W. This yields almost surely on $\mathcal{E}$

$$\sup_{t\in(0,T-\xi)}\sup_{\xi'\in(0,\xi)} D^{\xi}_{hh}(t,\xi') \le K_T(1+B)\|W'-W''\|_{T-\xi} \le K_T(1+B)\xi$$

as desired. Analogous bounds on $D^{\xi}_{hh}[\mathbf{W}], D^{\xi}_{xh}[W], D^{\xi}_{xh}[\mathbf{W}], D^{\xi}_{hy}[W], D^{\xi}_{hy}[\mathbf{W}]$ prove the lemma. □

D. Global optimality

Recall the definition of the preactivation between the first and the second layer:

$$H_{hh}(\vartheta;\mathbf{x},L) = \int W_{hh}(\vartheta,\vartheta')\sigma_h(W_{xh}(\vartheta')\cdot\mathbf{x}_{-L})P_h(\mathrm{d}\vartheta')$$

and define recursively the corresponding ℓ-preactivaton for $\ell \le L-1$

$$H_{hh}(\vartheta;\mathbf{x},\ell) := \int W_{hh}(\vartheta,\vartheta')\sigma_h(H_{hh}(\vartheta';\mathbf{x},\ell+1) + W_{xh}(\vartheta')\mathbf{x}_{-(\ell+1)})\,P_h(\mathrm{d}\vartheta')$$

For notational convenience, we define $\nu_{L,\ell} := \Pi_{\#}^{(-L,-L+\ell)}\nu$, where $\Pi^{(a,b)}$ is projection on coordinates ranging from a to b.

D.1. Expressivity at initialization

In this section we prove our main expressivity result. Defining throughout $\Theta := \mathrm{supp}(P_h)$ we we state the result as follows:

Proposition D.1. *Fix $L > 0$, for any $t > 0$ let $W = W(t)$ satisfy Assumption 3b) and c), let σ_h satisfy Assumption 3a) and let Assumptions 1 and 2 hold. Then*

$$\textit{span}\,\{\sigma_h(H_{hh}(\vartheta;\mathbf{x},0) + W_{xh}(\vartheta)\mathbf{x}_0) \;:\; \vartheta\in\Theta\} = L^2(\nu_{L,0})$$

The above result can readily be rephrased in the following, more explicit form:

Corollary D.2. *Under the conditions of Proposition D.1 above, the map*

$$\hat{F}(W;\mathbf{x}) = \int W_{hy}(\vartheta)\sigma_h(H_{hh}(\vartheta;\mathbf{x},0) + W_{xh}(\vartheta)\mathbf{x}_0)P_h(\mathrm{d}\vartheta)$$

intended as a functional of $W_{hy} \in L^2(P_h)$ is dense in the space $L^2(\nu_{L,0})$.

Proposition D.1 above proves that the network can express any function in $L^2(\nu_{L,0})$ provided that the support of the weights $W(t)$ is sufficiently varied as codified in Assumption 3b). We will show in the next subsection that this condition, if satisfied at initialization, is also satisfied at every finite time throughout the dynamics.

To prove the above result we first state the following

Lemma D.3. *Let $\mathcal{X} \subseteq \mathbb{R}^{d'}$ for $d' \in \mathbb{N}$ and let μ be a probability measure on $\mathcal{X}$. Assume that $\sigma_h : \mathbb{R}\to\mathbb{R}$ satisfies Assumption 3, that the set $\Phi := \{\phi_\vartheta \;:\; \vartheta\in\Theta\} \subseteq L^2(\mu)\cap L^\infty(\mu)$ is star-shaped at $0 \in L^2(\mu)$ and span $\{\phi_\vartheta \;:\; \vartheta\in\Theta\}$ is dense in $L^2(\mu)$, then so is span $\{\sigma_h(\phi_\vartheta) \;:\; \vartheta\in\Theta\}$.*

