Math. Stat. Learn. 6 (2023), 201–357
DOI 10.4171/MSL/42

A rigorous framework for the mean field limit of multilayer neural networks

Phan-Minh Nguyen and Huy Tuan Pham

Abstract. We develop a mathematically rigorous framework for multilayer neural networks in the mean field regime. As the network's widths increase, the network's learning trajectory is shown to be well captured by a meaningful and dynamically nonlinear limit (the *mean field* limit), which is characterized by a system of ODEs. Our framework applies to a broad range of network architectures, learning dynamics and network initializations. Central to the framework is the new idea of a *neuronal embedding*, which comprises of a non-evolving probability space that allows to embed neural networks of arbitrary widths. Using our framework, we prove several properties of large-width multilayer neural networks. Firstly we show that independent and identically distributed initializations cause strong degeneracy effects on the network's learning trajectory when the network's depth is at least four. Secondly we obtain several global convergence guarantees for feedforward multilayer networks under a number of different setups. These include two-layer and three-layer networks with independent and identically distributed initializations, and multilayer networks of arbitrary depths with a special type of correlated initializations that is motivated by the new concept of *bidirectional diversity*. Unlike previous works that rely on convexity, our results admit non-convex losses and hinge on a certain universal approximation property, which is a distinctive feature of infinite-width neural networks and it is shown to hold throughout the training process. Aside from being the first known results for global convergence of multilayer networks in the mean field regime, they demonstrate flexibility of our framework and incorporate several new ideas and insights that depart from the conventional convex optimization wisdom.

Contents

2020 Mathematics Subject Classification. Primary 62G08; Secondary 82C22,90C26,68T07.
Keywords. Neural network, deep learning, mean field limit, optimization, global convergence.

1. Introduction

A major outstanding theoretical challenge in deep learning is the understanding of the learning dynamics of multilayer neural networks. A precise characterization of the learning trajectory is typically hard, primarily due to the highly nonlinear and complex structure of deep learning architectures, which departs from convex optimization even when the loss function is convex. Recent progresses tackle this challenge with one simplification: they consider networks whose widths are very large, ideally approaching infinity. In particular, under suitable conditions, as the width increases, the network's behavior during training is expected to be captured by a meaningful limit.

One such type of analysis exploits exchangeability of neurons. Recent works [9, 22,32,34] show that under a suitable scaling limit, the learning dynamics of wide two-layer neural networks can be captured by a Wasserstein gradient flow of a probability measure over weights. In this limit, which is usually referred to as the *mean field* (MF) *limit*, the network weights evolve nonlinearly with time. The MF scaling of two-layer networks requires a certain normalization to be applied to the last layer, together with a learning rate that compensates for this normalization. The MF limit under the Wasserstein gradient flow formulation has led to a fruitful line of research that explains and uncovers interesting properties of two-layer networks, such as their optimization efficacy. Let us delve into a few further high-level details of the two-layer case, before discussing the interesting challenge in the multilayer case.

1.1. Two-layer MF network: a brief overview

Let us informally present a sampled subset of interesting results from this line of works. To fix ideas, we consider the usual two-layer neural network:

$$\hat{\mathbf{y}}_{\text{2-layer}}(x; W) = \frac{1}{n} \sum_{i=1}^{n} w_{2,i}\, \sigma(\langle w_{1,i}, x \rangle).$$

Here $x \in \mathbb{R}^d$ is the input, $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function, and $W = \{w_{1,i}, w_{2,i}\}_{i\in[n]}$ is the set of weights with $w_{1,i} \in \mathbb{R}^d$ and $w_{2,i} \in \mathbb{R}$, for $i \in [n]$, the set of integers from 1 and n. This network has n neurons; n is also referred to as the width of the network. The scaling factor $1/n$ is special to the MF scaling in two-layer networks and we will see its role shortly.

An "infinite-width" representation. One key idea in this line of work is to introduce the following representation:

$$\hat{y}_{\text{2-layer}}(x;\mu) = \int w_2\,\sigma(\langle w_1, x\rangle)\mu(dw_1, dw_2),$$

where μ is a probability measure on $\mathbb{R}^{d+1}$. It is easy to see that by choosing $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{w_{1,i},w_{2,i}}$ the empirical measure over the weights W, we have $\hat{y}_{\text{2-layer}}(x;\mu) = \hat{\mathbf{y}}_{\text{2-layer}}(x;W)$. This identification is possible thanks to the previously mentioned scaling factor n.

One way to rationalize this representation is as follows: there is a special symmetry in the two-layer neural network, in which

$$\hat{\mathbf{y}}_{\text{2-layer}}(x;W) = \hat{\mathbf{y}}_{\text{2-layer}}\big(x;\{w_{1,\Pi(i)}, w_{2,\Pi(i)}\}_{i\in[n]}\big)$$

for any permutation Π on the set of integers $[n]$. The representation via μ is a neat way to factor out this symmetry and capture the exchangeability of neurons. Of course, one is not restricted to only empirical measures for μ. Therefore, this representation allows one to reason about two-layer neural networks with *arbitrary* widths. In other words, it gives us the ability to take the infinite-width limit $n \to \infty$. This is an important observation that is central to this line of work.

The learning dynamics at infinite width: the MF limit. We are interested in understanding the learning dynamics of the network in the infinite-width limit. Consider the continuous-time gradient descent learning rule (with respect to W) for the loss ℓ:

$$\frac{d}{dt}W(t) = -n\nabla_W \mathbb{E}_Z[\ell(Y, \hat{\mathbf{y}}_{\text{2-layer}}(X;W(t)))].$$

Here t denotes the time and $Z = (X, Y)$ a random variable that represents the training data. Note that the scaling factor n compensates for the previously mentioned factor $1/n$ and therefore allows the learning update to be on the "correct" order. To see this, we rewrite the learning rule:

$$\begin{aligned}
\frac{d}{dt}w_{1,i}(t) &= -\mathbb{E}_Z\big[\partial_2\ell(Y, \hat{\mathbf{y}}_{\text{2-layer}}(X;W(t)))\cdot w_{2,i}(t)\sigma'(\langle w_{1,i}(t), X\rangle)X\big],\\
\frac{d}{dt}w_{2,i}(t) &= -\mathbb{E}_Z\big[\partial_2\ell(Y, \hat{\mathbf{y}}_{\text{2-layer}}(X;W(t)))\cdot \sigma(\langle w_{1,i}(t), X\rangle)\big],
\end{aligned}$$

where $\partial_2 \ell$ is the derivative of ℓ with respect to the second variable. In this form of the learning rule, one sees that if $w_{1,i}(t)$ and $w_{2,i}(t)$ all have magnitudes on order $O(1)$ independent of n, then so are their updates $\frac{d}{dt} w_{1,i}(t)$ and $\frac{d}{dt} w_{2,i}(t)$. Hence, if they are initialized to be on this order, one can expect to see the same order of weights and weight movements at any finite time t. This is a feature of the MF scaling.

Suppose that the initialization is sampled $(w_{1,i}(0), w_{2,i}(0)) \sim \mu_0$ independently for each $i \in [n]$, for a probability measure μ_0 on $\mathbb{R}^{d+1}$. We would like to study the empirical measure over the weights $W(t)$:

$$\mu_t^n = \frac{1}{n} \sum_{i=1}^{n} \delta_{w_{1,i}(t), w_{2,i}(t)}.$$

At $t = 0$, it is a standard result that $\mu_0^n \to \mu_0$ weakly as $n \to \infty$, under suitable regularity conditions. We are interested in a similar statement for any time t. To that end, we recall the "infinite-width" representation $\hat{y}_{\text{2-layer}}(x; \mu)$ and introduce the following distributional dynamics in the Wasserstein space of probability measures on $\mathbb{R}^{d+1}$:

$$\partial_t \mu_t(w_1, w_2) = \operatorname{div}\big(\mu_t(w_1, w_2) \nabla_{(w_1, w_2)} \Psi(w_1, w_2; \mu_t)\big),$$

in which $\Psi(w_1, w_2; \mu) = \mathbb{E}_Z[\partial_2 \ell(Y, \hat{y}_{\text{2-layer}}(X; \mu)) \cdot w_2 \sigma(\langle w_1, X \rangle)]$ and the initialization is μ_0. This dynamics is a Wasserstein gradient flow. Prior works [9, 22] prove the following type of result.

Theorem 1.1 (Two-layer MF network with distributional representation, $n \to \infty$, informal and simplified). *Under suitable regularity conditions, for any finite constant T, as $n \to \infty$, $\mu_t^n \to \mu_t$ weakly and uniformly over $t \in [0, T]$.*

The precise statement includes a quantitative convergence rate and more realistic learning rules, such as discrete-time stochastic gradient descent and other variations. Theorem 1.1 formalizes the notion of an infinite-width limit: we call μ_t the MF limit.

An application of the MF limit: *proving global convergence.* Theorem 1.1 conveys an interesting message: one can study the width-n neural network by analyzing the MF limit μ_t. One success story is the study of optimization efficacy. In particular, [9] proves the following type of result.

Theorem 1.2 (Two-layer MF network with distributional representation, $t \to \infty$, informal and simplified). *Suppose that the support of μ_0 is $\mathbb{R}^{d+1}$ (i.e., it has full support at initialization) and the loss ℓ is convex in the second variable. Under suitable regularity and convergence conditions, as $t \to \infty$,*

$$\mu_t \to \inf_{\mu} \mathbb{E}_Z[\ell(Y, \hat{y}_{\text{2-layer}}(X; \mu))].$$

This global convergence result affirms positively the message that taking $n \to \infty$ under the MF limit can lead to meaningful learning. Similar global convergence results have been established for different types of learning rules and, in special occasions, with quantitative convergence rates. To understand the significance of this result, we note a remarkable feature of the MF limit μ_t: it represents a genuinely nonlinear dynamics. To contrast the situation, other works (e.g., [8, 17]) show that under a different scaling, in the infinite-width limit, the neural network is equivalent to a parameterized model which is linear in its parameter. In that scaling regime, the learning dynamics hence simplifies into a linear dynamics, and consequently it is relatively clear how to attain global convergence using the usual convex optimization wisdom. The MF limit is distinct in this sense, but it also comes with a nontrivial problem: insights from convex optimization may no longer apply. This is indeed the case in the proof of the global convergence results of [9, 22].

We refer to [5, 24] for further overview discussions on two-layer MF neural networks. See also Section 9 for a partial list of works.

1.2. Multilayer MF network: the challenge and our contributions

Recall that an important milestone is to find a representation that allows to interpolate to an infinite-width limit. For two-layer networks, by exploiting exchangeability among neurons, one can achieve this goal and use the representation to successfully analyze properties of the neural networks in the infinite-width limit. In multilayer networks, exchangeability is, however, not a priori obvious and hence poses a highly non-trivial challenge. In particular, the presence of intermediate layers exhibits multiple symmetry groups with intertwined actions on the model. To illustrate the point, let us consider a simple three-layer fully-connected neural network which assumes the following form (modulo scaling factors):

$$\hat{\mathbf{y}}_{\text{3-layer}}(x) = \sum_{i=1}^{n_2} w_{3,i}\, \sigma\Big(\sum_{j=1}^{n_1} w_{2,ij}\, \sigma(\langle w_{1,j}, x\rangle)\Big),$$

for a set of parameters $\{w_{3,i}, w_{2,ij}, w_{1,j}\}_{i\in[n_2],\, j\in[n_1]}$. In the matrix notation,

$$\hat{\mathbf{y}}_{\text{3-layer}}(x) = w_3^\top \sigma(W_2 \sigma(W_1 x)).$$

Under any two permutations $\Pi_1 \colon [n_1] \to [n_1]$ and $\Pi_2 \colon [n_2] \to [n_2]$, we recognize

$$\hat{\mathbf{y}}_{\text{3-layer}}(x) = w_3^\top \Pi_2^\top \sigma(\Pi_2 W_2 \Pi_1^\top \sigma(\Pi_1 W_1 x)).$$

The fact that the weight matrix W_2 in the middle layer is under the simultaneous influence of both actions Π_1 and Π_2 is what makes the three-layer case specifically and

the multilayer case in general different from the two-layer case, more challenging and at the same time also a highly interesting problem. With this blocker on the strategy to extend the two-layer case, even the goal of obtaining a representation that captures networks with arbitrary widths becomes less approachable. Indeed, prior attempts in [4,23,35] arrive at quite complex solutions or require a certain strong assumption that leads to undesirable properties (see Section 9), and yet these attempts already have to do away with the Wasserstein gradient flow formulation.

In short, finding a suitable formulation that is amenable to the infinite-width limit-taking procedure, simultaneous at all layers, requires innovation beyond the Wasserstein gradient flow idea of the two-layer case. The formulation should faithfully describe settings where nonlinear and meaningful learning trajectories take place. To compound the difficulty, a useful formulation should lend a way to analyze properties of multilayer neural networks in the infinite-width limit, for instance, how well these networks could be optimized despite the strong presence of nonlinearity and the lack of convexity. These are the considerations one ought to keep in mind when tackling the challenge.

This work responses to this challenge with the proposal of a mathematically rigorous framework for the MF limit of multilayer neural networks. The framework is built on an innovative idea of a *neuronal embedding*. More importantly, using this framework, we prove several properties of multilayer networks, which incorporate new insights and ideas. Specifically, our key contributions can be summarized as follows.

In Sections 2, 3 and 4, we develop a framework for the MF limit of multilayer neural networks under stochastic gradient descent (SGD) training and suitable scalings. We introduce the concept of a neuronal embedding, which comprises of a non-evolving probability space that can embed neural networks of arbitrary widths. In this framework, the MF limit is described by a system of ordinary differential equations (ODEs), which govern the evolutions of different functions that represent the weights at different layers and are adapted to the given neuronal embedding. The complete framework is described in Section 2 and the well-posedness of the MF limit is proven in Section 3. Our main result in this thread is stated in Section 4, where the MF limit is proven to track closely characteristics of a wide multilayer network under SGD training, with quantitative bounds on the required widths.

In fact, our framework is quite general, admits a broad variety of initialization schemes (including, but not limited to, independent and identically distributed (i.i.d.) initializations) and operates in Hilbert spaces. This allows for firstly describing the MF behavior for generic multilayer setups (including fully-connected and convolutional networks in Euclidean spaces that are common in practice), and secondly obtaining *dimension-free* quantitative bounds.

In Section 5, using the neuronal embedding framework, we uncover strong degeneracy properties caused by i.i.d. initializations. Specifically, we prove that with at least four layers, the MF limits, and hence the neural networks, are substantially simplified under i.i.d. initializations: at an intermediate layer, each weight evolves as a function of only time, its own initialization and the initial biases associated with its connected neurons. An implication is that when the initial biases are constant, different intermediate layers evolve independently of each other. Remarkably, for common neural network architectures, all weights (or biases) at each intermediate layer then evolve by translation: they differ from their respective initializations by the same deterministic amount, and the effective number of parameters at each intermediate layer thus collapses to only one.

Our framework allows to study the optimization efficacy of multilayer neural networks trained under SGD in the infinite-width limit.

In particular, in Section 6, we prove convergence to the global optimum for two-layer and three-layer networks under i.i.d. initializations, with suitable regularity conditions and convergence assumptions. Some of these assumptions are mild and natural in neural network learning. The key convergence assumption in this section turns out to be necessary for global convergence to hold, i.e., it is impossible to attain global convergence if this convergence assumption fails.

In Section 7, avoiding the degeneracy effect of i.i.d. initializations, we prove global convergence for multilayer networks of arbitrary depths under a special type of correlated initializations and a similar set of assumptions. Here we introduce the new concept of *bidirectional diversity*.

In Section 8, we also establish global convergence in the above settings under Morse–Sard conditions that are usually assumed in the literature for MF two-layer networks. This demonstrates flexibility of our framework: it can handle situations where the two-layer Wasserstein gradient flow formulation works, as well as situations where such formulation finds difficulty.

Two novel features that our global convergence results have in common are firstly the role of a certain universal approximation property which is natural for nonlinear neural networks, and secondly the admission of non-convex losses. Importantly, the universal approximation property is shown to hold at any finite training time (but not necessarily at infinite time) via topological invariance arguments. These new insights signal the departure from conventional wisdoms of convex optimization.

The idea of bidirectional diversity that we introduce in Section 7 strikes directly to the universal approximation insight. Roughly speaking, it helps "propagating" the universal approximation property from the first layer to the second last layer. This is to be contrasted with i.i.d.-initialized networks: universal approximation at the first layer suffices when there are few layers, but as the number of layers increases, due to degeneracy by i.i.d. initializations, the middle layers become a bottleneck that generally

prohibits universal approximation to be propagated. Bidirectional diversity aims to break this bottleneck.

We defer a more technical discussion on the related literature to Section 9. Proofs of several intermediate results are deferred to the appendices. Readers who are interested in global convergence of networks with more than three layers may skip directly to Sections 7 and 8, which we have made relatively self-contained with minimal references to the previous sections.

1.3. Notations

For an integer n, we use $[n]$ to denote the set $\{1, \dots, n\}$. We shall use $\langle \cdot, \cdot \rangle$ and $|\cdot|$ to indicate, respectively, the inner product and its induced norm for a Hilbert space, and $|\cdot|$ to indicate the absolute value for $\mathbb{R}$. We use $\sigma_{\text{alg}}(U)$ to denote the sigma-algebra generated by a random variable U. We write $\text{cl}(S)$ to denote the closure of a set S in a topological space. We use K to denote a generic absolute constant that may change from line to line. For a probability space $(\Omega, \mathcal{F}, P)$, we will suppress the presence of the sigma-algebra $\mathcal{F}$ wherever unimportant. Given two events $\mathcal{E}$ and $\mathcal{E}'$, we say that $\mathcal{E}'$ occurs with probability at least $1 - \delta$ on the event $\mathcal{E}$ and write $\mathbb{P}(\mathcal{E}'; \mathcal{E}) \geq 1 - \delta$ if $\mathbb{P}((\neg \mathcal{E}') \cap \mathcal{E}) \leq \delta$.