Proof of Lemma D.3. Assume towards a contradiction that there exists $f^* \in L^2(\mu)$ such that for any sequence $\{\phi_n\}_n$ with $\phi_n \in L^2(\mu)$ we have $\int f^* \sigma_h(\phi_n)\mu(dx) = 0$ for all $n \in \mathbb{N}$. By the spanning assumption there exists $\phi^* \in \Phi$ with $\delta^* := |\int \phi^* f^* \mu(dx)| > 0$. We now consider the sequence of functions $\phi_n = (n\|\phi^*\|_\infty)^{-1}\phi^*$. By assumption on the star-like structure of Φ, $\phi_n \in \Phi$ for all $n > 0$. The result of the lemma follows by the Taylor expansion of σ_h around the point 0:

$$\sigma_h(\phi_\vartheta(x)) = 0 + \sigma_h'(0)\phi_\vartheta(x) + R[\phi_\vartheta](x)$$

where, denoting by $\mathcal{B}_\varrho^\infty(0)$ the ball of radius ϱ in the $L^\infty(\mu)$ norm around 0, there exists a constant $C > 0$ such that the remainder term satisfies $|R[\phi](x)| < C\phi(x)^2$ uniformly in $\phi \in \mathcal{B}_\varrho^\infty(0)$ and x for ϱ small enough. Then, along the sequence $\{\phi_n\}_n$ we have

$$\left|\int f^*(x)\sigma_h(\phi_n(x))\mu(dx)\right| \geq \left|\sigma_h'(0)\int f^*(x)\phi_n(x)\mu(dx)\right| - \left|\int f^*(x)R[\phi_n](x)\mu(dx)\right| \tag{D.1}$$

We notice that for $n \in \mathbb{N}$ sufficiently large we have,

$$\left|\sigma_h'(0)\int f^*(x)\phi_n(x)\mu(dx)\right| = \frac{\sigma_h'(0)}{n\|\phi^*\|_\infty}\delta^*$$
$$\left|\int f^*(x)R[\phi_n](x)\mu(dx)\right| \leq \|f^*\|_2\|R[\phi_n](x)\|_2 \leq \frac{C\|(\phi^*)^2\|_2}{n^2\|\phi^*\|_\infty^2}\|f^*\|_2 \leq \frac{1}{2}\frac{\sigma_h'(0)}{n\|\phi^*\|_\infty}\delta^*$$

so that the first term in the expansion dominates the second. Combining this with (D.1) implies that there exists n large enough such that

$$\left|\int f^*(x)\sigma_h(\phi_n(x))\mu(dx)\right| > \frac{1}{2}\frac{\sigma_h'(0)}{n\|\phi^*\|_\infty}\delta^* > 0$$

contradicting the fact that $\int f^*\phi_n\mu(dx) = 0$ for all $n \in \mathbb{N}$. □

Proof of Proposition D.1. We want to show that, for any $\ell \in \{0, \dots, L\}$,

$$\text{span}\{\sigma_h(H_{hh}(\vartheta; \mathbf{x}, \ell) + W_{xh}\mathbf{x}_{-\ell}) \;:\; \vartheta \in \text{supp}(P_h)\} = L^2(\nu_{-L,-\ell}) \tag{D.2}$$

By the deterministic nature of the dynamical system T the measure $\nu_{-L,-\ell}$ can be written, in the sense of distributions, as

$$\begin{aligned}\nu_{-L,-\ell}(\mathrm{d}\mathbf{x}) &= \nu(\mathrm{d}\mathbf{x}_{-\ell}|\mathbf{x}_{-\ell})\dots\nu(\mathrm{d}\mathbf{x}_{-L+1}|\mathbf{x}_{-L})\nu_0(\mathrm{d}\mathbf{x}_{-L})\\ &= \nu_0(\mathrm{d}\mathbf{x}_{-L})\prod_{j=1}^{\ell}\delta(\mathbf{x}_{L-j} - T^j(\mathbf{x}_{-L}))\end{aligned}$$

so that, integrating on $\mathbf{x}_{-L}, \dots, \mathbf{x}_{-\ell}$ we write

$$H_{hh}(\vartheta; x, \ell) := H_{hh}(\vartheta; (x, T(x), \dots, T^{L-\ell+1}(x)), \ell)\,.$$

Then, condition (D.2) can be written as

$$\text{span}\{\sigma_h(H_{hh}(\vartheta; x, \ell) + W_{xh}T^{L-\ell}(x)) \;:\; \vartheta \in \text{supp}(P_h)\} = L^2(\nu_0)$$

which, since $\{0\} \times L_R^\infty(P_h) \subset \text{supp}\{W_{xh}(t;\vartheta), W_{hh}(t;\vartheta,\cdot) \;:\; \vartheta \in \Theta\}$ by Lemma D.4 follows if

$$\text{span}_\vartheta\left[\int \sigma\left(W_{hh}(\vartheta,\vartheta')\sigma\left(W_{hh}(\vartheta',\vartheta'')\sigma\left(\dots\sigma_h(W_{xh}(\vartheta^{(L)})x)\right)\right)\right)P_h^{\otimes L}(d\vartheta',\dots,d\vartheta^{(L)})\right] = L^2(\nu_0) \tag{D.3}$$

We prove (D.3) by induction on the depth of the unrolled network.

Base case $\ell = L$: In this case we simply need to show that $\text{span}\{\sigma_h(W_{xh}(\vartheta)\mathbf{x}_{-L}) \;:\; \vartheta \in \Theta\}$ is dense in $L^2(\nu_{-L}) = L^2(\nu_0)$. This, however, is immediately true by the global approximation property Assumption 3b).

Induction step $\ell \to \ell - 1$: By Lemma D.3 it is sufficient to show that

$$H_{hh}'(\vartheta; x, \ell) := \int W_{hh}(\vartheta,\vartheta')\sigma_h\left(W_{hh}(\vartheta',\vartheta'')\sigma_h\left(\dots\sigma_h(W_{xh}(\vartheta^{(L)})x)\right)\right)P_h^{\otimes(L-\ell+1)}(d\vartheta^{(\ell)},\dots,d\vartheta^{(L)})$$

spans the desired space. This claim is true if having

$$\int \bar{g}(x) H'_{hh}(\vartheta; x, \ell-1)\nu_0(\mathrm{d}x) = \int \bar{g}(x) \int W_{hh}(\vartheta, \vartheta')\sigma_h(H_{hh}(\vartheta'; x, \ell))P_h(\mathrm{d}\vartheta')\nu_0(\mathrm{d}x) = 0$$

for almost all $\vartheta \in \Omega_h$ implies that the function $\bar{g} \; : \; \mathbb{R}^d \to \mathbb{R}$ must satisfy $\bar{g}(x) \equiv 0$. Using Lemma D.4 to establish that $\{W_{hh}(t; \vartheta, \cdot)\}_\vartheta$ is dense in $L^\infty_R(P_h)$ we can rewrite the above condition as

$$\begin{aligned}\int \bar{g}(x) H'_{hh}(\vartheta; x, \ell-1)\nu_0(\mathrm{d}x) &= \int \bar{g}(x) \int f(\vartheta')\sigma_h(H'_{hh}(\vartheta'; x, \ell))P_h(\mathrm{d}\vartheta')\nu_0(\mathrm{d}x) \\ &= \int f(\vartheta') \int \bar{g}(x)\sigma_h(H'_{hh}(\vartheta'; x, \ell))\nu_0(\mathrm{d}x)\, P_h(\mathrm{d}\vartheta') = 0\end{aligned}$$

for all $f \in L^\infty_R(P_h)$, where in the last line we have applied Fubini's theorem. This is true only if

$$\int \bar{g}(x)\sigma_h(H'_{hh}(\vartheta'; x, \ell))\nu_0(\mathrm{d}x) = 0 \qquad \text{for } P_h\text{-almost all } \vartheta' \in \Omega_h \,. \tag{D.4}$$

which, by the induction assumption, is only true if $\bar{g}(x) \equiv 0$, showing (D.3) and therefore the claim.