2. A general framework

In this section, we describe our setup of a general multilayer neural network with a generalized (stochastic) learning dynamics. In particular, it covers several common neural network architectures as well as the SGD training dynamics. We then describe the corresponding MF limit.

2.1. Multilayer neural network and generalized learning dynamics

We consider the following generalized neural network with L layers:

$$\hat{\mathbf{y}}(k, x) \equiv \hat{\mathbf{y}}(x; \mathbf{W}(k)) = \phi_{L+1}(\mathbf{H}_L(k, x, 1)), \tag{2.1}$$

in which we define $\mathbf{H}_L(k, x, 1)$ recursively:

$$\begin{aligned}
\mathbf{H}_1(k, x, j_1) &\equiv \mathbf{H}_1(x, j_1; \mathbf{W}(k)) = \phi_1(\mathbf{w}_1(k, j_1), x), \quad j_1 \in [n_1], \\
\mathbf{H}_i(k, x, j_i) &\equiv \mathbf{H}_i(x, j_i; \mathbf{W}(k)) \\
&= \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \phi_i(\mathbf{w}_i(k, j_{i-1}, j_i), \mathbf{b}_i(k, j_i), \mathbf{H}_{i-1}(k, x, j_{i-1})), \\
&\qquad\qquad j_i \in [n_i],\ i = 2, \dots, L.
\end{aligned}$$

The above equations describe the forward pass in the neural network. We explain the quantities in the following:

- $x \in \mathbb{X}$ is the input, and $\mathbb{X}$ is the input space.
- $k \in \mathbb{N}_{\geq 0}$ is the (discrete) time.
- $\mathbf{W}(k) = \{\mathbf{w}_1(k,\cdot), \mathbf{w}_i(k,\cdot,\cdot), \mathbf{b}_i(k,\cdot),\ i = 2,\dots,L\}$ is the collection of neural network parameters (weights and biases) at time k.
- $\mathbf{w}_1 \colon \mathbb{N}_{\geq 0} \times [n_1] \to \mathbb{W}_1$ is the weight of the first layer (which also includes the bias). Similarly for $i = 2,\dots,L$, $\mathbf{w}_i \colon \mathbb{N}_{\geq 0} \times [n_{i-1}] \times [n_i] \to \mathbb{W}_i$ and $\mathbf{b}_i \colon \mathbb{N}_{\geq 0} \times [n_i] \to \mathbb{B}_i$ are the weight and bias of the i-th layer. Here n_i is the number of neurons at the i-th layer, $\mathbb{W}_i$ and $\mathbb{B}_i$ are separable Hilbert spaces, and we take $n_L = 1$.
- $\phi_1 \colon \mathbb{W}_1 \times \mathbb{X} \to \mathbb{H}_1$, $\phi_i \colon \mathbb{W}_i \times \mathbb{B}_i \times \mathbb{H}_{i-1} \to \mathbb{H}_i$ for $i = 2,\dots,L$, $\phi_{L+1} \colon \mathbb{H}_L \to \hat{\mathbb{Y}}$, where again $\mathbb{H}_i$ and $\hat{\mathbb{Y}}$ are separable Hilbert spaces.

In other words, the network $\hat{\mathbf{y}}(k,x)$ is a state-dependent mapping that takes x as input and it is dependent on the state $\mathbf{W}(k)$, which is allowed to vary with time k.

The network is trained by the following (discrete-time) stochastic learning dynamics. At each time k, we draw independently a data sample $z(k) = (x(k), y(k)) \sim \mathcal{P}$, where $\mathcal{P}$ is the data distribution on $\mathbb{X} \times \mathbb{Y}$ and $y(k) \in \mathbb{Y}$, a separable Hilbert space. Given an initialization $\mathbf{W}(0)$, we update $\mathbf{W}(k)$ into $\mathbf{W}(k+1)$ as follows:

$$\begin{aligned}
\mathbf{w}_1(k+1, j_1) &= \mathbf{w}_1(k, j_1) - \epsilon \xi_1^{\mathbf{w}}(k\epsilon) \Delta_1^{\mathbf{w}}(k, z(k), j_1) \quad \text{for all } j_1 \in [n_1], \\
\mathbf{w}_i(k+1, j_{i-1}, j_i) &= \mathbf{w}_i(k, j_{i-1}, j_i) - \epsilon \xi_i^{\mathbf{w}}(k\epsilon) \Delta_i^{\mathbf{w}}(k, z(k), j_{i-1}, j_i), \\
\mathbf{b}_i(k+1, j_i) &= \mathbf{b}_i(k, j_i) - \epsilon \xi_i^{\mathbf{b}}(k\epsilon) \Delta_i^{\mathbf{b}}(k, z(k), j_i) \\
&\qquad \text{for all } j_{i-1} \in [n_{i-1}],\ j_i \in [n_i],\ i = 2,\dots,L.
\end{aligned}$$

Here $\epsilon \in \mathbb{R}_{>0}$ is the learning rate, and $\xi_i^{\mathbf{w}}$ and $\xi_i^{\mathbf{b}}$ are mappings from $\mathbb{R}$ to $\mathbb{R}$, representing the different learning rate schedules for each of the weights and biases. Note that we allow the learning rate schedules to take non-positive values.

To define the updates $\Delta_i^{\mathbf{w}}$ and $\Delta_i^{\mathbf{b}}$ requires additional definitions. Firstly, for $z = (x, y)$, we define

$$\Delta_L^{\mathbf{H}}(k, z, 1) \equiv \Delta_L^{\mathbf{H}}(z, 1; \mathbf{W}(k)) = \sigma_L^{\mathbf{H}}(y, \hat{\mathbf{y}}(k, x), \mathbf{H}_L(k, x, 1)).$$

Then we define recursively

$$\begin{aligned}
\Delta_i^{\mathbf{w}}(k, z, j_{i-1}, j_i) &\equiv \Delta_i^{\mathbf{w}}(z, j_{i-1}, j_i; \mathbf{W}(k)) \\
&= \sigma_i^{\mathbf{w}}\big(\Delta_i^{\mathbf{H}}(k, z, j_i), \mathbf{w}_i(k, j_{i-1}, j_i), \mathbf{b}_i(k, j_i), \\
&\qquad \mathbf{H}_i(k, x, j_i), \mathbf{H}_{i-1}(k, x, j_{i-1})\big),
\end{aligned}$$

$$\begin{aligned}
\Delta_i^{\mathbf{b}}(k,z,j_i) &\equiv \Delta_i^{\mathbf{b}}(z,j_i;\mathbf{W}(k)) \\
&= \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \sigma_i^{\mathbf{b}}\big(\Delta_i^{\mathbf{H}}(k,z,j_i), \mathbf{w}_i(k,j_{i-1},j_i), \mathbf{b}_i(k,j_i), \\
&\qquad\qquad \mathbf{H}_i(k,x,j_i), \mathbf{H}_{i-1}(k,x,j_{i-1})\big), \\
\Delta_{i-1}^{\mathbf{H}}(k,z,j_{i-1}) &\equiv \Delta_{i-1}^{\mathbf{H}}(z,j_{i-1};\mathbf{W}(k)) \\
&= \frac{1}{n_i} \sum_{j_i=1}^{n_i} \sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{\mathbf{H}}(k,z,j_i), \mathbf{w}_i(k,j_{i-1},j_i), \mathbf{b}_i(k,j_i), \\
&\qquad\qquad \mathbf{H}_i(k,x,j_i), \mathbf{H}_{i-1}(k,x,j_{i-1})\big), \quad i = L,\dots,2, \\
\Delta_1^{\mathbf{w}}(k,z,j_1) &\equiv \Delta_1^{\mathbf{w}}(z,j_1;\mathbf{W}(k)) = \sigma_1^{\mathbf{w}}(\Delta_1^{\mathbf{H}}(k,z,j_i), \mathbf{w}_1(k,j_1), x),
\end{aligned}$$

in which the functions are

$$\begin{aligned}
\sigma_L^{\mathbf{H}} &: \mathbb{Y} \times \hat{\mathbb{Y}} \times \mathbb{H}_L \to \hat{\mathbb{H}}_L, \\
\sigma_i^{\mathbf{w}} &: \hat{\mathbb{H}}_i \times \mathbb{W}_i \times \mathbb{B}_i \times \mathbb{H}_i \times \mathbb{H}_{i-1} \to \mathbb{W}_i, \\
\sigma_i^{\mathbf{b}} &: \hat{\mathbb{H}}_i \times \mathbb{W}_i \times \mathbb{B}_i \times \mathbb{H}_i \times \mathbb{H}_{i-1} \to \mathbb{B}_i, \\
\sigma_{i-1}^{\mathbf{H}} &: \hat{\mathbb{H}}_i \times \mathbb{W}_i \times \mathbb{B}_i \times \mathbb{H}_i \times \mathbb{H}_{i-1} \to \hat{\mathbb{H}}_{i-1}, \quad i = L,\dots,2, \\
\sigma_1^{\mathbf{w}} &: \hat{\mathbb{H}}_1 \times \mathbb{W}_1 \times \mathbb{X} \to \mathbb{W}_1,
\end{aligned}$$

for separable Hilbert spaces $\hat{\mathbb{H}}_i$. Note that the above equations describe the backward pass in the neural network.

The introduced framework is quite general, while certain assumptions can be further relaxed. We observe that several common network architectures and training processes can be cast as special cases.

Example 2.1 (Fully-connected networks). We describe the simple setting of a fully-connected network with 1-dimensional output and an activation function $\varphi_i : \mathbb{R} \to \mathbb{R}$ at the i-th layer. Specifically, the network output assumes the form

$$\hat{\mathbf{y}}(x;\mathbf{W}) = \frac{1}{n_{L-1}} \Big\langle \mathbf{w}_L, \varphi_{L-1}\Big(\mathbf{b}_{L-1} + \frac{1}{n_{L-2}} \mathbf{W}_{L-1}\varphi_{L-2}\Big(\cdots \varphi_1\Big(\mathbf{W}_1 \begin{bmatrix} x \\ 1 \end{bmatrix}\Big)\Big)\Big)\Big\rangle + \mathbf{b}_L,$$

in which $x \in \mathbb{R}^d$, $\mathbf{W} = \{\mathbf{w}_L, \mathbf{W}_{L-1}, \dots, \mathbf{W}_1, \mathbf{b}_L, \dots, \mathbf{b}_2\}$, $\mathbf{w}_L \in \mathbb{R}^{n_{L-1}}$, $\mathbf{W}_i \in \mathbb{R}^{n_i \times n_{i-1}}$, with $n_0 = d + 1$, $n_L = 1$ and $\mathbf{b}_i \in \mathbb{R}^{n_i}$. This case fits into our framework with $\mathbb{X} = \mathbb{R}^d$, $\mathbb{W}_1 = \mathbb{R}^{d+1}$, $\mathbb{H}_1 = \mathbb{R}$ and $\mathbb{W}_i = \mathbb{B}_i = \mathbb{H}_i = \mathbb{Y} = \hat{\mathbb{Y}} = \mathbb{R}$ for $2 \le i \le L$. We also have

$$\begin{aligned}
\phi_1(w,x) &= \langle w_{1:d}, x\rangle + w_{d+1}, \\
\phi_i(w,b,h) &= w\varphi_{i-1}(h) + b, \quad 2 \le i \le L, \\
\phi_{L+1}(h) &= h.
\end{aligned}$$

Consider the regularized loss function

$$\begin{aligned}\mathrm{Loss}(\mathbf{W}; z) = \mathscr{L}(y, \hat{y}(x; \mathbf{W})) + \frac{1}{n_1}\sum_{j_1=1}^{n_1}\Phi_1(w_{1,j_1}) + \frac{1}{n_{L-1}}\sum_{j_{L-1}=1}^{n_{L-1}}\Phi_L(w_{L,j_{L-1}}) \\ + \sum_{i=2}^{L-1}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\Phi_i(w_{i,j_{i-1}j_i})\Big) + \sum_{i=2}^{L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\Psi_i(b_{i,j_i})\Big),\end{aligned}$$

where $\mathscr{L}\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}_{\geq 0}$, $\Phi_i\colon \mathbb{R}\to\mathbb{R}_{\geq 0}$ for $i \geq 1$, $\Phi_1\colon \mathbb{R}^{d+1}\to\mathbb{R}_{\geq 0}$, $\Psi_i\colon \mathbb{R}\to\mathbb{R}_{\geq 0}$, w_{1,j_1} is the j_1-th row of $\mathbf{W}_1$, $w_{i,j_{i-1}j_i}$ is the (j_{i-1}, j_i)-th entry of $\mathbf{W}_i$ for $2 \leq i \leq L-1$, $w_{L,j_{L-1}}$ is the j_{L-1}-th entry of $\mathbf{w}_L$, and b_{i,j_i} is the j_i-th entry of $\mathbf{b}_i$. If we train the network by SGD with respect to this loss, then $\hat{\mathbb{H}}_i = \mathbb{R}$ and

$$\begin{aligned}\sigma_L^{\boldsymbol{H}}(y, \hat{y}, h) &= \partial_2\mathscr{L}(y, \hat{y}),\\ \sigma_i^{\mathbf{w}}(\Delta, w, b, g, h) &= \Delta\varphi_{i-1}(h) + \Phi_i'(w),\\ \sigma_i^{\mathbf{b}}(\Delta, w, b, g, h) &= \Delta + \Psi_i'(b),\\ \sigma_{i-1}^{\mathbf{H}}(\Delta, w, b, g, h) &= \Delta w\varphi_{i-1}'(h), \quad 2 \leq i \leq L,\\ \sigma_1^{\mathbf{w}}(\Delta, w, x) &= \Delta\begin{bmatrix} x \\ 1 \end{bmatrix} + \nabla\Phi_1(w).\end{aligned}$$

Observe that when there is no regularization (i.e., no Φ_i and Ψ_i), $\sigma_i^{\mathbf{w}}(\Delta, w, b, g, h)$ is independent of w and b, and the same holds for $\sigma_i^{\mathbf{b}}$ and $\sigma_1^{\mathbf{w}}$.

Example 2.2 (Convolutional networks). Our framework can also describe networks that are not of the fully-connected type. For illustration, we consider the first two layers of a convolutional network with an activation $\varphi\colon \mathbb{R}\to\mathbb{R}$ and pooling operation $\mathrm{pool}(\cdot)$; a description of the complete network (which may contain fully-connected layers) can be done in a similar fashion to Example 2.1. Here $\mathbb{X} = (\mathbb{R}^{p\times p})^{n_c}$, where in the context of a square image input, p is the number of pixels per row and n_c is the number of channels (which is 3 for RGB images and 1 for gray-scale images). We take $\mathbb{W}_1 = \mathbb{R}^{f_1\times f_1\times n_c}\times\mathbb{R}$ and $\mathbb{W}_2 = \mathbb{R}^{f_2\times f_2}$, where f_1 and f_2 are the filter sizes, $\mathbb{B}_2 = \mathbb{R}$, $\mathbb{H}_1 = \mathbb{R}^{p_1\times p_1}$ and $\mathbb{H}_2 = \mathbb{R}^{p_2\times p_2}$. Then

$$\begin{aligned}\phi_1((w, b), x) &= w * x + b\mathbf{1}_{p_1},\\ \phi_2(w, b, h) &= w * \mathrm{pool}(\varphi(h)) + b\mathbf{1}_{p_2},\end{aligned}$$

where $*$ denotes (strided) convolution and $\mathbf{1}_{p_i}$ is an all-one matrix in $\mathbb{R}^{p_i\times p_i}$. The dimensions p_1 and p_2 are determined by the actual convolution operation, its stride size, its padding type and the input size. In this context, n_1 and n_2 are the numbers of filters at the first and second layer, respectively. One can also specify the forms of $\sigma_i^{\mathbf{w}}$, $\sigma_i^{\mathbf{b}}$ and $\sigma_i^{\mathbf{H}}$ upon the choice of a loss function, with SGD training.

The examples of fully-connected and convolutional neural networks serve as the main motivation to study the generalized neural network model as described. In both of these examples, the spaces are finite-dimensional Euclidean spaces, while in the generalized model, the spaces are allowed to be infinite-dimensional. Similarly, while SGD with respect to a loss function is the typical choice of learning dynamics for these examples, in our framework, the learning dynamics is more general. We shall see that the key ideas hold regardless of the specific details. In particular, the ultimate goal is to understand the properties of $\mathbf{W}(k)$ in the limit of large n_i and small ϵ, via a limiting object that is well-defined and has an explicit form. To this end, we introduce the mean field limit in the next section.

2.2. Mean field limit

We now describe the mean field (MF) limit. Given a probability space $(\Omega, \mathcal{F}, P) = \prod_{i=1}^{L}(\Omega_i, \mathcal{F}_i, P_i)$ with $\Omega_L = \{1\}$, we independently sample $C_i \sim P_i$, $1 \leq i \leq L$. From here onwards, we hide the sigma-algebras $\mathcal{F}$, $\mathcal{F}_i$ wherever unimportant. In the following, we use $\mathbb{E}_{C_i}$ to denote the expectation with respect to the random variable $C_i \sim P_i$ and c_i to denote a dummy variable $c_i \in \Omega_i$. The space (Ω, P) is key to our MF formulation and is referred to as the *neuronal ensemble*. The choice of the neuronal ensemble provides a bridge between the earlier described neural network and the MF limit; this connection shall be established later in Section 4. For the moment we treat the MF limit as an independent object from the neural network.