□

D.2. Preservation of expressivity during training

Recalling the definition of $L^\infty_R(P_h) = \{f \in L^2(P_h) \; : \; \sup_\Theta |f| \leq R\}$ we have

Lemma D.4 (Bidirectional diversity, Step 1 in (Nguyen & Pham, 2020), proof of Thm. 46)**.** *Let $W_{hh}(t; \cdot, \cdot), W_{xh}(t; \cdot)$ be the mean-field parameter functions solving (2.7) with initial condition $W^0_{hh}(\cdot, \cdot), W^0_{xh}(\cdot)$. If Assumption 3 holds, then at any time $t > 0$ we have that*

$$\textit{supp}(W_{xh}(t; \vartheta), W_{hh}(t; \cdot, \vartheta), W_{hh}(t; \vartheta, \cdot) \; : \; \vartheta \in \Theta) = \mathbb{R}^d \times L^\infty_R(P_h) \times L^\infty_R(P_h)$$

To prove the bidirectional diversity result we will consider the flow induced by (2.7) on any value of the (parametric) initial condition. From now on we denote by $\langle f, g\rangle$ the inner product in $L^2(P_h)$.

Proof of Lemma D.4. Consider a MF trajectory $W(t)$ and a triple $u = (u_1, u_2, u_3) \in \mathbb{R}^d \times L^\infty_R(P_h) \times L^\infty_R(P_h)$, representing respectively values of $(W_{xh}(\vartheta), W_{hh}(\cdot, \vartheta), W_{hh}(\vartheta, \cdot))$. To characterize the evolution of a triple u we consider the flow

$$\begin{aligned}\frac{\partial}{\partial t} a^+_{xh}(t; u) &= -\beta(t) \int \Delta F(W(t), \mathbf{x}) \\ &\qquad \sum_{i=1}^{L+1} \langle \Gamma_{i-1}(W(t), \cdot, \mathbf{x}), a^+_{h1}(t, \cdot; u)\rangle \sigma'_h(H_\sigma[W](\mathbf{x}, a^+_{h2}(t, \cdot; u), a^+_{xh}(t; u), i))\mathbf{x}_i \,\nu(\mathrm{d}\mathbf{x}) \\ \frac{\partial}{\partial t} a^+_{h1}(t, \vartheta; u) &= -\beta(t)\chi_R(a^+_{h1}(t, \vartheta; u)) \int \Delta F(W(t), \mathbf{x}) \\ &\qquad \sum_{i=1}^{L+1} \langle \Gamma_{i-1}(W(t), \cdot, \mathbf{x}), a^+_{h1}(t, \cdot; u)\rangle \sigma_h(H_\sigma[W](\mathbf{x}, a^+_{h2}(t, \cdot; u), a^+_{xh}(t; u), i))\nu(\mathrm{d}\mathbf{x}) \\ \frac{\partial}{\partial t} a^+_{h2}(t, \vartheta'; u) &= -\beta(t)\chi_R(a^+_{h2}(t, \vartheta'; u)) \int \Delta F(W(t), \mathbf{x}) \\ &\qquad \sum_{i=1}^{L+1} \langle \Gamma_{i-1}(W(t), \cdot, \mathbf{x}), a^+_{h1}(t, \cdot; u)\rangle \sigma_h(H_\sigma[W](\vartheta', \mathbf{x}, W, i+1))\nu(\mathrm{d}\mathbf{x})\end{aligned}$$

with initial conditions $a_{xh}(0; u)^+ = u_1$, $a_{h2}(0, \cdot; u)^+ = u_2$, $a_{h1}(0, \cdot; u)^+ = u_3$, where $\Gamma_i(W, \vartheta, \mathbf{x}), \Delta F(W, \mathbf{x})$ were defined in Appendix A and

$$H_\sigma[W](\mathbf{x}, a^+_{h2}(t, \cdot; u), a^+_{xh}(t; u), i) := \langle a^+_{h2}(t, \cdot; u)\, ,\, \sigma_h(H_\sigma[W](\cdot, \mathbf{x}, i+1))\rangle + a^+_{xh}(t; u)\mathbf{x}_{-i}$$

These flows track the evolution of mean-field parameters in the space where their evolution is naturally embedded: we see that the MF trajectory solving (2.7) satisfies

$$\begin{aligned} W_{xh}(t,\vartheta) &= a^+_{xh}(t; W_{xh}(0;\vartheta), W_{hh}(0;\vartheta,\cdot), W_{hh}(0;\cdot,\vartheta)) \\ W_{hh}(t,\vartheta,\cdot) &= a^+_{h1}(t; W_{xh}(0;\vartheta), W_{hh}(0;\vartheta,\cdot), W_{hh}(0;\cdot,\vartheta)) \\ W_{hh}(t,\cdot,\vartheta) &= a^+_{h2}(t; W_{xh}(0;\vartheta), W_{hh}(0;\vartheta,\cdot), W_{hh}(0;\cdot,\vartheta)) \end{aligned} \tag{D.5}$$

We proceed construct, for all finite $T > 0$ and every $u^+ = (u^+_1, u^+_2, u^+_3) \in \mathbb{R}^d \times L^\infty_R(P_h) \times L^\infty_R(P_h)$ an initial condition $u^- = (u^-_1, u^-_2, u^-_3) \in \mathbb{R}^d \times L^\infty_R(P_h) \times L^\infty_R(P_h)$ that reaches u^+ after time T, *i.e.*, such that

$$a^+_{xh}(T; u^-) = u^+_1 \qquad a^+_{h1}(T,\cdot; u^-) = u^+_2 \qquad a^+_{h2}(T,\cdot; u^-) = u^+_3 \tag{D.6}$$

To do so we consider the reverse-time dynamics on the interval $(0, T)$, described by the flow

$$\begin{aligned} \frac{\partial}{\partial t} a^-_{xh}(t;u) &= -\beta(T-t)\int \Delta F(W(T-t),\mathbf{x}) \\ &\qquad \sum_{i=1}^{L+1} \langle \Gamma_{i-1}(W(T-t),\cdot,\mathbf{x}), a^-_{h1}(t,\cdot;u)\rangle \sigma'_h(H_\sigma(\mathbf{x}, a^-_{h2}(t,\cdot;u), a^-_{xh}(t;u), i))\mathbf{x}_i \,\nu(\mathrm{d}\mathbf{x}) \\ \frac{\partial}{\partial t} a^-_{h1}(t,\vartheta;u) &= -\beta(T-t)\chi_R(a^-_{h1}(t,\vartheta;u))\int \Delta F(W(T-t),\mathbf{x}) \\ &\qquad \sum_{i=1}^{L+1} \langle \Gamma_{i-1}(W(T-t),\cdot,\mathbf{x}), a^-_{h1}(t,\cdot;u)\rangle \sigma_h(H_\sigma(\mathbf{x}, a^-_{h2}(t,\cdot;u), a^-_{xh}(t;u), i))\nu(\mathrm{d}\mathbf{x}) \\ \frac{\partial}{\partial t} a^-_{h2}(t,\vartheta';u) &= -\beta(T-t)\chi_R(a^-_{h2}(t,\vartheta';u))\int \Delta F(W(T-t),\mathbf{x}) \\ &\qquad \sum_{i=1}^{L+1} \langle \Gamma_{i-1}(W(T-t),\cdot,\mathbf{x}), a^-_{h1}(t,\cdot;u)\rangle \sigma_h(H_\sigma(\vartheta', \mathbf{x}, W, i))\nu(\mathrm{d}\mathbf{x}) \end{aligned}$$