Given the neuronal ensemble, we obtain the MF limit as follows. It entails the following quantity:

$$\hat{y}(t, x) \equiv \hat{y}(x; W(t)) = \phi_{L+1}(H_L(t, x, 1)),$$

in which $H_L(t, x, 1)$ is computed recursively:

$$\begin{aligned} H_1(t, x, c_1) &\equiv H_1(x, c_1; W(t)) = \phi_1(w_1(t, c_1), x) \quad \text{for all } c_1 \in \Omega_1, \\ H_i(t, x, c_i) &\equiv H_i(x, c_i; W(t)) \\ &= \mathbb{E}_{C_{i-1}}[\phi_i(w_i(t, C_{i-1}, c_i), b_i(t, c_i), H_{i-1}(t, x, C_{i-1}))] \\ &\quad \text{for all } c_i \in \Omega_i,\ i = 2, \ldots, L. \end{aligned}$$

This corresponds to the forward pass of the neural network. We note the similarity with the corresponding quantities of the neural network:

- $x \in \mathbb{X}$ is the input and $t \in \mathbb{R}_{\geq 0}$ is the (continuous) time.
- $W(t) = \{w_1(t, \cdot), w_i(t, \cdot, \cdot), b_i(t, \cdot),\ i = 2, \ldots, L\}$ is the collection of the MF parameters at time t.

- $w_1 \colon \mathbb{R}_{\geq 0} \times \Omega_1 \to \mathbb{W}_1$, and for $i = 2, \dots, L$, $w_i \colon \mathbb{R}_{\geq 0} \times \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$ and $b_i \colon \mathbb{R}_{\geq 0} \times \Omega_i \to \mathbb{B}_i$.

In correspondence with the neural network's learning dynamics for $\mathbf{W}(k)$, the MF limit also entails a continuous-time evolution dynamics for $W(t)$. This dynamics takes the form of a system of ODEs, which we refer to as the *MF ODEs*, given an initialization $W(0)$:

$$\begin{aligned}
\frac{\partial}{\partial t} w_1(t, c_1) &= -\xi_1^{\mathbf{w}}(t) \mathbb{E}_Z[\Delta_1^w(t, Z, c_1)] \quad \text{for all } c_1 \in \Omega_1, \\
\frac{\partial}{\partial t} w_i(t, c_{i-1}, c_i) &= -\xi_i^{\mathbf{w}}(t) \mathbb{E}_Z[\Delta_i^w(t, Z, c_{i-1}, c_i)], \\
\frac{\partial}{\partial t} b_i(t, c_i) &= -\xi_i^{\mathbf{b}}(t) \mathbb{E}_Z[\Delta_i^b(t, Z, c_i)] \quad \text{for all } c_{i-1} \in \Omega_{i-1},\ c_i \in \Omega_i,
\end{aligned}$$

for $i = 2, \dots, L$, where $\mathbb{E}_Z$ denotes the expectation with respect to the data $Z = (X, Y) \sim \mathcal{P}$, and the update quantities are defined by the following recursion:

$$\begin{aligned}
\Delta_L^H(t, z, 1) &\equiv \Delta_L^H(z, 1; W(t)) = \sigma_L^{\mathbf{H}}(y, \hat{y}(t, x), H_L(t, x, 1)), \\
\Delta_i^w(t, z, c_{i-1}, c_i) &\equiv \Delta_i^w(z, c_{i-1}, c_i; W(t)) \\
&= \sigma_i^{\mathbf{w}}\big(\Delta_i^H(t, z, c_i), w_i(t, c_{i-1}, c_i), b_i(t, c_i), \\
&\qquad H_i(t, x, c_i), H_{i-1}(t, x, c_{i-1})\big), \\
\Delta_i^b(t, z, c_i) &\equiv \Delta_i^b(z, c_i; W(t)) \\
&= \mathbb{E}_{C_{i-1}}\big[\sigma_i^{\mathbf{b}}\big(\Delta_i^H(t, z, c_i), w_i(t, C_{i-1}, c_i), b_i(t, c_i), \\
&\qquad H_i(t, x, c_i), H_{i-1}(t, x, C_{i-1})\big)\big], \\
\Delta_{i-1}^H(t, z, c_{i-1}) &\equiv \Delta_{i-1}^H(z, c_{i-1}; W(t)) \\
&= \mathbb{E}_{C_i}\big[\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^H(t, z, C_i), w_i(t, c_{i-1}, C_i), b_i(t, C_i), \\
&\qquad H_i(t, x, C_i), H_{i-1}(t, x, c_{i-1})\big)\big], \quad i = L, \dots, 2, \\
\Delta_1^w(t, z, c_1) &\equiv \Delta_1^w(z, c_1; W(t)) = \sigma_1^{\mathbf{w}}(\Delta_1^H(t, z, c_1), w_1(t, c_1), x).
\end{aligned}$$

This recursion corresponds to the backward pass of the neural network.

Remark 2.3. The definition of a MF limit model $\hat{y}(t, x)$ based on the neuronal ensemble (Ω, P) gives a way to define a large class of neural networks that encapsulates networks of arbitrary sizes. More specifically, let us write $W = \{w_1, w_i, b_i : i = 2, \dots, L\}$ in place of $W(t)$ and $\hat{y}(x; W, \Omega, P)$ in place of $\hat{y}(t, x)$ to ignore the time t and make explicit the dependency on the neuronal ensemble. Similarly, here let us also write $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_i, \mathbf{b}_i : i = 2, \dots, L\}$ in place of $\mathbf{W}(k)$ and $\hat{\mathbf{y}}(x; \mathbf{W}, n_1, \dots, n_L)$ in place of $\hat{\mathbf{y}}(k, x)$. Then, by defining the class $\mathtt{NN}_\infty = \{\hat{y}(\cdot; W, \Omega, P)\}_{W,\Omega,P}$ that is indexed by (W, Ω, P) while fixing other parameters (such as the number of layers L), one sees that any finite-sized neural network $\hat{\mathbf{y}}(\cdot; \mathbf{W}, n_1, \dots, n_L)$ belongs to $\mathtt{NN}_\infty$.

This correspondence can be seen by the following identification: $\Omega = \prod_{i=1}^{L} \Omega_i$ with $\Omega_i = [n_i]$, $P = \prod_{i=1}^{L} P_i$, with P_i a uniform measure on $[n_i]$, and

$$\begin{aligned} w_i(j_{i-1}, j_i) &= \mathbf{w}_i(j_{i-1}, j_i) \quad \text{for all } j_{i-1} \in \Omega_{i-1} = [n_{i-1}],\ j_i \in \Omega_i = [n_i], \\ b_i(j_i) &= \mathbf{b}_i(j_i) \quad \text{for all } j_i \in \Omega_i = [n_i], \\ w_1(j_1) &= \mathbf{w}_1(j_1) \quad \text{for all } j_1 \in \Omega_1 = [n_1], \end{aligned}$$

for $2 \le i \le L$. In particular, there exists $\hat{y}(\cdot; W, \Omega, P) \in \mathsf{NN}_\infty$ such that

$$\hat{y}(\cdot; W, \Omega, P) = \hat{\mathbf{y}}(\cdot; \mathbf{W}, n_1, \ldots, n_L).$$

More generally one may observe that a similar correspondence holds for both the forward pass and the backward pass, for example,

$$\begin{aligned} H_i(x, j_i; W, \Omega, P) &= \mathbf{H}_i(x, j_i; \mathbf{W}, n_1, \ldots, n_L), \\ \Delta_i^H(z, j_i; W, \Omega, P) &= \Delta_i^{\mathbf{H}}(z, j_i; \mathbf{W}, n_1, \ldots, n_L) \quad \text{for all } j_i \in \Omega_i = [n_i], \end{aligned}$$

where the quantities are rewritten forms of $H_i(t, x, c_i)$, $\Delta_i^H(t, z, c_i)$, $\mathbf{H}_i(k, x, j_i)$ and $\Delta_i^{\mathbf{H}}(k, z, c_i)$, respectively. As such, roughly speaking, the dynamics of any finite-sized neural network can be identified with a MF dynamics, modulo the differences in time discretization and stochastic sampling of the data. The same observation is made in [13], which instead studies it from the function space approximation perspective.

2.3. Preliminaries

We describe several preliminaries that are necessary for the next steps. First we consider several structural assumptions.

Assumption 2.4. The learning rate schedules are bounded and Lipschitz:

$$\begin{aligned} \max_{1 \le i \le L} |\xi_i^{\mathbf{w}}(t)|, \max_{2 \le i \le L} |\xi_i^{\mathbf{b}}(t)| \le K, \\ \max_{1 \le i \le L} |\xi_i^{\mathbf{w}}(t) - \xi_i^{\mathbf{w}}(t')|, \max_{2 \le i \le L} |\xi_i^{\mathbf{b}}(t) - \xi_i^{\mathbf{b}}(t')| \le K|t - t'|. \end{aligned}$$

Assumption 2.5 (Forward pass assumptions). The function ϕ_1 satisfies

$$|\phi_1(w, x) - \phi_1(w', x)| \le K|w - w'|,$$

for all $w, w' \in \mathbb{W}_1$ and for $\mathcal{P}$-almost every x. For $i = 2, \ldots, L$, ϕ_i satisfies

$$\begin{aligned} |\phi_i(w, b, h)| &\le K(1 + |w| + |b|), \\ |\phi_i(w, b, h) - \phi_i(w', b', h')| &\le K(1 + |w| + |w'| + |b| + |b'|)|h - h'| \\ &\quad + K(|w - w'| + |b - b'|), \end{aligned}$$

for all $w, w' \in \mathbb{W}_i$, $b, b' \in \mathbb{B}_i$, and $h, h' \in \mathbb{H}_{i-1}$. Finally, ϕ_{L+1} satisfies

$$|\phi_{L+1}(h) - \phi_{L+1}(h')| \leq K|h - h'|,$$

for all $h, h' \in \mathbb{H}_L$.

Assumption 2.6 (Backward pass assumptions). The function $\sigma_1^{\mathbf{w}}$ satisfies

$$|\sigma_1^{\mathbf{w}}(\Delta, w, x)| \leq K(1 + |\Delta|),$$
$$|\sigma_1^{\mathbf{w}}(\Delta, w, x) - \sigma_1^{\mathbf{w}}(\Delta', w', x)| \leq K(|\Delta - \Delta'| + |w - w'|),$$

for all $w, w' \in \mathbb{W}_1$, $\Delta, \Delta' \in \hat{\mathbb{H}}_1$ and for $\mathcal{P}$-almost every x. For $i = 2, \ldots, L$, $\sigma_i^{\mathbf{w}}$ and $\sigma_i^{\mathbf{b}}$ satisfy the following growth bounds:

$$\max(|\sigma_i^{\mathbf{w}}(\Delta, w, b, g, h)|, |\sigma_i^{\mathbf{b}}(\Delta, w, b, g, h)|) \leq K(1 + |\Delta|),$$

as well as the following perturbation bounds:

$$\begin{aligned}
&\max\big(|\sigma_i^{\mathbf{w}}(\Delta, w, b, g, h) - \sigma_i^{\mathbf{w}}(\Delta', w', b', g', h')|, \\
&\qquad |\sigma_i^{\mathbf{b}}(\Delta, w, b, g, h) - \sigma_i^{\mathbf{b}}(\Delta', w', b', g', h')|\big) \\
&\leq K(1 + |\Delta| + |\Delta'|)|h - h'| + K(|\Delta - \Delta'| + |w - w'| + |b - b'| + |g - g'|).
\end{aligned}$$

For $i = 2, \ldots, L$, $\sigma_{i-1}^{\mathbf{H}}$ satisfies the growth bound

$$|\sigma_{i-1}^{\mathbf{H}}(\Delta, w, b, g, h)| \leq K(1 + |\Delta|)(1 + |w| + |b|),$$

and the perturbation bound

$$\begin{aligned}
&|\sigma_{i-1}^{\mathbf{H}}(\Delta, w, b, g, h) - \sigma_{i-1}^{\mathbf{H}}(\Delta', w', b', g', h')| \\
&\quad \leq K(1 + |w| + |w'| + |b| + |b'|)|\Delta - \Delta'| \\
&\qquad + K(1 + |\Delta| + |\Delta'|)(|w - w'| + |b - b'|) \\
&\qquad + K(1 + |\Delta| + |\Delta'|)(1 + |w| + |w'| + |b| + |b'|)(|g - g'| + |h - h'|).
\end{aligned}$$

Finally, $\sigma_L^{\mathbf{H}}$ satisfies

$$|\sigma_L^{\mathbf{H}}(y, \hat{y}, h)| \leq K, \quad |\sigma_L^{\mathbf{H}}(y, \hat{y}, h) - \sigma_L^{\mathbf{H}}(y, \hat{y}', h')| \leq K(|h - h'| + |\hat{y} - \hat{y}'|),$$

for $\mathcal{P}$-almost every y.

Remark 2.7. We remark that these assumptions can be relaxed, e.g., $\phi_i(w, b, h)$ may be allowed to grow super-linearly with the variables, at the expense of suitable additional assumptions.[1] Here we pay attention to a simpler setting, which covers neural network setups of interest that are relevant to Sections 6 and 7.

[1] Indeed, this has been done in our previous iterate of the paper, posted on arXiv.

We also equip the neural network and its MF limit with several norms. In particular, we define, for the neural network parameters,

$$|||\mathbf{w}_i|||_t = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\le t}|\mathbf{w}_i(\lfloor s/\epsilon\rfloor, j_{i-1}, j_i)|^{50}\Big)^{1/50},$$

$$|||\mathbf{b}_i|||_t = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}|\mathbf{b}_i(\lfloor s/\epsilon\rfloor, j_i)|^{50}\Big)^{1/50}, \qquad i = 2,\dots,L,$$

$$|||\mathbf{w}_1|||_t = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{s\le t}|\mathbf{w}_1(\lfloor s/\epsilon\rfloor, j_1)|^{50}\Big)^{1/50}.$$

We also introduce the notation:

$$|||\mathbf{W}|||_t = \max\big(\max_{1\le i\le L}|||\mathbf{w}_i|||_t, \max_{2\le i\le L}|||\mathbf{b}_i\ |||_t\big).$$

We also have similarly for the MF limit:

$$|||w_i|||_t = \mathbb{E}\big[\sup_{s\le t}|w_i(s, C_{i-1}, C_i)|^{50}\big]^{1/50},$$

$$|||b_i|||_t = \mathbb{E}\big[\sup_{s\le t}|b_i(s, C_i)|^{50}\big]^{1/50}, \quad i = 2,\dots,L,$$

$$|||w_1|||_t = \mathbb{E}\big[\sup_{s\le t}|w_1(s, C_1)|^{50}\big]^{1/50},$$

as well as

$$|||W|||_t = \max\big(\max_{1\le i\le L}|||w_i|||_t, \max_{2\le i\le L}|||b_i\ |||_t\big).$$

For convenience, let us define the quantities:

$$\mathsf{max}_t^w(W) = \max_{2\le i\le L}\sup_{s\le t}|w_i(s, C_{i-1}, C_i)|,$$

$$\mathsf{max}_t^b(W) = \max_{2\le i\le L}\sup_{s\le t}|b_i(s, C_i)|,$$

which are random variables. Note that $\mathsf{max}_t^w(W)$ does not involve w_1.

For a set of MF parameters W, we define

$$\|W\|_t = \max\big(\max_{1\le i\le L}\|w_i\|_t, \max_{2\le i\le L}\|b_i\|_t\big),$$

$$\|w_i\|_t = \mathbb{E}\big[\sup_{s\le t}|w_i(s, C_{i-1}, C_i)|^2\big]^{1/2},$$

$$\|b_i\|_t = \mathbb{E}\big[\sup_{s\le t}|b_i(s, C_i)|^2\big]^{1/2}, \quad i = 2,\dots,L,$$

$$\|w_1\|_t = \mathbb{E}\big[\sup_{s\le t}|w_1(s, C_1)|^2\big]^{1/2}.$$

Note that this defines a norm on the space of MF parameters. As such, we can define the following distance for two sets of MF parameters W and W':

$$\begin{aligned}
\|W - W'\|_t &= \max\big(\max_{1\le i\le L} \|w_i - w_i'\|_t, \max_{2\le i\le L} \|b_i - b_i'\|_t\big),\\
\|w_i - w_i'\|_t &= \mathbb{E}\big[\sup_{s\le t} |w_i(s, C_{i-1}, C_i) - w_i'(s, C_{i-1}, C_i)|^2\big]^{1/2},\\
\|b_i - b_i'\|_t &= \mathbb{E}\big[\sup_{s\le t} |b_i(s, C_i) - b_i'(s, C_i)|^2\big]^{1/2}, \quad i = 2, \dots, L,\\
\|w_1 - w_1'\|_t &= \mathbb{E}\big[\sup_{s\le t} |w_1(s, C_1) - w_1'(s, C_1)|^2\big]^{1/2}.
\end{aligned}$$

3. Existence and uniqueness of the solution of the MF ODEs

We study the well-posedness of the solution of the MF ODEs introduced in Section 2.2. For this purpose specifically, we consider the following sub-Gaussian norm for w_i, $i \ge 2$:

$$[\![w_i]\!]_{\psi,t} = \sqrt{50} \sup_{m\ge 1} \frac{1}{\sqrt{m}} \mathbb{E}\big[\sup_{s\le t} |w_i(s, C_{i-1}, C_i)|^m\big]^{1/m}, \quad i = 2, \dots, L,$$

and accordingly we define

$$[\![W]\!]_{\psi,t} = \max\big(\max_{2\le i\le L} [\![w_i]\!]_{\psi,t}, \max_{2\le i\le L} |\!|\!| b_i |\!|\!|_t, |\!|\!| w_1 |\!|\!|_t\big).$$

The factor $\sqrt{50}$ is taken for convenience to guarantee that $[\![w_i]\!]_{\psi,t} \ge |\!|\!| w_i |\!|\!|_t$ and hence $[\![W]\!]_{\psi,t} \ge |\!|\!| W |\!|\!|_t$.