initialized at $a^-_{xh}(0;u) = u_1$, $a^-_{h1}(0,\cdot;u) = u_2$ and $a^-_{h2}(0,\cdot;u) = u_3$. Note that, by construction, $\tilde{a_{h1}}^-(t) = a^-_{h1}(T - t,\vartheta;u)$, $\tilde{a_{h2}}^-(t) = a^-_{h2}(T-t,\vartheta;u)$ and $\tilde{a_{xh}}^-(t) = a^-_{xh}(t;u)$ solve the same equation as $a^+_{xh}(t;u)$, $a^+_{h1}(t,\cdot;u)$, $a^+_{h1}(t,\cdot;u)$ with initial condition $\tilde{a_{h1}}^-(0,\cdot) = a^-_{h1}(T,\cdot;u^+)$, $\tilde{a_{h2}}^-(0,\cdot) = a^-_{h2}(T,\cdot;u^+)$, $\tilde{a_{xh}}^-(0) = a^-_{xh}(T;u^+)$. By existence and uniqueness of the solution of this system of ODEs both forward and backward in time proven in Section B, we must have that, setting $u^- = (u^-_1, u^-_2, u^-_3) := (a^-_{xh}(T,\cdot;u^+), a^-_{h1}(T,\cdot;u^+), a^-_{h2}(T,\cdot;u^+))$ as the initial condition of (D.5), the endpoint of the trajectory of satisfies (D.6) as desired. Finally, we show that the point u^- is in $\mathbb{R} \times L^\infty_R(P_h) \times L^\infty_R(P_h)$. This follows immediately upon showing that the set $\mathbb{R} \times L^\infty_R(P_h) \times L^\infty_R(P_h)$ is invariant with respect to the flow maps $(a^+_{xh}, a^+_{h1}, a^+_{h2})$, $(a^-_{xh}, a^-_{h1}, a^-_{h2})$ induced by the ODEs. The forward invariance of $\mathbb{R}$ for W_{xh} under both forward and backward flow maps follows from the Lipschitz bounds on the RHS of the corresponding ODEs, established in Section B. It remains to prove forward invariance of $L^\infty_R(P_h)$, which we now do by contradiction. Assuming that $L^\infty_R(P_h)$ is not invariant with respect to $(a^+_{xh}, a^+_{h1}, a^+_{h2})$, $(a^-_{xh}, a^-_{h1}, a^-_{h2})$, then by the continuity of the flow maps, there must exist $\vartheta, \vartheta' \in \operatorname{supp} P_h$ with $|W_{hh}(t;\vartheta,\vartheta')| = K$ such that $\partial_t |W_{hh}(t;\vartheta,\vartheta')| > 0$, which is impossible given that $\partial_t |W_{hh}(t;\vartheta,\vartheta')| = 0$, since $\chi_R(W_{hh}(\vartheta,\vartheta')) = 0$, for all such ϑ, ϑ'.

By continuity of the solution map $u \mapsto (a^+_{xh}(T;u), a^+_{h1}(T,\cdot;u), a^+_{h2}(T,\cdot;u))$, for any $\varepsilon > 0$ there exists a neighborhood U of $u^- \in \mathbb{R} \times L^\infty_R(P_h) \times L^\infty_R(P_h)$ such that

$$\|(a^+_{xh}(T;u), a^+_{h1}(T,\cdot;u), a^+_{h2}(T,\cdot;u)) - u^+\| < \varepsilon$$

for all $u \in U$. This finally implies, by Assumption 3c), that $(W_{xh}(T;\vartheta), W_{hh}(T;\vartheta,\cdot), W_{hh}(T;\cdot,\vartheta))$ has full support in $\mathbb{R}^d \times L^\infty_R(P_h) \times L^\infty_R(P_h)$, which in turn proves the claim.