Denote by $\mathfrak{W}_T$ the space of MF parameters W such that $\|W\|_T < \infty$. Given a terminal time $T \ge 0$ and an initialization $W(0)$, we define the mapping F that associates $W' \in \mathfrak{W}_T$ with

$$\begin{aligned}
F(W')(t, c_1, \dots, c_L) = \big\{&F_1^w(W')(t, c_1), F_2^w(W')(t, c_1, c_2), F_2^b(W')(t, c_2),\\
&\dots, F_L^w(W')(t, c_{L-1}, c_L), F_L^b(W')(t, c_L)\big\},
\end{aligned}$$

in which

$$\begin{aligned}
F_1^w(W')(t, c_1) &= w_1(0, c_1) - \int_0^t \xi_1^{\mathbf{w}}(s)\mathbb{E}_Z[\Delta_1^w(Z, c_1; W'(s))]\, ds,\\
F_i^w(W')(t, c_{i-1}, c_i) &= w_i(0, c_{i-1}, c_i) - \int_0^t \xi_i^{\mathbf{w}}(s)\mathbb{E}_Z[\Delta_i^w(Z, c_{i-1}, c_i; W'(s))]\, ds,\\
F_i^b(W')(t, c_i) &= b_i(0, c_i) - \int_0^t \xi_i^{\mathbf{b}}(s)\mathbb{E}_Z[\Delta_i^b(Z, c_i; W'(s))]\, ds,\ i = 2, \dots, L.
\end{aligned}$$

Observe that at initialization $F(W')(0,\cdot,\ldots,\cdot) = W(0)$, whereas the quantities in the above time integrals are computed with respect to W'. In the following, when referring to a solution W to the MF ODEs on $[0,T]$, we mean an element of $\mathfrak{W}_T$ satisfying $F(W) = W$. We say that W is a solution to the MF ODEs on $t \in [0,\infty)$ if its restriction to $[0,T]$ is a solution to the MF ODEs on $[0,T]$ for all $T > 0$.

Theorem 3.1. *Assume the initialization $W(0)$ of the MF ODEs satisfies $[\![W]\!]_{\psi,0} \le K$. Then under Assumptions 2.4–2.6, there exists a unique solution to the MF ODEs on $t \in [0,\infty)$.*

The rest of this section is devoted to the proof of this theorem. To prove the theorem, we first collect a useful a priori estimate.

Lemma 3.2. *Under Assumptions 2.4 and 2.6, given an initialization $W(0)$, a solution W to the MF ODEs, if exists, must satisfy that for any $t \in [0,\infty)$,*

$$\begin{aligned}|||W|||_t, \max_{1\le i\le L}\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z,C_i;W(s))|^{50}\big]^{1/50}\\ \le K^{\kappa_L}(1+t^{\kappa_L})(1+|||W|||_0^{\kappa_L}),\end{aligned}$$

where $\kappa_L = K^L$ for some constant $K > 1$ sufficiently large.

A similar result holds for $[\![\cdot]\!]_{\psi,t}$ norm. Under Assumptions 2.4 and 2.6, given an initialization $W(0)$, for any $t \in [0,\infty)$, there exists $K_0(t) \ge 1$ of the form

$$K_0(t) = K^{\kappa_L}(1+t^{\kappa_L})(1+[\![W]\!]_{\psi,0}^{\kappa_L}),$$

where $\kappa_L = K^L$ for some constant $K > 1$ sufficiently large, such that the following holds. A solution W to the MF ODEs, if exists, must satisfy that for any $t \in [0,\infty)$, $|||W|||_t \le [\![W]\!]_{\psi,t} \le K_0(t)$. Furthermore, by assuming $[\![W]\!]_{\psi,0} < \infty$, for any $B \ge 0$,

$$\mathbb{P}(\mathsf{max}_t^w(W) \ge K_0(t)B) \le 2Le^{1-K_1B^2},$$

for some universal constant $K_1 > 0$.

Recall the bounds in Lemma 3.2 are given by $K_0(t)$, which is a function of the initialization $W(0)$ and non-decreasing with t. These a priori bounds lead us to consider the following spaces, given an initialization $W(0)$ and an arbitrary terminal time $T > 0$:

- The space $\mathcal{W}_T$ of MF parameters

$$W' = \{W'(t)\}_{t\le T} = \big\{w_1'(t,\cdot), w_i'(t,\cdot,\cdot), b_i'(t,\cdot),\ i = 2,\ldots,L\big\}_{t\le T}$$

such that

$$|||W'|||_T \le K_0(T).$$

- The space $\mathcal{W}_T^0 \subset \mathcal{W}_T$ of MF parameters $W' \in \mathcal{W}_T$ such that

$$[\![W']\!]_{\psi,T} \le K_0(T),$$
$$\mathbb{P}\big(\mathsf{max}_T^w(W') \ge K_0(T)B\big) \le 2Le^{1-K_1B^2} \quad \text{for all } B \ge 0,$$

  and $W'(0) = W(0)$ (and hence every elements W' in $\mathcal{W}_T^0$ share the same initialization $W(0)$). It is easy to see that $\mathcal{W}_T^0 \subset \mathcal{W}_T$ is valid since $|||W'|||_T \le [\![W']\!]_{\psi,T}$.

We equip these spaces with the metric $(W', W'') \mapsto \|W' - W''\|_T$. By Lemma 3.2, we know that any solution W to the MF ODEs, if exists, must belong to $\mathcal{W}_T^0$.

The proof of Theorem 3.1 follows from a Picard-type iteration. It is easy to see that a solution to the MF ODEs is a fixed point of F and vice versa. Also note that by the same argument of Lemma 3.2, one can prove the following.

Lemma 3.3. *Under Assumptions 2.4 and 2.6, for any $W' \in \mathcal{W}_T^0$, $F(W') \in \mathcal{W}_T^0$.*

We have the following key result.

Lemma 3.4. *For a given $B \ge 0$, consider two collections of MF parameters $W', W'' \in \mathcal{W}_T$ such that*

$$\mathbb{P}\big(\mathsf{max}_T^w(W') \ge K_0(T)B\big) \le 2Le^{1-K_1B^2},$$
$$\mathbb{P}\big(\mathsf{max}_T^w(W'') \ge K_0(T)B\big) \le 2Le^{1-K_1B^2}.$$

Under Assumptions 2.4–2.6, for any $t \le T$,

$$\|F(W') - F(W'')\|_t \le (KK_0(T))^{2L+2} \int_0^t \big((1+B)\|W' - W''\|_s + \sqrt{L}e^{-K_1B^2/2}\big)\, ds.$$

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. We perform a Picard-type iteration argument. For an arbitrary finite $T \ge 0$ and $W', W'' \in \mathcal{W}_T^0$, from Lemma 3.4, we have

$$\begin{aligned}
&\|F(W') - F(W'')\|_t \\
&\quad \le (KK_0(T))^{2L+2}\Big((1+B)\int_0^t \|W' - W''\|_s\, ds + T\sqrt{L}e^{-K_1B^2/2}\Big) \\
&\quad \equiv k_1(1+B)\int_0^t \|W' - W''\|_s\, ds + k_2 e^{-k_3B^2},
\end{aligned}$$

for any $B > 0$. By Lemma 3.3, F maps $\mathcal{W}_T^0$ to $\mathcal{W}_T^0$. As such, we can iterate this inequality to obtain

$$\begin{aligned}
&\|F^{(m)}(W') - F^{(m)}(W'')\|_T \\
&\quad \le k_1(1+B)\int_0^T \|F^{(m-1)}(W') - F^{(m-1)}(W'')\|_{T_2}\, dT_2 + k_2 e^{-k_3B^2}
\end{aligned}$$

$$
\begin{aligned}
&\le k_1^2(1+B)^2 \int_0^T \int_0^{T_2} \|F^{(m-2)}(W') - F^{(m-2)}(W'')\|_{T_3} \mathbb{I}(T_2 \le T)\, dT_3\, dT_2 \\
&\qquad + k_2 \sum_{\ell=1}^{2} \frac{(Tk_1k_2(1+B))^{\ell-1}}{\ell!}\, e^{-k_3B^2} \\
&\ \vdots \\
&\le k_1^m(1+B)^m \int_0^T \int_0^{T_2} \cdots \int_0^{T_m} \|W' - W''\|_{T_{m+1}} \mathbb{I}(T_m \le \cdots \le T_2 \le T)\, dT_{m+1} \cdots dT_2 \\
&\qquad + k_2 \sum_{\ell=1}^{m} \frac{(Tk_1k_2(1+B))^{\ell-1}}{\ell!}\, e^{-k_3B^2} \\
&\le \frac{1}{m!} T^m k_1^m (1+B)^m \|W' - W''\|_T + k_2 e^{Tk_1k_2(1+B)-k_3B^2} \\
&\le \frac{1}{m!} T^m k_1^m (1+\sqrt{m})^m \|W' - W''\|_T + k_2 e^{Tk_1k_2(1+\sqrt{m})-k_3m},
\end{aligned}
$$

where we choose $B = \sqrt{m}$ in the last display. Note that since $|||W|||_0 < \infty$, $K_0(T)$ and hence k_1, k_2 are finite for finite T. By substituting $W'' = F(W')$, we obtain

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

Hence, as $m \to \infty$, $F^{(m)}(W')$ converges in $\|\cdot\|_T$ to a limit $W \in \mathfrak{W}_T$, which is a fixed point of F. By Lemma 3.2, W belongs to $\mathcal{W}_T^0$.

The uniqueness of the fixed point comes from the above estimate, since if W' and W'' are fixed points of F, then they are both in $\mathcal{W}_T^0$, and

$$
\begin{aligned}
\|W' - W''\|_T &= \|F^{(m)}(W') - F^{(m)}(W'')\|_T \\
&\le \frac{1}{m!} T^m k_1^m (1+\sqrt{m})^m \|W' - W''\|_T + k_2 e^{Tk_1k_2(1+\sqrt{m})-k_3m},
\end{aligned}
$$

and one can take m arbitrarily large. This proves that the solution exists and is unique on $t \in [0, T]$. Since T is arbitrary, we have existence and uniqueness of the solution to the MF ODEs on the time interval $[0, \infty)$. ∎

The proofs of the lemmas are in Appendix B.

4. Main result: connection between neural network and MF limit

4.1. Neuronal embedding and the coupling procedure

Neuronal embedding. To formalize a connection between the neural network and its MF limit, we consider their initializations. In practical scenarios, to set the initial parameters $\mathbf{W}(0)$ of the neural network, one typically randomizes $\mathbf{W}(0)$ according

to some distributional law ρ. We note that since the neural network is defined with respect to a set of finite integers $\mathbf{n} = \{n_1, \dots, n_L\}$ that represents its size, so is ρ. In the context of infinite-width limits of neural networks, we would like to accommodate a sequence of neural networks of diverging sizes $\mathbf{n}$ (where $n_1, \dots, n_{L-1} \to \infty$ and $n_L = 1$). As such, it is useful to also consider a family Init of initialization laws, each of which is indexed by the set of finite integers $\mathbf{n} = \{n_1, \dots, n_L\}$ (with $n_L = 1$):

$$\mathsf{Init} = \big\{\rho : \rho \text{ is the initialization law of a neural network of size } \mathbf{n} = \{n_1, \dots, n_L\}, \\ n_1, \dots, n_L \in \mathbb{N}_{>0},\, n_L = 1\big\}.$$

We make the following crucial definitions.

Definition 4.1 (Unit neuronal embedding). Given an initialization law ρ of a neural network of size $\mathbf{n} = \{n_1, \dots, n_L\}$ (where $n_L = 1$), we call $(\Omega, P, \{w_i^0\}_{i\in[L]}, \{b_i^0\}_{2\le i\le L})$ a *unit neuronal embedding* for this neural network if there exists a sampling rule $\overline{P}_{\mathbf{n}} = \prod_{i=1}^{L} \overline{P}_{n_i}$ such that the following hold:

(1) $(\Omega, P) = \prod_{i=1}^{L}(\Omega_i, P_i)$ is a product space and $\Omega_L = \{1\}$. We recall that (Ω, P) is called a neuronal ensemble.

(2) $\overline{P}_{n_i}$ is a distribution over $\Omega_i^{n_i}$ whose marginals are given by P_i. Note it is not necessary that $\overline{P}_{n_i}$ is factored as a product of P_i's.

(3) The deterministic functions $w_1^0 : \Omega_1 \to \mathbb{W}_1$, $w_i^0 : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$ and $b_i^0 : \Omega_i \to \mathbb{B}_i$, $2 \le i \le L$ are such that if – with an abuse of notations – we sample $(C_i(j_i))_{i\in[L], j_i\in[n_i]} \sim \overline{P}_{\mathbf{n}}$, then

$$\mathrm{Law}\big(w_1^0(C_1(j_1)), w_i^0(C_{i-1}(j_{i-1}), C_i(j_i)), b_i^0(C_i(j_i)), \\ j_1 \in [n_1], j_i \in [n_i], i = 2, \dots, L\big) = \rho.$$

Definition 4.2 (Neuronal embedding). Given a family of initialization laws Init, we call $(\Omega, P, \{w_i^0\}_{i\in[L]}, \{b_i^0\}_{2\le i\le L})$ a *neuronal embedding* for Init if it is a unit neuronal embedding for any law ρ in Init.

On one hand, we concern chiefly with the notion of a neuronal embedding, which carries the idea of infinite-width limits. On the other hand, the unit neuronal embedding – as a standalone notion – is useful when one is to obtain a quantitative (finite-width) result, such as Theorem 4.7 and Corollary 4.9 below. Note also that if the family Init contains only one initialization law, then a unit neuronal embedding for this law is obviously a neuronal embedding for Init. We shall thus routinely refer to a unit neuronal embedding as a neuronal embedding, whenever there is no risk of confusion.

η-independence. An important aspect of the neuronal embedding is the sampling rule $\overline{P}_{\mathbf{n}}$. The product structure $\overline{P}_{\mathbf{n}} = \prod_{i=1}^{L} \overline{P}_{n_i}$ implies layer-wise independence. At

each layer $i \in [L]$, a canonical example of a sampling rule is one in which the samples are i.i.d., i.e., $\bar{P}_{n_i} = P_i \times \cdots \times P_i$ (n_i-time product). In fact, we shall require a weaker condition, given in the following.

Definition 4.3 (η-independence). We say that $(X_1, \ldots, X_n) \in \Omega_0^n$ are *η-independent* if for all 1-bounded functions f that maps from Ω_0 to a separable Hilbert space, for any $i \in [n]$, almost surely,

$$\left|\mathbb{E}[f(X_i) \mid \{X_{i'}, i' < i\}] - \mathbb{E}[f(X_i)]\right| \leq \eta.$$

Assumption 4.4 ($\bar{\eta}$-independence for neuronal embedding). Let $\bar{\eta} = (\eta_1, \ldots, \eta_{L-1})$, where $\eta_i = n_i^{-0.501}$. For the neuronal embedding in Definition 4.2 (or Definition 4.1), for each index $\mathbf{n}$ in the family Init, the sampling rule $\bar{P}_{\mathbf{n}}$ satisfies that $(C_i(j_i))_{j_i \in [n_i]} \sim \bar{P}_{n_i}$ are η_{i-1}-independent for all $i \in [L-1]$. In this case, we say the neuronal embedding satisfies $\bar{\eta}$-independence.

It is easy to see that in the canonical example where $\bar{P}_{n_i} = P_i \times \cdots \times P_i$ for all $i \in [L-1]$ and all indices $\mathbf{n}$ from Init, the above assumption is trivially satisfied; that is, any n independent random variables are n^{-c}-independent with $c = \infty$.

Remark 4.5. When Init contains more than one law, if a neuronal embedding exists, then Init must satisfy a certain consistency property. For instance, under the canonical example where $\bar{P}_{n_i} = P_i \times \cdots \times P_i$ for all $i \in [L-1]$ and all indices $\mathbf{n}$ from Init, if a neuronal embedding with this sampling rule exists, then the following must hold. Suppose that ρ indexed by $\{n_1, \ldots, n_L\}$ and ρ' indexed by $\{n'_1, \ldots, n'_L\}$ are elements of Init such that $n_1 \leq n'_1, \ldots, n_{L-1} \leq n'_{L-1}$, and suppose that

$$\mathrm{Law}\big(\mathbf{w}'_1(0, j_1), \mathbf{w}'_i(0, j_{i-1}, j_i), \mathbf{b}'_i(0, j_i) : j_i \in [n'_i], i = 1, \ldots, L\big) = \rho'.$$

Then we must have that

$$\mathrm{Law}\big(\mathbf{w}'_1(0, j_1), \mathbf{w}'_i(0, j_{i-1}, j_i), \mathbf{b}'_i(0, j_i) : j_i \in S_i, i = 1, \ldots, L\big) = \rho,$$

for any collection of L sets S_i, $i = 1, \ldots, L$, where each S_i is a subset of $[n'_i]$ with size $|S_i| = n_i$.

Coupling procedure. To proceed, we perform the following *coupling procedure*:

(1) Given a family of initialization laws Init, let $(\Omega, P, \{w_i^0\}_{i \in [L]}, \{b_i^0\}_{2 \leq i \leq L})$ be a neuronal embedding of Init.

(2) We form the MF ODEs' initialization $W(0)$ by setting $w_1(0, \cdot) = w_1^0(\cdot)$, $w_i(0, \cdot, \cdot) = w_i^0(\cdot, \cdot)$ and $b_i(0, \cdot) = b_i^0(\cdot)$ for $2 \leq i \leq L$. With this initialization, we obtain the MF limit's trajectory $W(t)$, for $t \in \mathbb{R}_{\geq 0}$, according to the neuronal ensemble (Ω, P).

(3) Given $\mathbf{n} = \{n_1, \dots, n_L\}$, we find a sampling rule $\bar{P}_{\mathbf{n}} = \prod_{i=1}^{L} \bar{P}_{n_i}$. For each $i \in [L]$, we sample $(C_i(j_1), \dots, C_i(j_{n_i})) \sim \bar{P}_{n_i}$. We then form the neural network initialization $\mathbf{W}(0)$ by setting $\mathbf{w}_1(0, j_1) = w_1^0(C_1(j_1))$, $\mathbf{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))$ and $\mathbf{b}_i(0, j_i) = b_i^0(C_i(j_i))$ for $j_1 \in [n_1]$, $j_i \in [n_i]$, $2 \le i \le L$. With this initialization, we obtain the neural network's trajectory $\mathbf{W}(k)$ for $k \in \mathbb{N}_{\ge 0}$, with the data $z(k)$ being generated independently of $C_i(j_i)$'s and hence of $\mathbf{W}(0)$.

Hence, we see that the connection is formalized on the basis of the initialization, and in particular, the neuronal ensemble (Ω, P). Note that $W(t)$ is a deterministic trajectory for $t \in \mathbb{R}_{\ge 0}$ and it is independent of $\{n_1, \dots, n_L\}$, whereas $\mathbf{W}(k)$ is random for all $k \in \mathbb{N}_{\ge 0}$ due to the randomness of $C_i(j_i)$ and the generation of the training data $z(k)$. We define a measure of closeness between $\mathbf{W}(\lfloor t/\epsilon \rfloor)$ and $W(t)$ for the whole interval $t \in [0, T]$:

$$\mathcal{D}_T(W, \mathbf{W}) = \max(\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3), \tag{4.1}$$

in which

$$\mathrm{I}_1 = \max_{2 \le i \le L} \Big(\frac{1}{n_{i-1} n_i} \sum_{j_{i-1}=1}^{n_{i-1}} \sum_{j_i=1}^{n_i} \sup_{t \le T} |\mathbf{w}_i(\lfloor t/\epsilon \rfloor, j_{i-1}, j_i) - w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))|^2 \Big)^{1/2},$$

$$\mathrm{I}_2 = \max_{2 \le i \le L} \Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \sup_{t \le T} |\mathbf{b}_i(\lfloor t/\epsilon \rfloor, j_i) - b_i(t, C_i(j_i))|^2 \Big)^{1/2},$$

$$\mathrm{I}_3 = \Big(\frac{1}{n_1} \sum_{j_1=1}^{n_1} \sup_{t \le T} |\mathbf{w}_1(\lfloor t/\epsilon \rfloor, j_1) - w_1(t, C_1(j_1))|^2 \Big)^{1/2}.$$

Note that, by definition, $\mathcal{D}_T(W, \mathbf{W})$ is a random quantity due to the randomness of $\{C_i(j_i)\}_{i \in [L]}$ and $\{\mathbf{W}(\lfloor t/\epsilon \rfloor)\}_{t \in [0,T]}$.

The idea of the coupling procedure is closely related to the "propagation of chaos" argument [36]. Here, instead of playing the role of a proof technique, the coupling serves as a vehicle to establish the connection between the neural network's trajectory and the MF trajectory on the basis of the neuronal embedding.

4.2. Main theorem

Let us consider an assumption on the initialization.

Assumption 4.6 (Initialization). The functions w_i^0 and b_i^0 of the neuronal embedding satisfy

$$\max_{2 \le i \le L} \sup_{m \ge 1} \frac{1}{\sqrt{m}} \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^m]^{1/m} \le K,$$

$$\max_{2\le i\le L}\sup_{m\ge 1}\frac{1}{\sqrt{m}}\mathbb{E}[|b_i^0(C_i)|^m]^{1/m}\le K,$$
$$\sup_{m\ge 1}\frac{1}{\sqrt{m}}\mathbb{E}[|w_1^0(C_1)|^m]^{1/m}\le K.$$

As such, following the coupling procedure, the initialization $W(0)$ of the MF ODEs satisfies $|||W|||_0 \le K < \infty$.

We are now ready to state the main theorem.

Theorem 4.7. *Given a family* Init *of initialization laws and a tuple of positive integers* $\{n_1,\dots,n_L\}$ *with* $n_L = 1$, *perform the coupling procedure as described in Section 4.1. Under Assumptions 2.4–2.6, 4.4 and 4.6, there exist constants* $c_1 \in (0, 0.5)$ *and* $c_2 \in (0, 1/52)$ *such that for any* $\delta > 0$, $L \ge 1$ *and* $T \in \epsilon\mathbb{N}_{\ge 0}$, *the following holds. There exist* $n^* = n^*(T, L, c_1, c_2) \ge 1$ *and* $\epsilon^* = \epsilon^*(T, L, c_1, c_2) \le 1$ *such that for any integer* $n_{\min} \ge n^*$ *and any* $\epsilon \in (0, \epsilon^*)$,

$$\mathbb{P}\left(\mathcal{D}_T(W,\mathbf{W}) \ge K(n_{\min}^{-c_1}+\epsilon^{c_1})\sqrt{\log\Big(\frac{1}{\delta}n_{\max}^2+e\Big)}\right) \le 2\delta + KLn_{\max}\exp(-Kn_{\min}^{c_2}).$$

Here $n_{\min} = \min_{1\le j\le L-1} n_j$ *and* $n_{\max} = \max_{1\le j\le L} n_j$.

Roughly speaking, with $n_i = \Theta(n)$ for $i \in [L-1]$ and $\epsilon \ll 1/\log(n)$, we have $\mathbf{W}(\lfloor t/\epsilon\rfloor) \approx W(t)$ for all $t \in [0, T]$ and large n. We note that the exponents c_1 and c_2 are independent of the terminal time T and the number of layers L. It is an interesting task to derive explicit constant values for c_1 and c_2, which we have not done given the complex dependency of these exponents on other hidden constants in our current analysis.

Remark 4.8. Under the stronger assumption of boundedness of the initial weight distributions at all except the first layer, in our work's previous preprint, we show that a similar result to Theorem 4.7 holds with $c_1 = 0.5$. Therein an even stronger result is achieved, in which we define $\mathcal{D}_T(W,\mathbf{W})$ via L^∞ distance, instead of L^2 distance as done in equation (4.1).

The theorem gives a connection between $\mathbf{W}(\lfloor t/\epsilon\rfloor)$, which involves finitely many neurons, and the MF limit $W(t)$, whose description is independent of the number of neurons. It lends a way to extract properties of the neural network in the many-neurons limit.

Corollary 4.9. *Consider any test function* $\psi\colon \mathbb{H}_i \to \mathbb{S}$ *which is* K*-Lipschitz and* K*-bounded, i.e.,*

$$|\psi(h)-\psi(h')| \le K|h-h'|, \quad |\psi(h)| \le K,$$

where $\mathbb{S}$ *is a separable Hilbert space. Under the same setting of Theorem 4.7, for*

any $\delta > 0$*, we have, with probability at least* $1 - 3\delta - KLn_{\max}\exp(-Kn_{\min}^{c_2})$,

$$\sup_{t\le T}\left|\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[\psi(\mathbf{H}_i(\lfloor t/\epsilon\rfloor, X, j_i))] - \mathbb{E}_Z\mathbb{E}_{C_i}[\psi(H_i(t, X, C_i))]\right| = \tilde{O}(n_{\min}^{-c_1} + \epsilon^{c_1}),$$

where $\tilde{O}$ *hides the dependency on* T, L *and* δ *as well as the logarithmic factors* $\log n_{\max}$ *and* $\log(1/\epsilon)$. *Furthermore, for any test function* $\psi\colon \mathbb{Y}\times\hat{\mathbb{Y}}\to\mathbb{S}$ *which is* K*-Lipschitz in the second variable, uniformly in the first variable,*

$$\sup_{t\le T}\left|\mathbb{E}_Z[\psi(Y, \hat{\mathbf{y}}(\lfloor t/\epsilon\rfloor, X))] - \mathbb{E}_Z[\psi(Y, \hat{y}(t, X))]\right| = \tilde{O}(n_{\min}^{-c_1} + \epsilon^{c_1}),$$

with probability at least $1 - 2\delta - KLn_{\max}\exp(-Kn_{\min}^{c_2})$.

As per Remark 4.5, we note that the statements in Theorem 4.7 and Corollary 4.9 have explicit quantitative dependence on the hidden widths n_i, and hence one may consider Init that contains only one initialization law.

We observe that while the MF trajectory $W(t)$ is defined as per the choice of the neuronal embedding $(\Omega, P, \{w_i^0\}_{i\in[L]}, \{b_i^0\}_{2\le i\le L})$, which may not be unique. On the other hand, the neural network's trajectory $\mathbf{W}(t)$ depends on the randomization of the initial parameters $\mathbf{W}(0)$ according to an initialization law from the family Init (as well as the data $z(t)$) and hence it is independent of this choice. Another corollary of Theorem 4.7 is that given the same family Init, the MF trajectory is insensitive to the choice of the neuronal embedding of Init.

Corollary 4.10. *Consider a family* Init *of initialization laws such that it contains a sequence of indices* $\{\{n_1(n), \dots, n_L(n)\} : n \in \mathbb{N}\}$*, in which* $n_{\min}(n)\to\infty$ *and* $n_{\min}^{-c}(n)\log n_{\max}(n)\to 0$ *as* $n\to\infty$ *for any* $c > 0$*, with* $n_{\min}(n) = \min_{1\le i\le L-1} n_i(n)$ *and* $n_{\max}(n) = \max_{1\le i\le L-1} n_i(n)$.

Let $W(t)$ *and* $\hat{W}(t)$ *be two MF trajectories associated with two choices of neuronal embeddings of* Init*, say* $(\Omega, P, \{w_i^0\}_{i\in[L]}, \{b_i^0\}_{2\le i\le L})$ *and* $(\hat{\Omega}, \hat{P}, \{\hat{w}_i^0\}_{i\in[L]}, \{\hat{b}_i^0\}_{2\le i\le L})$*, respectively. Suppose that both neuronal embeddings satisfy Assumptions 4.4 and 4.6. Let us also assume Assumptions 2.4–2.6.*

For any $T\in\mathbb{R}_{\ge 0}$ *and any set of positive integers* $\{n_1, \dots, n_L\}$ *with* $n_L = 1$*, if we independently sample* $U_i(j_i)\sim P_i$ *and* $\hat{U}_i(j_i)\sim\hat{P}_i$ *for* $j_i\in[n_i]$ *and* $i\in[L]$*, then* $\mathrm{Law}(\mathcal{W}(n_1, \dots, n_L, T)) = \mathrm{Law}(\hat{\mathcal{W}}(n_1, \dots, n_L, T))$*, where* $\mathcal{W}(n_1, \dots, n_L, T)$ *denotes the following collection on* $W(t)$:

$$\mathcal{W}(n_1, \dots, n_L, T) = \big\{w_1(t, U_1(j_1)), w_i(t, U_{i-1}(j_{i-1}), U_i(j_i)), b_i(t, U_i(j_i)) : j_i\in[n_i], i\in[L], t\in[0, T]\big\},$$

and $\hat{\mathcal{W}}(n_1, \dots, n_L, T)$ *denotes a similar collection on* $\hat{W}(t)$.

In the case $L=2$, by looking at the induced distribution of $(w_1(t,C_1), w_2(t,C_1,1))$ over $C_1 \sim P_1$, we immediately recover the distributional equation in [22] describing the MF limit.

Corollary 4.11. *Assume the same setting as Theorem 4.7, and let us consider $L=2$. For simplicity, let us disregard the bias of the second layer by considering $\xi_2^{\mathbf{b}}(\cdot)=0$ and $b_2(0,\cdot)=0$. Assume $\mathbb{W}_1=\mathbb{R}^{d_1}$, $\mathbb{W}_2=\mathbb{R}^{d_2}$ for some integers $d_1,d_2>0$. Let ρ_t denote the law of $(w_1(t,C_1), w_2(t,C_1,1))$ over $C_1 \sim P_1$. Then ρ_t satisfies the following distributional partial differential equation in the weak sense*:

$$\partial_t \rho_t(u_1,u_2) = \operatorname{div}[\rho_t(u_1,u_2) G(u_1,u_2;\rho_t)],$$

in which

$$G(u_1,u_2;\rho_t) = \begin{bmatrix} \xi_1^{\mathbf{w}}(t)\mathbb{E}_Z[\underline{\Delta}_1^w(u_1,u_2;Z,\rho_t)] \\ \xi_2^{\mathbf{w}}(t)\mathbb{E}_Z[\underline{\Delta}_2^w(u_1,u_2;Z,\rho_t)] \end{bmatrix},$$

and we define

$$\begin{aligned}
\underline{H}_2(x;\rho_t) &= \int \phi_2(u_2,0,\phi_1(u_1,x))\, d\rho_t(u_1,u_2), \\
\underline{\hat{y}}(x;\rho_t) &= \phi_3(H_2(x;\rho_t)), \\
\underline{\Delta}_2^H(z;\rho_t) &= \sigma_2^{\mathbf{H}}(y, \underline{\hat{y}}(x;\rho_t), \underline{H}_2(x;\rho_t)), \\
\underline{\Delta}_2^w(u_1,u_2;z,\rho_t) &= \sigma_2^{\mathbf{w}}(\underline{\Delta}_2^H(z;\rho_t), u_2, 0, \underline{H}_2(x;\rho_t), \phi_1(u_1,x)), \\
\underline{\Delta}_1^H(u_1,u_2;z,\rho_t) &= \sigma_1^{\mathbf{H}}(\underline{\Delta}_2^H(z;\rho_t), u_2, 0, \underline{H}_2(x;\rho_t), \phi_1(u_1,x)), \\
\underline{\Delta}_1^w(u_1,u_2;z,\rho_t) &= \sigma_1^{\mathbf{w}}(\underline{\Delta}_1^H(u_1,u_2;z,\rho_t), u_1, x).
\end{aligned}$$

In particular, for any $\delta>0$ and any K-Lipschitz and K-bounded test function $\psi\colon \mathbb{W}_1 \times \mathbb{W}_2 \to \mathbb{S}$, where $\mathbb{S}$ is a separable Hilbert space,

$$\begin{aligned}
&\sup_{t\le T}\left|\frac{1}{n_1}\sum_{j_1=1}^{n_1}\psi(\mathbf{w}_1(\lfloor t/\epsilon\rfloor,j_1),\mathbf{w}_2(\lfloor t/\epsilon\rfloor,j_1,1)) - \int \psi(u_1,u_2)\, d\rho_t(u_1,u_2)\right| \\
&\quad = \tilde{O}(n_1^{-c_1}+\epsilon^{c_1}).
\end{aligned}$$

with probability at least $1-3\delta-Kn_1\exp(-Kn_1^{c_2})$, where $\tilde{O}$ hides the dependency on T and δ as well as the logarithmic factors $\log n_1$ and $\log(1/\epsilon)$. Similarly, for any test function $\psi\colon \mathbb{Y}\times\hat{\mathbb{Y}}\to\mathbb{S}$ which is K-Lipschitz in the second variable, uniformly in the first variable,

$$\sup_{t\le T}\left|\mathbb{E}_Z[\psi(Y,\hat{\mathbf{y}}(\lfloor t/\epsilon\rfloor,X))] - \mathbb{E}_Z[\psi(Y,\underline{\hat{y}}(X;\rho_t))]\right| = \tilde{O}(n_1^{-c_1}+\epsilon^{c_1}),$$

with probability at least $1-3\delta-Kn_1\exp(-Kn_1^{c_2})$.

4.3. Proof of Theorem 4.7

We construct an auxiliary trajectory, which we call the *particle ODEs*:

$$\begin{aligned}
\frac{\partial}{\partial t}\tilde{w}_1(t, j_1) &= -\xi_1^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_1^{\mathbf{w}}(Z, j_1; \tilde{W}(t))] \quad \text{for all } j_1 \in [n_1],\\
\frac{\partial}{\partial t}\tilde{w}_i(t, j_{i-1}, j_i) &= -\xi_i^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_i^{\mathbf{w}}(Z, j_{i-1}, j_i; \tilde{W}(t))],\\
\frac{\partial}{\partial t}\tilde{b}_i(t, j_i) &= -\xi_i^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_i^{\mathbf{b}}(Z, j_i; \tilde{W}(t))] \quad \text{for all } j_{i-1} \in [n_{i-1}],\, j_i \in [n_i],
\end{aligned}$$

for $i = 2, \dots, L$, in which $\tilde{W}(t) = \{\tilde{w}_1(t, \cdot), \tilde{w}_i(t, \cdot, \cdot), \tilde{b}_i(t, \cdot),\ i = 2, \dots, L\}$, and $t \in \mathbb{R}_{\geq 0}$. We specify the initialization $\tilde{W}(0)$ as follows: $\tilde{w}_1(0, j_1) = w_1^0(C_1(j_1))$, $\tilde{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))$ and $\tilde{b}_i(0, j_i) = b_i^0(C_i(j_i))$. That is, it shares the same initialization with the neural network one $\mathbf{W}(0)$, and hence it is coupled with the neural network and the MF ODEs. Roughly speaking, the particle ODEs are continuous-time trajectories of finitely many neurons, averaged over the data distribution. We note that $\tilde{W}(t)$ is random for all $t \in \mathbb{R}_{\geq 0}$ due to the randomness of $C_i(j_i)$'s.