□

D.3. Proof of Theorem 3.4

The proof of Theorem 3.4 is carried out by adapting the argument from Theorem 50 in (Nguyen & Pham, 2020), to the present setting. We recall that, writing $\Delta F[W](\mathbf{x}) := \hat{F}(\mathbf{x}; W(t)) - F^*(\mathbf{x})$ and using the definition (A.2) we have

$$\partial_t W_{hy}(t;\vartheta) = -\int \Delta F[W](\mathbf{x})\sigma_h\big(H_\sigma(\vartheta;\mathbf{x},0)\big)\nu(\mathrm{d}\mathbf{x})$$

so that, by the convergence assumption, we have that for every $\varepsilon > 0$ there exists a $T > 0$ such that for almost every $\vartheta \in \mathrm{supp}(P_h)$

$$|\int \Delta F[W](\mathbf{x})\sigma_h\big(H_\sigma(\vartheta;\mathbf{x},0)\big)\nu(\mathrm{d}\mathbf{x})| \le \varepsilon\,.$$

We proceed to prove that $\Delta F[W]$ converges in $L^2(\nu)$ to $\Delta F[\bar W]$ as $t \to \infty$. To do so we define

$$\delta_i(t,\mathbf{x},\vartheta) = \big|\sigma_h(H_\sigma[\bar W](\vartheta;\mathbf{x},i)) - \sigma_h(H_\sigma[W(t)](\vartheta;\mathbf{x},i))\big|$$

for which by boundedness and Lipschitz continuity of σ_h we have

$$\begin{aligned}
\delta_L(t,\mathbf{x},\vartheta) &\le K|\bar W_{xh}(\vartheta)\mathbf{x}_{-L} - W_{xh}(t;\vartheta)\mathbf{x}_{-L}|\\
\delta_i(t,\mathbf{x},\vartheta) &\le K\bigg(|\bar W_{xh}(\vartheta)\mathbf{x}_{-L} - W_{xh}(t;\vartheta)\mathbf{x}_{-L}| + K\int |\bar W_{hh}(\vartheta,\vartheta') - W_{hh}(t;\vartheta,\vartheta')|P_h(\mathrm{d}\vartheta')\\
&\qquad\qquad + \int |\bar W_{hh}(\vartheta,\vartheta')\delta_{i+1}(t,\mathbf{x},\vartheta')|P_h(\mathrm{d}\vartheta')\bigg)
\end{aligned}$$

Therefore, denoting by $\mathrm{d}\boldsymbol{\vartheta}$ the differential $\mathrm{d}\vartheta^{(0)},\dots,\mathrm{d}\vartheta^{(L)}$ we have that