The existence and uniqueness of the solution to the particle ODEs follows from the same proof as in Theorem 3.1, which we shall not repeat here.[2] We equip $\tilde{W}(t)$ with the norms

$$\begin{aligned}
|||\tilde{w}_i|||_t &= \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\leq t}|\tilde{w}_i(s, j_{i-1}, j_i)|^{50}\Big)^{1/50},\\
|||\tilde{b}_i|||_t &= \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\leq t}|\tilde{b}_i(s, j_i)|^{50}\Big)^{1/50}, \quad i = 2, \dots, L,\\
|||\tilde{w}_1|||_t &= \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{s\leq t}|\tilde{w}_1(s, j_1)|^{50}\Big)^{1/50}.
\end{aligned}$$

as well as

$$|||\tilde{W}|||_t = \max\Big(\max_{1\leq i\leq L}|||\tilde{w}_i|||_t, \max_{2\leq i\leq L}|||\tilde{b}_i|||_t\Big).$$

One can also define the measures $\mathcal{D}_T(W, \tilde{W})$ and $\mathcal{D}_T(\tilde{W}, \mathbf{W})$ similar to equation (4.1):

$$\begin{aligned}
\mathcal{D}_T(W, \tilde{W}) &= \max(\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3),\\
\mathcal{D}_T(\tilde{W}, \mathbf{W}) &= \max(\mathrm{I}_4, \mathrm{I}_5, \mathrm{I}_6),
\end{aligned}$$

[2] On a more technical note, we can view the particle ODEs as a new system of MF ODEs whose neuronal ensemble $(\Omega_{\text{new}}, P_{\text{new}}) = \prod_{i=1}^{L}(\Omega_{i,\text{new}}, P_{i,\text{new}})$ takes the following specific form: $\Omega_{i,\text{new}} = \{C_i(1), \dots, C_i(n_i)\}$ and $P_{i,\text{new}}$ is a uniform probability measure on $\Omega_{i,\text{new}}$. In light of this view, the existence and uniqueness of the solution to the particle ODEs follows from Theorem 3.1.

in which

$$\mathrm{I}_1 = \max_{2\le i\le L}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))-\tilde{w}_i(t,j_{i-1},j_i)|^2\Big)^{1/2},$$

$$\mathrm{I}_2 = \max_{2\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{t\le T}|b_i(t,C_i(j_i))-\tilde{b}_i(t,j_i)|^2\Big)^{1/2},$$

$$\mathrm{I}_3 = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{t\le T}|w_1(t,C_1(j_1))-\tilde{w}_1(t,j_1)|^2\Big)^{1/2},$$

$$\mathrm{I}_4 = \max_{2\le i\le L}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)-\tilde{w}_i(t,j_{i-1},j_i)|^2\Big)^{1/2},$$

$$\mathrm{I}_5 = \max_{2\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{t\le T}|\mathbf{b}_i(\lfloor t/\epsilon\rfloor,j_i)-\tilde{b}_i(t,j_i)|^2\Big)^{1/2},$$

$$\mathrm{I}_6 = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{t\le T}|\mathbf{w}_1(\lfloor t/\epsilon\rfloor,j_1)-\tilde{w}_1(t,j_1)|^2\Big)^{1/2}.$$

We have the following results.

Theorem 4.12. *Under the same setting as Theorem 4.7, there exist constants $c_1 \in (0, 0.5)$ and $c_2 \in (0, 1/52)$ such that for any $\delta > 0$, $L \ge 1$ and $T \ge 1$, the following holds. There exists $n^* = n^*(T, L, c_1, c_2) \ge 1$ such that for any $n_{\min} \ge n^*$,*

$$\mathbb{P}\Big(\mathcal{D}_T(W,\tilde{W}) \ge K n_{\min}^{-c_1}\sqrt{\log\Big(\frac{1}{\delta}n_{\max}^2+e\Big)}\Big) \le \delta + KLn_{\max}\exp(-Kn_{\min}^{c_2}).$$

Here $n_{\min} = \min_{1\le j\le L-1} n_j$ and $n_{\max} = \max_{1\le j\le L} n_j$.

Theorem 4.13. *Under the same setting as Theorem 4.7, there exist constants $c_1 \in (0, 0.5)$ and $c_2 \in (0, 1/52)$ such that for any $\delta > 0$, $L \ge 1$ and $T \ge 1$, the following holds. There exists $\epsilon^* = \epsilon^*(T, L, c_1, c_2) \le 1$ such that for any $\epsilon \in (0, \epsilon^*)$,*

$$\mathbb{P}\Big(\mathcal{D}_T(\tilde{W},\mathbf{W}) \ge K\epsilon^{c_1}\sqrt{\log\Big(\frac{1}{\delta}n_{\max}^2+e\Big)}\Big) \le \delta + KLn_{\max}\exp(-Kn_{\min}^{c_2}).$$

Here $n_{\min} = \min_{1\le j\le L-1} n_j$ and $n_{\max} = \max_{1\le j\le L} n_j$.

Proof of Theorem 4.7. Using the fact

$$\mathcal{D}_T(W,\mathbf{W}) \le \mathcal{D}_T(W,\tilde{W}) + \mathcal{D}_T(\tilde{W},\mathbf{W}),$$

the thesis is immediate from Theorems 4.12 and 4.13. ∎

4.4. Proof of Theorems 4.12 and 4.13

The proof of Theorem 4.12 rests in the following proposition, which is essentially a version of Theorem 4.12 with an extra boundedness condition at initialization.

Proposition 4.14. *Under the same setting as Theorem 4.7, for a given $B > 0$, further assume that*

$$\text{ess-sup}\, \max_0^w(W) = \text{ess-sup} \max_{2\le i\le L} |w_i^0(C_{i-1}, C_i)| \le B,$$
$$\text{ess-sup}\, \max_0^b(W) = \text{ess-sup} \max_{2\le i\le L} |b_i^0(C_i)| \le B.$$

Then for any $\delta > 0$, with probability at least $1 - \delta - KLn_{\max}\exp(-Kn_{\min}^{1/52})$,

$$\mathcal{D}_T(W, \tilde{W}) \le \sqrt{\frac{1}{n_{\min}} \log\Big(\frac{2TL}{\delta} n_{\max}^2 + e\Big)}\exp(K^{\bar{K}}(1 + T^{\bar{K}})(1 + B)),$$

in which $n_{\min} = \min_{1\le j\le L-1} n_j$, $n_{\max} = \max_{1\le j\le L} n_j$, and $\bar{K}$ is a constant that depends on L such that $\bar{K} \le K^L$ for some sufficiently large constant K.

Similar to Proposition 4.14, the following proposition is essentially a version of Theorem 4.13 with an extra boundedness condition at initialization.

Proposition 4.15. *Under the same setting as Theorem 4.7, for a given $B > 0$, further assume that*

$$\text{ess-sup}\, \max_0^w(W) = \text{ess-sup} \max_{2\le i\le L} |w_i^0(C_{i-1}, C_i)| \le B,$$
$$\text{ess-sup}\, \max_0^b(W) = \text{ess-sup} \max_{2\le i\le L} |b_i^0(C_i)| \le B.$$

Then for any $\delta > 0$ and $\epsilon < 1$, with probability at least $1 - \delta - KLn_{\max}\exp(-Kn_{\min}^{1/52})$,

$$\mathcal{D}_T(\tilde{W}, \mathbf{W}) \le \sqrt{\epsilon \log\Big(\frac{2L}{\delta} n_{\max}^2 + e\Big)}\exp(K^{\bar{K}}(1 + T^{\bar{K}})(1 + B)),$$

in which $n_{\min} = \min_{1\le j\le L-1} n_j$, $n_{\max} = \max_{1\le j\le L} n_j$, and $\bar{K}$ is a constant that depends on L such that $\bar{K} \le K^L$ for some sufficiently large constant K.

The following proposition bridges the last two propositions with their respective theorems.

Proposition 4.16. *Assume the same setting as Theorem 4.7. Let $\underline{W}(t) = \{\underline{w}_1(t,\cdot), \underline{w}_i(t,\cdot,\cdot), \underline{b}_i(t,\cdot), i = 2,\dots,L\}$ be the MF ODEs' solution for which its initialization $\underline{W}(0)$ is a truncated version of $W(0)$, for a given $B > 0$:*

$$\underline{w}_1(0, c_1) = w_1^0(c_1), \quad \underline{w}_i(0, c_{i-1}, c_i) = \text{Trunc}_B(w_i^0(c_{i-1}, c_i)),$$
$$\underline{b}_i(0, c_i) = \text{Trunc}_B(b_i^0(c_i)),$$

for $2 \le i \le L$, *where* $\mathrm{Trunc}_B(u) = u\mathbb{I}(|u| \le B) + B\mathrm{sign}(u)\mathbb{I}(|u| > B)$. *Then*

$$\|W - \underline{W}\|_T \le K \exp(-KB^2 + K^{\bar{K}}(1 + T^{\bar{K}})(1 + B)),$$

for $\bar{K}$ *a constant that depends on* L *such that* $\bar{K} \le K^L$ *for some sufficiently large constant* K. *Similarly, let* $\underline{\tilde{W}}$ *and* $\underline{\mathbf{W}}$ *be the particle ODEs' solution and the neural network's dynamics with a similarly truncated initialization*:

$$\begin{aligned}
\underline{\tilde{w}}_1(0, j_1) = \underline{\mathbf{w}}_1(0, j_1) &= w_1^0(C_1(j_1)), \\
\underline{\tilde{w}}_i(0, j_{i-1}, j_i) = \underline{\mathbf{w}}_i(0, j_{i-1}, j_i) &= \mathrm{Trunc}_B(w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))), \\
\underline{\tilde{b}}_i(0, j_i) = \underline{\mathbf{b}}_i(0, j_i) &= \mathrm{Trunc}_B(b_i^0(C_i(j_i))).
\end{aligned}$$

Then, with probability at least $1 - KLn_{\max}\exp(-Ke^{-KB^2}n_{\min}^{1/52})$,

$$\|\tilde{W} - \underline{\tilde{W}}\|_T,\ \|\mathbf{W} - \underline{\mathbf{W}}\|_T \le K \exp(-KB^2 + K^{\bar{K}}(1 + T^{\bar{K}})(1 + B)).$$

Here $n_{\max} = \max(n_1, \ldots, n_L)$, $n_{\min} = \min(n_1, \ldots, n_{L-1})$,

$$\|\mathbf{W} - \underline{\mathbf{W}}\|_t = \max\Big(\max_{2\le i\le L}\|\mathbf{w}_i - \underline{\mathbf{w}}_i\|_t, \max_{2\le i\le L}\|\mathbf{b}_i - \underline{\mathbf{b}}_i\|_t, \|\mathbf{w}_1 - \underline{\mathbf{w}}_1\|_t\Big),$$

$$\|\mathbf{w}_i - \underline{\mathbf{w}}_i\|_t = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\le t}|\mathbf{w}_i(\lfloor s/\epsilon\rfloor, j_{i-1}, j_i) - \underline{\mathbf{w}}_i(\lfloor s/\epsilon\rfloor, j_{i-1}, j_i)|^2\Big)^{1/2},$$

$$\|\mathbf{b}_i - \underline{\mathbf{b}}_i\|_t = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}|\mathbf{b}_i(\lfloor s/\epsilon\rfloor, j_i) - \underline{\mathbf{b}}_i(\lfloor s/\epsilon\rfloor, j_i)|^2\Big)^{1/2}, \quad i = 2, \ldots, L,$$

$$\|\mathbf{w}_1 - \underline{\mathbf{w}}_1\|_t = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{s\le t}|\mathbf{w}_1(\lfloor s/\epsilon\rfloor, j_1) - \underline{\mathbf{w}}_1(\lfloor s/\epsilon\rfloor, j_1)|^2\Big)^{1/2},$$

and $\|\tilde{W} - \underline{\tilde{W}}\|_t$ *is defined similarly.*

We can now prove Theorems 4.12 and 4.13.

Proof of Theorem 4.12. Let $K_T = K^{\bar{K}}(1 + T^{\bar{K}})$. For a given $B > 0$, let $\underline{W}$ and $\underline{\tilde{W}}$ be the initialization-truncated versions of W and $\tilde{W}$, respectively, as per Proposition 4.16. Then Proposition 4.14 states that for any $\delta > 0$, with probability at least $1 - \delta - KLn_{\max}\exp(-Kn_{\min}^{1/52})$,

$$\mathcal{D}_T(\underline{W}, \underline{\tilde{W}}) \le \sqrt{\frac{1}{n_{\min}}\log\Big(\frac{2TL}{\delta}n_{\max}^2 + e\Big)}e^{K_T(1+B)}.$$

In addition, by Proposition 4.16, the following holds:

$$\|\tilde{W} - \underline{\tilde{W}}\|_T,\ \|W - \underline{W}\|_T \le Ke^{-KB^2 + K_T(1+B)}$$

with probability at least $1 - KLn_{\max}\exp(-Ke^{-KB^2}n_{\min}^{1/52})$. Also notice that

$$\mathcal{D}_T(W,\tilde{W}) \le \mathcal{D}_T(\underline{W},\underline{\tilde{W}}) + \|W-\underline{W}\|_T + \|\tilde{W}-\underline{\tilde{W}}\|_T.$$

As such,

$$\mathcal{D}_T(W,\tilde{W}) \le \left(\sqrt{\frac{1}{n_{\min}}\log\Big(\frac{2TL}{\delta}n_{\max}^2+e\Big)} + e^{-KB^2}\right)e^{K_T(1+B)},$$

with probability at least $1-\delta-KLn_{\max}\exp(-Ke^{-KB^2}n_{\min}^{1/52})$, for any fixed $B>0$. Then upon choosing $B = c_0\sqrt{\log n_{\min}}$ for some suitable constant $c_0>0$ independent of T, it is easy to see that there exist constants $c_1 \in (0,0.5)$ and $c_2 \in (0,1/52)$ independent of T and some $n^* = n^*(T,L,c_1,c_2) \ge 1$ such that for any $n_{\min} \ge n^*$, we have

$$\mathbb{P}\left(\mathcal{D}_T(W,\tilde{W}) \ge Kn_{\min}^{-c_1}\sqrt{\log\Big(\frac{1}{\delta}n_{\max}^2+e\Big)}\right) \le \delta + KLn_{\max}\exp(-Kn_{\min}^{c_2}). \quad \blacksquare$$

Proof of Theorem 4.13. This comes from Propositions 4.15 and 4.16, similar to the proof of Theorem 4.12. ∎

Let us mention again the correspondence between Theorem 4.12 and Proposition 4.14, and that between Theorem 4.13 and Proposition 4.15. The truncation at initialization allows for technical feasibility and it is then bridged by Proposition 4.16. The proofs of Propositions 4.14 and 4.15 are necessarily lengthy, so let us defer them (as well as missing proofs of other results) to Appendix C. Let us describe briefly the argument for Proposition 4.14. Recall that $w_i(0, C_{i-1}(j_{i-1}), C_i(j_i)) = \tilde{w}_i(0, j_{i-1}, j_i)$ at initialization $t=0$, and hence one hopes to prove

$$w_i(t, C_{i-1}(j_{i-1}), C_i(j_i)) \approx \tilde{w}_i(t, j_{i-1}, j_i)$$

at any finite t. In other words, we would like to show

$$\mathbb{E}_Z[\Delta_i^w(Z, C_{i-1}(j_{i-1}), C_i(j_i); W(t))] \approx \mathbb{E}_Z[\Delta_i^{\mathbf{w}}(Z, j_{i-1}, j_i; \tilde{W}(t))].$$

Both of these quantities share very similar structures. Roughly speaking, the left-hand side involves quantities that assume the form of an expectation $\mathbb{E}_{C_r}[g(C_r)]$ and the right-hand side correspondingly involves quantities of the form of an empirical average $(1/n_r)\cdot\sum_{j_r=1}^{n_r} g(C_r(j_r))$, for some function g. An invocation of concentration of measure bounds links the two sides, and if done correctly over the training horizon (i.e., over $t \le T$), the depth of the network (i.e., over index $i \le L$) and the width at each layer (i.e., over neuron $j_i \le n_i$), it gives the desired estimation. One also recognizes that the neural network $\mathbf{W}$ is essentially a time discretization version of $\tilde{W}$ where the learning rate ϵ plays the role of the discretization level. A martingale-type argument then suffices to prove Proposition 4.15 for small ϵ.

5. Simplifications under independent and identically distributed initialization

In this section, we prove that the MF limit under an independent and identically distributed (i.i.d.) initialization degenerates to a simple structured dynamics. Let us first state the definition of i.i.d. initializations.

Definition 5.1. An initialization law ρ for a neural network of size $\{n_1, \dots, n_L\}$ is called $(\rho_{\mathbf{w}}^1, \dots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \dots, \rho_{\mathbf{b}}^L)$-i.i.d. initialization (or i.i.d. initialization, for brevity), where $\rho_{\mathbf{w}}^i$ is a probability measure over $\mathbb{W}_i$ and $\rho_{\mathbf{b}}^i$ is a probability measure over $\mathbb{B}_i$, if it satisfies the following:

- $\{\mathbf{w}_1(0, j_1)\}_{j_1 \in [n_1]}$ are generated i.i.d. according to $\rho_{\mathbf{w}}^1$,
- for each $i = 2, \dots, L$, $\{\mathbf{w}_i(0, j_{i-1}, j_i)\}_{j_{i-1} \in [n_{i-1}], j_i \in [n_i]}$ are generated i.i.d. according to $\rho_{\mathbf{w}}^i$, and $\{\mathbf{b}_i(0, j_i)\}_{j_i \in [n_i]}$ are generated i.i.d. according to $\rho_{\mathbf{b}}^i$,
- all these generations are independent of each other, and $\rho_{\mathbf{b}}^L$ is a single point mass.

Observe that, given $(\rho_{\mathbf{w}}^1, \dots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \dots, \rho_{\mathbf{b}}^L)$, one can build a family Init of i.i.d. initialization laws that contains any index tuple $\{n_1, \dots, n_L\}$.

In the following, we construct a canonical MF limit under i.i.d. initialization and show that the MF dynamics can be significantly simplified. Our plan is as follows:

(1) We first construct a sequence (in increasing M) of neuronal embeddings, which we call *canonical neuronal embeddings*. In particular, each of these – indexed by M – allows to embed i.i.d.-initialized neural networks of sizes at most M. Each canonical neuronal embedding is associated with a MF limit, which we call *a canonical MF limit*.