$$\begin{aligned}
&\int |\Delta F[\bar W](\mathbf{x}) - \Delta F[W(t)](\mathbf{x})|^2\nu(\mathrm{d}\mathbf{x})\\
&\quad = \int |\hat F(\bar W;\mathbf{x}) - \hat F(W(t);\mathbf{x})|^2\nu(\mathrm{d}\mathbf{x})\\
&\quad \le \int\bigg(K\int|\bar W_{hy}(\vartheta) - W_{hy}(t;\vartheta)|P_h(\mathrm{d}\vartheta) + \int \bar W_{hy}(\vartheta)\delta_0(t,\mathbf{x},\vartheta)P_h(\mathrm{d}\vartheta)\bigg)^2\nu(\mathrm{d}\mathbf{x})\\
&\quad \le K^{2L}\sum_{i=0}^{L}\int |\bar W_{hy}(\vartheta^{(0)})|^2\left(\prod_{j=1}^{i-1}|\bar W_{hh}(\vartheta^{(j-1)},\vartheta^{(j)})|\right)^2\Big|\bar W_{hh}(\vartheta^{(i-1)},\vartheta^{(i)}) - W_{hh}(t;\vartheta^{(i-1)},\vartheta^{(i)})\Big|^2 P_h^{\otimes L+1}(\mathrm{d}\boldsymbol{\vartheta})\\
&\qquad + K^{2L}\sum_{i=0}^{L}\int |\bar W_{hy}(\vartheta^{(0)})|^2\left(\prod_{j=1}^{i-1}|\bar W_{hh}(\vartheta^{(j-1)},\vartheta^{(j)})|\right)^2\Big|\bar W_{xh}(\vartheta^{(i-1)}) - W_{xh}(t;\vartheta^{(i-1)})\Big|^2 P_h^{\otimes L+1}(d\boldsymbol{\vartheta})\mathbb{E}_X[\|\mathbf{x}\|^2]\\
&\qquad + K^2\int|\bar W_{hy}(\vartheta) - W_{hy}(t;\vartheta)|^2P_h(d\vartheta)
\end{aligned}\tag{D.7}$$

and by Assumption 3 we have that the above goes to 0 as $t \to \infty$.

Having proven the convergence of $\Delta F[W(t)]$ to $\Delta F[\bar W]$ we proceed to prove the claim of the theorem. By boundedness of σ_h we have that for every $\vartheta \in \mathrm{supp}(P_h)$

$$\begin{aligned}
&|\int \Delta F[\bar W]\sigma_h\big(H_\sigma(\vartheta;\mathbf{x},0)\big)\nu(\mathrm{d}\mathbf{x})|\\
&\qquad\qquad \le K|\int(\Delta F[\bar W](\mathbf{x}) - \Delta F[W](\mathbf{x}))\nu(\mathrm{d}\mathbf{x})| + |\int \Delta F[W](\mathbf{x})\sigma_h\big(H_\sigma(\vartheta;\mathbf{x},0)\big)\nu(\mathrm{d}\mathbf{x})|\\
&\qquad\qquad \le K\varepsilon
\end{aligned}$$

By continuity of σ_h we have that for every $\varepsilon > 0$

$$|\int \Delta F[\bar W](\mathbf{x})f(\mathbf{x})\nu(\mathrm{d}\mathbf{x})| \le K\varepsilon$$

uniformly over $f(\mathbf{x}) \in S$ where $S = \{\sigma_h(H_\sigma(\vartheta;\mathbf{x},0)) \;:\; \vartheta \in P_h\}$, implying that $|\int \Delta F[\bar W](\mathbf{x})f(\mathbf{x})\nu(\mathrm{d}\mathbf{x})| = 0$ for all $f \in S$. Since from Proposition D.1 we have that $\mathrm{span}(\sigma_h(H_\sigma(\vartheta;\mathbf{x},0))) = L^2(\nu)$, the above result immediately yields that for ν-almost every $\mathbf{x}$, $\Delta F[\bar W](\mathbf{x}) = 0$, so that $\mathcal{L}(\bar W) = 0$.

Finally, we prove the desired result by connecting $\mathcal{L}(\bar W)$ and $\mathcal{L}(W(t))$:

$$|\mathcal{L}(\bar W) - \mathcal{L}(W(t))| = |\int \Delta F[\bar W](\mathbf{x})^2 - \Delta F[W(t)](\mathbf{x})^2\nu(\mathrm{d}\mathbf{x})| \le 2K\|\hat F(\bar W;\cdot) - \hat F(W(t);\cdot)\|_\nu\,,$$

which by (D.7) goes to 0 with $t \to \infty$. Combining the above we have

$$\lim_{t\to\infty} \mathcal{L}(W(t)) \leq \mathcal{L}(\bar{W}) + \lim_{t\to\infty} |\mathcal{L}(\bar{W}) - \mathcal{L}(W(t))| = 0$$

which proves the claim. $\square$