(2) We present a dynamics which is shown to be the infinite-M limit of the canonical MF limits. This dynamics displays the simplifying properties that we wish to show. In particular, the dynamics of i.i.d.-initialized neural networks of large widths are well-approximated by the infinite-M limit, and asymptotically displays the same simplifying properties.

This plan streamlines our studies of i.i.d.-initialized networks in the infinite-width limit. As we shall see, the construction of the canonical neuronal embedding is quite natural due to the cap at finite M. More importantly, on one hand, the fact that the canonical MF limit tracks closely the neural network of size less than M demonstrates flexibility of Theorem 4.7 from Section 4, in that its applicability is not limited to abstract infinite-width limits. On the other hand, the fact that the simplifying properties are shown in the infinite-M limit demonstrates the advantage of working with these abstract infinite-width dynamics: they reveal properties that are virtually invisible at the finite-width level.

5.1. Neuronal embedding construction and main results

5.1.1. Canonical neuronal embeddings and canonical MF limits. We describe the construction in three steps with a given positive integer M and a set of measures $(\rho_{\mathbf{w}}^1, \dots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \dots, \rho_{\mathbf{b}}^L)$.

Step 1. We first give a description of a σ-finite measure space. Consider a probability space (Λ, P_0) of the random processes $\mathbb{W}_1$-valued $\mathfrak{p}_1(\theta_1)$, $\mathbb{W}_i$-valued $\mathfrak{q}_i(\theta_{i-1}, \theta_i)$ and $\mathbb{B}_i$-valued $\mathfrak{p}_i(\theta_i)$ for $2 \le i \le L$. These processes are indexed by $\theta_i \in \mathbb{N}_{>0}$ and satisfy the following property. Let $m_1, \dots, m_{L-1}$ be $L-1$ arbitrary finite positive integers and, with these integers, let $\{\theta_i^{(k_i)} \in \mathbb{N}_{>0} : k_i \in [m_i],\ i = 1, \dots, L-1\}$ be an arbitrary collection. Let $m_L = 1$ and $\theta_L^{(1)} = 1$. For each $i = 1, \dots, L$, let S_i be the set of unique elements in $\{\theta_i^{(k_i)} : k_i \in [m_i]\}$. Similarly, for each $i = 2, \dots, L$, let R_i be the set of unique pairs in $\{(\theta_{i-1}^{(k_{i-1})}, \theta_i^{(k_i)}) : k_{i-1} \in [m_{i-1}], k_i \in [m_i]\}$. The space (Λ, P_0) satisfies that $\{\mathfrak{p}_i(\theta_i) : \theta_i \in S_i, i = 1, \dots, L\}$ and $\{\mathfrak{q}_i(\theta_{i-1}, \theta_i) : (\theta_{i-1}, \theta_i) \in R_i, i = 2, \dots, L\}$ are all mutually independent. In addition, we also have

$$\mathrm{Law}(\mathfrak{p}_1(\theta_1)) = \rho_{\mathbf{w}}^1, \quad \mathrm{Law}(\mathfrak{p}_i(\theta_i)) = \rho_{\mathbf{b}}^i, \quad \mathrm{Law}(\mathfrak{q}_i(\theta'_{i-1}, \theta'_i)) = \rho_{\mathbf{w}}^i$$

for any $\theta_1 \in S_1$, $\theta_i \in S_i$ and $(\theta'_{i-1}, \theta'_i) \in R_i$, for $i = 2, \dots, L$. Such a space (Λ, P_0) exists by Kolmogorov's extension theorem.

Step 2. With this space, given the integer M, for each $i \in [L-1]$, we define $\Omega_i^M = \Lambda \times [M]$ equipped with the product measure $P_i^M = P_0 \times \mathrm{Unif}([M])$, where $\mathrm{Unif}([M])$ is the uniform measure over the finite set $[M]$. We also let $\Omega_L^M = \{1\}$ and $P_L^M = \mathbb{I}_{\Omega_L^M}$. We construct $\Omega^M = \prod_{i=1}^L \Omega_i^M$, equipped with the product measure $P^M = \prod_{i=1}^L P_i^M$. The space (Ω^M, P^M) gives a *canonical neuronal ensemble*.

Step 3. Let $\Omega_i = \Lambda \times \mathbb{N}_{>0}$ and observe $\Omega_i^M \subset \Omega_i$ for any M. We define the deterministic functions $w_1^0 \colon \Omega_1 \to \mathbb{W}_1$, $w_i^0 \colon \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$ and $b_i^0 \colon \Omega_i \to \mathbb{B}_i$, for $i = 2, \dots, L$:

$$w_1^0((\lambda_1, \theta_1)) = \mathfrak{p}_1(\theta_1)(\lambda_1), \tag{5.1}$$

$$w_i^0((\lambda_{i-1}, \theta_{i-1}), (\lambda_i, \theta_i)) = \mathfrak{q}_i(\theta_{i-1}, \theta_i)(\lambda_i), \quad i = 2, \dots, L-1, \tag{5.2}$$

$$w_L^0((\lambda_{L-1}, \theta_{L-1}), 1) = \mathfrak{q}_L(\theta_{L-1}, 1)(\lambda_{L-1}), \tag{5.3}$$

$$b_i^0((\lambda_i, \theta_i)) = \mathfrak{p}_i(\theta_i)(\lambda_i), \quad i = 2, \dots, L-1, \tag{5.4}$$

$$b_L^0(1) = \mathfrak{p}_L(1). \tag{5.5}$$

These functions, together with (Ω^M, P^M) , give a *canonical neuronal embedding*. Per Section 2.2, given this neuronal embedding, one obtains a *canonical MF limit* $W^M(t) = \{w_1^M(t, \cdot), w_i^M(t, \cdot, \cdot), b_i^M(t, \cdot),\ i = 2, \dots, L\}$, defined on (Ω^M, P^M), with initialization $W^M(0) = \{w_1^0, w_i^0, b_i^0 : i = 2, \dots, L\}$. With $(C_1, \dots, C_L) \sim P^M$,

one observes that

$$\mathrm{Law}\big(w_1^0(C_1), w_2^0(C_1, C_2), b_2^0(C_2), \ldots, w_L^0(C_{L-1}, 1), b_L^0(1)\big) = \rho_{\mathbf{w}}^1 \times \prod_{i=2}^{L} \rho_{\mathbf{w}}^i \times \rho_{\mathbf{b}}^i.$$

We also consider the sampling rule $\bar{P}_{\mathbf{n}}^M$, defined, for each $\mathbf{n} = (n_1, \ldots, n_L)$ with $n_i \le M$ for $i \in [L-1]$ and $n_L = 1$, by independently sampling $\{C_i(j_i)\}_{j_i \in [n_i]}$ from $(P_i^M)^{n_i}$ conditioned on that $\{\theta_i(j_i)\}_{j_i \in [n_i]}$ are all distinct, for each $i \in [L]$, where $C_i(j_i) = (\lambda_i(j_i), \theta_i(j_i))$.

The constructed embedding indeed gives a valid neuronal embedding for neural networks of sizes at most M.

Proposition 5.2. *For $\mathbf{n} = \{n_1, \ldots, n_L\}$ with $n_i \le M$ and $n_L = 1$, the space (Ω^M, P^M) together with the functions $(\{w_i^0\}_{i \in [L]}, \{b_i^0\}_{2 \le i \le L})$ form a neuronal embedding for the neural network of size $\mathbf{n}$ under $(\rho_{\mathbf{w}}^1, \ldots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \ldots, \rho_{\mathbf{b}}^L)$-i.i.d. initialization, in which the associated sampling rule is $\bar{P}_{\mathbf{n}}^M$. Furthermore, $\bar{P}_{\mathbf{n}}^M$ is $(2n_{\max}/M)$-independent, where $n_{\max} = \max(n_1, \ldots, n_L)$.*

The proof of the proposition is deferred to Appendix D. This result, together with Theorem 4.7, suggests that for large M, the canonical MF limit tracks closely the trajectory of an i.i.d.-initialized neural network, as long as its (large) size is much smaller than M. Equivalently an i.i.d.-initialized large neural network can be closely tracked by any canonical MF limit with sufficiently large M. This motivates the studies of the canonical MF limits in the limit $M \to \infty$, which display simplified structures.

5.1.2. Infinite-M limit of canonical MF limits. Recall that the space (Ω^M, P^M) depends on M and only gives an embedding of networks whose widths are at most M. More specifically, while the space (Λ, P_0) is independent of M and Ω^M can be extended to infinite M, the measure P^M would become an improper probability measure for infinite M. Nevertheless, one can still define a dynamics that is independent of M.

Let $W^*(t) = \{w_1^*(t, \cdot), w_i^*(t, \cdot), b_i^*(t, \cdot), i = 2, \ldots, L\}$ be a dynamics to be described shortly, which we shall prove to be the "infinite-M" limit of W^M. The full description is lengthy and is deferred to Appendix D.1; let us give a snapshot description for $L \ge 5$ and $i = 3, \ldots, L-2$:

$$\frac{\partial}{\partial t} w_i^*(t, u_i, v_{i-1}, v_i) = -\xi_i^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_i^{w*}(t, Z, u_i, v_{i-1}, v_i)],$$
$$\frac{\partial}{\partial t} b_i^*(t, v_i) = -\xi_i^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_i^{b*}(t, Z)],$$

for all $u_i \in \mathrm{supp}(\rho_{\mathbf{w}}^i)$, $v_i \in \mathrm{supp}(\rho_{\mathbf{b}}^i)$, with the initialization $w_i^*(0, u_i, \cdot, \cdot) = u_i$ and

$b_i^*(0, v_i) = v_i$. Here the quantities are defined by

$$H_i^*(t,x,v_i) = \int \phi_i\big(w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_{i-1}^*(t,x,v_{i-1})\big) \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1}),$$

$$\Delta_i^{w*}(t,z,u_i,v_{i-1},v_i) = \sigma_i^{\mathbf{w}}\big(\Delta_i^{H*}(t,z,v_i), w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_i^*(t,x,v_i), H_{i-1}^*(t,x,v_{i-1})\big),$$

$$\Delta_i^{b*}(t,z,v_i) = \int \sigma_i^{\mathbf{b}}\big(\Delta_i^{H*}(t,z,v_i), w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_i^*(t,x,v_i), H_{i-1}^*(t,x,v_{i-1})\big) \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1}),$$

$$\Delta_{i-1}^{H*}(t,z,v_{i-1}) = \int \sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{H*}(t,z,v_i), w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_i^*(t,x,v_i), H_{i-1}^*(t,x,v_{i-1})\big) \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^i(dv_i).$$

The existence and uniqueness of such dynamics follow similarly to the proof of Theorem 3.1. We state the main result of this section, which shows that the dynamics W^* is the infinite-M limit of W^M. (Again we refer to Appendix D.1, specifically Theorem D.1, for the complete statement of this theorem.)

Theorem 5.3 (Snapshot statement). *Given $(\rho_{\mathbf{w}}^1, \dots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \dots, \rho_{\mathbf{b}}^L)$ and an integer M, construct the canonical neuronal ensemble (Ω^M, P^M), the random variables $(C_1, \dots, C_L) \sim P^M = \prod_{i=1}^L P_i^M$ and the canonical MF limit W^M as described in Section 5.1.1. Also construct the dynamics W^* described in Section 5.1.2. Define the following*:

$$w_i^\infty(t, c_{i-1}, c_i) = w_i^*(t, w_i^0(c_{i-1}, c_i), b_{i-1}^0(c_{i-1}), b_i^0(c_i)),$$
$$b_i^\infty(t, c_i) = b_i^*(t, b_i^0(c_i)) \quad \text{for all } c_i \in \Omega_i = \Lambda \times \mathbb{N}_{>0}, i = 3, \dots, L-2.$$

We also let $W^\infty(t) = \{w_1^\infty(t,\cdot), w_i^\infty(t,\cdot,\cdot), b_i^\infty(t,\cdot),\ i = 2, \dots, L\}$. Let us consider

$$\langle W^M - W^\infty\rangle_t = \max\Big(\max_{1\le i\le L}\langle w_i^M - w_i^\infty\rangle_t, \max_{2\le i\le L}\langle b_i^M - b_i^\infty\rangle_t\Big),$$
$$\langle w_i^M - w_i^\infty\rangle_t = \mathbb{E}\big[|w_i^M(t, C_{i-1}, C_i) - w_i^\infty(t, C_{i-1}, C_i)|^2\big]^{1/2},$$
$$\langle b_i^M - b_i^\infty\rangle_t = \mathbb{E}\big[|b_i^M(t, C_i) - b_i^\infty(t, C_i)|^2\big]^{1/2}.$$

Then, under Assumptions 2.4–2.6 and 4.6, for any $T \ge 0$ and $L \ge 2$,

$$\sup_{t\le T}\langle W^M - W^\infty\rangle_t \le \frac{K_{T,L}}{M^{0.499}},$$

for sufficiently large $M = M(T, L)$, where $K_{T,L}$ is a constant that depends on T and L. Furthermore, for $L \ge 4$ and $2 \le i \le L-2$,

$$\sup_{t\le T}\mathbb{E}\big[|H_i(X, C_i; W^M(t)) - H_i^*(t, X, b_i^0(C_i))|^2\big]^{1/2} \le \frac{K_{T,L}}{M^{0.499}}.$$

We give a sketch of the proof in Section 5.2. We now discuss the implications of Theorem 5.3, and in particular, the simplifying properties induced by i.i.d. initializations. The complete proofs of this theorem and its corollaries are deferred to Appendix D.

Tracking i.i.d.-initialized neural nets via W^*. For large M, the canonical MF limit W^M is well approximated by W^* (and equivalently by W^∞ as defined in Theorem 5.3), while we recall from Theorem 4.7 that W^M tracks closely the trajectory $\mathbf{W}$ of a large-width i.i.d.-initialized neural network. As such, viewing the bridge through W^M as an intermediate step and taking $M \to \infty$, one can track $\mathbf{W}$ via W^*. To be precise, by combining Proposition 5.2 and Corollary 4.9 with Theorem 5.3, we immediately obtain the following result.

Corollary 5.4. *Under Assumptions 2.4–2.6 and for a set of probability measures $(\rho_{\mathbf{w}}^1, \dots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \dots, \rho_{\mathbf{b}}^L)$ such that*

$$\max_{1\le i\le L} \sup_{m\ge 1} \frac{1}{\sqrt{m}} \Big(\int |u|^m \rho_{\mathbf{w}}^i(du) \Big)^{1/m} \le K, \quad \max_{2\le i\le L} \sup_{m\ge 1} \frac{1}{\sqrt{m}} \Big(\int |v|^m \rho_{\mathbf{b}}^i(dv) \Big)^{1/m} \le K,$$

there exist constants $c_1 \in (0, 0.5)$ and $c_2 \in (0, 1/52)$ such that the following statements hold.

Consider any positive integer $L \ge 2$ and a tuple of positive integers $\mathbf{n} = \{n_1, \dots, n_L\}$ with $n_{L=1}$. Let $n_{\min} = \min_{1\le j\le L-1} n_j$ and $n_{\max} = \max_{1\le j\le L} n_j$. Consider a neural network (2.1) of size $\mathbf{n}$ under $(\rho_{\mathbf{w}}^1, \dots, \rho_{\mathbf{w}}^L, \rho_{\mathbf{b}}^2, \dots, \rho_{\mathbf{b}}^L)$-i.i.d. initialization, and let $\mathbf{W}$ be its trajectory. Also construct the dynamics W^*, as well as the associated quantities, described in Section 5.1.2. Then for any $\delta > 0$ and $T \in \epsilon\mathbb{N}_{\ge 0}$, there exist $n^* = n^*(T, L, c_1, c_2) \ge 1$ and $\epsilon^* = \epsilon^*(T, L, c_1, c_2) \le 1$ such that if $n_{\min} \ge n^*$ and the learning rate $\epsilon \in (0, \epsilon^*)$, for $3 \le i \le L-2$, for any K-Lipschitz and K-bounded test function $\psi : \mathbb{H}_i \to \mathbb{S}$ (where $\mathbb{S}$ is a separable Hilbert space), for any $\delta > 0$, we have, with probability at least $1 - 3\delta - KLn_{\max} \exp(-Kn_{\min}^{c_2})$,

$$\sup_{t\le T} \Big| \frac{1}{n_i} \sum_{j_i=1}^{n_i} \mathbb{E}_Z[\psi(\mathbf{H}_i(\lfloor t/\epsilon \rfloor, X, j_i))] - \mathbb{E}_Z\Big[\int \psi(H_i^*(t, X, v)) \rho_{\mathbf{b}}^i(dv) \Big] \Big| = \tilde{O}(n_{\min}^{-c_1} + \epsilon^{c_1}),$$

where $\tilde{O}$ hides the dependency on T, L and δ as well as the logarithmic factors $\log n_{\max}$ and $\log(1/\epsilon)$. A similar statement holds for $i = 1, 2, L-1, L$. In addition, for any test function $\psi : \mathbb{Y} \times \hat{\mathbb{Y}} \to \mathbb{S}$ which is K-Lipschitz in the second variable, uniformly in the first variable,

$$\sup_{t\le T} \big| \mathbb{E}_Z[\psi(Y, \hat{\mathbf{y}}(\lfloor t/\epsilon \rfloor, X))] - \mathbb{E}_Z[\psi(Y, \hat{y}^*(t, X))] \big| = \tilde{O}(n_{\min}^{-c_1} + \epsilon^{c_1}),$$

with probability at least $1 - 2\delta - KLn_{\max} \exp(-Kn_{\min}^{c_2})$.

Degeneracy of the dynamics. By looking closely at W^*, we observe a simplifying property. By Theorem 5.3, under i.i.d. initialization, for each intermediate layer $i = 3, \dots, L-2$, the weight $w_i^\infty(t, C_{i-1}, C_i)$ is a function of only the time t, its own initialization $w_i^0(C_{i-1}, C_i)$ and the initializations of the adjacent biases $b_{i-1}^0(C_{i-1})$ and $b_i^0(C_i)$, and the bias $b_i^\infty(t, C_i)$ is a function of only the time t and its own initialization $b_i^0(C_i)$. When we further assume constant initial biases (i.e., $b_i^0(C_i) = B_i$ is a constant almost surely for all $i \geq 2$), $w_i^\infty(t, C_{i-1}, C_i)$ is a function of only the time t and its own initialization, and $b_i^\infty(t, C_i)$ is almost surely only a function of time t, regardless of C_i. Consequently, in this scenario, because the initialization is independent across layers, the weights of intermediate layers remain mutually independent at all time, for depth $L \geq 5$, in the infinite-width limit.

The theorem in fact further asserts that degeneracy can already be observed for $L \geq 4$. In particular, for $2 \leq i \leq L-2$, if the initial bias $b_i^0(\cdot) = B_i$ is a constant, then

$$\mathbb{E}\big[|H_i(X, C_i; W^M(t)) - H_i^*(t, X, B_i)|^2\big]^{1/2} \leq \frac{K_{T,L}}{M^{0.499}}.$$

Note that $H_i^*(t, X, B_i)$ is independent of C_i. This suggests that at any training time t, the neurons of each intermediate layer i compute the same function of the data input $x \mapsto H_i^*(t, x, B_i)$ in the infinite-width limit. This is formalized directly for the neural network $\mathbf{W}$ in the following.

Corollary 5.5. *Consider the same setting as Corollary 5.4 with $L \geq 4$. For $2 \leq i \leq L-2$, supposing that $b_i^0(C_i) = B_i$ is a constant almost surely, then we have, for any $t \leq T$, with probability at least $1 - 3\delta - KLn_{\max}\exp(-Kn_{\min}^{c_2})$,*

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} \mathbb{E}_Z\big[|\mathbf{H}_i(\lfloor t/\epsilon \rfloor, X, j_i) - H_i^*(t, X, B_i)|^2\big]\Big)^{1/2} = \tilde{O}(n_{\min}^{-c_1} + \epsilon^{c_1}).$$

Thus, by Markov's inequality, if one is to pick at random a neuron $j_i \in [n_i]$ at layer i from the neural network $\mathbf{W}$ at the training step $\lfloor t/\epsilon \rfloor$, for $2 \leq i \leq L-2$, then with high probability, this neuron would compute the function $x \mapsto H_i^*(t, x, B_i)$ which is independent of the index j_i.

Collapse to effectively one parameter per layer. Further consideration to standard neural network architectures reveals a stronger simplifying property. The next consequence of Theorem 5.3 is that with i.i.d. initialization and constant initial biases, for each intermediate layer $i = 3, \dots, L-2$, the weight $w_i^\infty(t, c_{i-1}, c_i)$ translates by a quantity that is independent of c_{i-1} and c_i, provided that $\sigma_i^{\mathbf{w}}$ satisfies a certain condition. This condition holds for unregularized standard fully-connected or convolutional neural networks (see Examples 2.1–2.2). So, for these networks, in the infinite-width limit, with i.i.d. initialization and constant initial biases, the dynamics of the weight at each intermediate layer reduces to a single deterministic translation parameter.

Corollary 5.6. *Under the same setting as Theorem 5.3 with $L \geq 5$, assume that $b_i^0(C_i) = B_i$ is a constant almost surely for all $i \geq 2$. Further assume that for each $i \in \{3, \ldots, L-2\}$, there exists a function $\bar{\sigma}_i^{\mathbf{w}}$ that satisfies*

$$\sigma_i^{\mathbf{w}}(\Delta, w, b, g, h) = \bar{\sigma}_i^{\mathbf{w}}(\Delta, b, g, h),$$

i.e., $\sigma_i^{\mathbf{w}}$ does not depend on the second variable. Then there are differentiable functions $w_i^{\#}(t)$ such that for $3 \leq i \leq L-2$, almost surely, for any $t \geq 0$,

$$w_i^{\infty}(t, C_{i-1}, C_i) - w_i^{\infty}(0, C_{i-1}, C_i) = w_i^{\#}(t).$$

5.2. Proof sketch of Theorem 5.3

Sketch of proof for Theorem 5.3. We will use $K_{T,L}$ to denote a generic constant that depends on T and L and may change from line to line. The main argument exploits the construction in Section 5.1.1 of the canonical neuronal embedding in a suitable way. To illustrate the idea, consider

$$D_i(t) = \mathbb{E}\big[|H_i(X, C_i; W^{\infty}(t)) - H_i^*(t, X, b_i^0(C_i))|^2\big].$$

We aim to show that for $t \leq T$,

$$D_i(t) \leq \frac{K_{T,L}}{M}.$$

For brevity, define

$$g(u_i, v_{i-1}, v_i) = \phi_i\big(w_i^*(t, u_i, v_{i-1}, v_i), b_i^*(t, v_i), H_{i-1}^*(t, x, v_{i-1})\big).$$

We recall $w_i^0(C_{i-1}, C_i) = \mathfrak{q}_i(\theta_{i-1}, \theta_i)(\lambda_i)$, $b_i^0(C_i) = \mathfrak{p}_i(\theta_i)(\lambda_i)$ and $b_{i-1}^0(C_{i-1}) = \mathfrak{p}_{i-1}(\theta_{i-1})(\lambda_{i-1})$ from the construction (5.1)–(5.5). To make use of the canonical neuronal embedding's construction, we consider a decomposition of the following squared quantity:

$$\begin{aligned}
&\mathbb{E}_{C_i}\big[\big|\mathbb{E}_{C_{i-1}}[g(w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i))]\big|^2\big] \\
&\overset{(a)}{=} \mathbb{E}_{\theta_{i-1}, \lambda_{i-1}, \theta'_{i-1}, \lambda'_{i-1}, \theta_i, \lambda_i}\big[\big\langle g\big(\mathfrak{q}_i(\theta_{i-1}, \theta_i)(\lambda_i), \mathfrak{p}_{i-1}(\theta_{i-1})(\lambda_{i-1}), \mathfrak{p}_i(\theta_i)(\lambda_i)\big), \\
&\qquad g\big(\mathfrak{q}_i(\theta'_{i-1}, \theta_i)(\lambda_i), \mathfrak{p}_{i-1}(\theta'_{i-1})(\lambda'_{i-1}), \mathfrak{p}_i(\theta_i)(\lambda_i)\big)\big\rangle\big] \\
&\overset{(b)}{=} \mathbb{E}_{\theta_{i-1}, \theta'_{i-1}}\Big[\mathbb{I}_{\theta_{i-1} = \theta'_{i-1}} \int \langle g(u_i, v_{i-1}, v_i), g(u_i, v'_{i-1}, v_i)\rangle \rho_{\mathbf{b}}^{i-1}(dv_{i-1}) \rho_{\mathbf{b}}^{i-1}(dv'_{i-1}) \\
&\qquad \times \rho_{\mathbf{b}}^{i}(dv_i) \rho_{\mathbf{w}}^{i}(du_i) \\
&\qquad + \mathbb{I}_{\theta_{i-1} \neq \theta'_{i-1}} \int \langle g(u_i, v_{i-1}, v_i), g(u'_i, v'_{i-1}, v_i)\rangle \rho_{\mathbf{b}}^{i-1}(dv_{i-1}) \rho_{\mathbf{b}}^{i-1}(dv'_{i-1}) \\
&\qquad \times \rho_{\mathbf{b}}^{i}(dv_i) \rho_{\mathbf{w}}^{i}(du_i) \rho_{\mathbf{w}}^{i}(du'_i)\Big]
\end{aligned}$$

$$= \frac{1}{M}\int \langle g(u_i, v_{i-1}, v_i), g(u_i, v'_{i-1}, v_i)\rangle \rho_{\mathbf{b}}^{i-1}(dv_{i-1})\rho_{\mathbf{b}}^{i-1}(dv'_{i-1})\rho_{\mathbf{b}}^{i}(dv_i)\rho_{\mathbf{w}}^{i}(du_i)$$
$$+ \frac{M-1}{M}\int \Big|\int g(u_i, v_{i-1}, v_i)\rho_{\mathbf{w}}^{i}(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big|^2 \rho_{\mathbf{b}}^{i}(dv_i),$$

where in (a), $(\theta'_{i-1}, \lambda'_{i-1}) \sim \mathrm{Unif}([M]) \times P_0$ is an independent copy of $(\theta_{i-1}, \lambda_{i-1})$ and it is independent of (θ_i, λ_i), and (b) follows by the construction of $\mathfrak{p}_{i-1}$, $\mathfrak{p}_i$ and $\mathfrak{q}_i$. We also notice that

$$\mathbb{E}_{C_i}\Big[\Big\langle \mathbb{E}_{C_{i-1}}\big[g\big(w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i)\big)\big],$$
$$\int g(u_i, v_{i-1}, b_i^0(C_i))\rho_{\mathbf{w}}^{i}(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big\rangle\Big]$$
$$= \int \Big|\int g(u_i, v_{i-1}, v_i)\rho_{\mathbf{w}}^{i}(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big|^2 \rho_{\mathbf{b}}^{i}(dv_i).$$

Putting the last two displays together, one easily arrives at the following:

$$\mathbb{E}_{C_i}\big[\big|\mathbb{E}_{C_{i-1}}\big[g\big(w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i)\big)\big] - H_i^*(t, X, b_i^0(C_i))\big|^2\big]$$
$$= \mathbb{E}_{C_i}\Big[\Big|\mathbb{E}_{C_{i-1}}\big[g\big(w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i)\big)\big]$$
$$- \int g(u_i, v_{i-1}, b_i^0(C_i))\rho_{\mathbf{w}}^{i}(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big|^2\Big]$$
$$\le \frac{K}{M}\int |g(u_i, v_{i-1}, v_i)|^2 \rho_{\mathbf{w}}^{i}(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\rho_{\mathbf{b}}^{i}(dv_i) \le \frac{K_{T,L}}{M}.$$

This illustrates the main use of the canonical neuronal embedding's construction. Now from Assumption 2.5, one can show that

$$\mathbb{E}\big[\big|\mathbb{E}_{C_{i-1}}\big[g\big(w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i)\big)\big] - H_i(X, C_i; W^\infty(t))\big|^2\big]$$
$$\le K_{T,L} D_{i-1}(t).$$

Therefore,

$$D_i(t) \le \frac{K_{T,L}}{M} + K_{T,L} D_{i-1}(t).$$

One arrives at the claim from this relation.

The rest of the proof involves similar estimates and Gronwall's inequality. Let us quickly describe the steps for completeness. Similar to the above argument, for

$$D_i^H(t) = \mathbb{E}\big[\big|\Delta_i^H(Z, C_i; W^\infty(t)) - \Delta_i^{H*}(t, Z, b_i^0(C_i))\big|^2\big],$$

we can show that for $t \le T$,

$$D_i^H(t) \le K_{T,L}\frac{\log^{1/2} M}{M}.$$

With the previous two claims, one easily shows:

$$D_i^w(t) \le K_{T,L}\frac{\log^{1/2} M}{M}, \quad D_i^b(t) \le K_{T,L}\frac{\log^{1/2} M}{M},$$

where we define

$$\begin{aligned} D_i^w(t) &= \mathbb{E}\big[\big|\Delta_i^w(Z, C_{i-1}, C_i; W^\infty(t)) \\ &\qquad - \Delta_i^{w*}(t, Z, w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i))\big|^2\big], \\ D_i^b(t) &= \mathbb{E}\big[|\Delta_i^b(Z, C_i; W^\infty(t)) - \Delta_i^{b*}(t, Z, b_i^0(C_i))\big|^2\big]. \end{aligned}$$

The next step is to show that for $2 \le i \le L$, any $t \le T$ and any $B \ge 0$,

$$\begin{aligned} &\mathbb{E}\big[\big|\mathbb{E}_Z\big[\Delta_i^w(Z, C_{i-1}, C_i; W^M(t)) - \Delta_i^w(Z, C_{i-1}, C_i; W^\infty(t))\big]\big|^2\big]^{1/2} \\ &\quad \le K_{T,L}((1+B)\langle W^M - W^\infty\rangle_t + e^{-KB^2}). \end{aligned}$$

With this, we then arrive at the following:

$$\begin{aligned} &\mathbb{E}\big[\big|\Delta_i^w(Z, C_{i-1}, C_i; W^M(t)) - \Delta_i^{w*}(t, Z, w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i))\big|^2\big]^{1/2} \\ &\le |D_i^w(t)|^{1/2} + \mathbb{E}\big[\big|\Delta_i^w(Z, C_{i-1}, C_i; W^M(t)) - \Delta_i^w(Z, C_{i-1}, C_i; W^\infty(t))\big|^2\big]^{1/2} \\ &\le K_{T,L}\Big(\frac{\log^{1/4} M}{M^{1/2}} + (1+B)\langle W^M - W^\infty\rangle_t + e^{-KB^2}\Big). \end{aligned}$$

A similar result holds for Δ_i^b. Hence, we obtain that for all $t \le T$,

$$\langle W^M - W^\infty\rangle_t \le K_{T,L}\int_0^t \Big(\frac{\log^{1/4} M}{M^{1/2}} + (1+B)\langle W^M - W^\infty\rangle_t + e^{-KB^2}\Big)\, ds.$$

Since $\langle W^M - W^\infty\rangle_0 = 0$, Gronwall's inequality implies that

$$\sup_{t\le T}\langle W^M - W^\infty\rangle_t \le K_{T,L}\inf_{B>0}\Big[\Big(\frac{\log^{1/4} M}{M^{1/2}} + e^{-KB^2}\Big)e^{K_{T,L}(1+B)}\Big] \le K_{T,L}\frac{1}{M^{0.499}},$$

for sufficiently large M. This proves the main statement in Theorem 5.3; the other statement follows easily. ∎

6. Convergence to global optimum: two-layer and three-layer networks with i.i.d. initialization

In this section, we prove several global convergence guarantees for fully-connected neural networks (without biases) with $L \le 3$ and i.i.d. initializations. A key element here is a certain universal approximation property that holds at *any* finite training time. This is shown using a tool from algebraic topology.

6.1. Warm-up: the case $L = 2$

Our first result is that in the case of two-layer fully-connected neural networks, the MF limit converges to the global optimum under some genericity assumptions on the initialization distribution. Before we proceed, we specify the two-layer network under consideration and its training:

$$\begin{aligned} \hat{\mathbf{y}}(k,x) &= \varphi_2(\mathbf{H}_2(k,x,1)), \\ \mathbf{H}_2(k,x,1) &= \frac{1}{n_1}\sum_{j_1=1}^{n_1} \mathbf{w}_2(k,j_1,1)\,\varphi_1(\langle \mathbf{w}_1(k,j_1),x\rangle), \end{aligned} \tag{6.1}$$

in which $\mathbf{w}_1(k,j_1) \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, $\varphi_1 \colon \mathbb{R} \to \mathbb{R}$, $\mathbf{w}_2(k,j_1,1) \in \mathbb{R}$ and $\varphi_2 \colon \mathbb{R} \to \mathbb{R}$. We train the network with SGD with respect to the loss $\mathscr{L} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ and the data $z(k) = (x(k), y(k))$ drawn independently at time k:

$$\begin{aligned} \mathbf{w}_2(k+1,j_1,1) - \mathbf{w}_2(k,j_1,1) &= -\epsilon \partial_2 \mathscr{L}(y(k), \hat{\mathbf{y}}(t,x(k)))\varphi_2'(\mathbf{H}_2(t,x(k),1)) \\ &\quad \times \varphi_1(\mathbf{w}_1(k,j_1),x(k)), \\ \mathbf{w}_1(k+1,j_1) - \mathbf{w}_1(k,j_1) &= -\epsilon \partial_2 \mathscr{L}(y(k), \hat{\mathbf{y}}(t,x(k)))\varphi_2'(\mathbf{H}_2(t,x(k),1)) \\ &\quad \times \mathbf{w}_2(k,j_1,1)\varphi_1'(\langle \mathbf{w}_1(k,j_1),x(k)\rangle)x(k). \end{aligned}$$

Here $\epsilon \in \mathbb{R}_{>0}$ is the learning rate. The corresponding MF ODEs are

$$\begin{aligned} \frac{\partial}{\partial t} w_2(t,c_1,1) &= -\mathbb{E}_Z\big[\partial_2 \mathscr{L}(Y, \hat{y}(t,X))\varphi_2'(H_2(t,X,1))\varphi_1(w_1(t,c_1),X)\big], \\ \frac{\partial}{\partial t} w_1(t,c_1) &= -\mathbb{E}_Z\big[\partial_2 \mathscr{L}(Y, \hat{y}(t,X))\varphi_2'(H_2(t,X,1))w_2(t,c_1,1) \\ &\quad \times \varphi_1'(\langle w_1(t,c_1),X\rangle)X\big], \end{aligned}$$

in which for $f_1 \colon \Omega_1 \to \mathbb{R}^d$ and $f_2 \colon \Omega_1 \to \mathbb{R}$, we define

$$\hat{y}(x; f_1, f_2) = \varphi_2(H_2(x; f_1, f_2)), \quad H_2(x; f_1, f_2) = \mathbb{E}_{C_1}[f_2(C_1)\varphi_1(\langle f_1(C_1), x\rangle)],$$

and $\hat{y}(t,x)$ and $H_2(t,x,1)$ are short-hands notations when $f_1 = w_1(t,\cdot)$ and $f_2 = w_2(t,\cdot,1)$. It is easy to see that this network fits into our framework. In particular, under the coupling procedure in Section 4.1, our framework allows to study the following initialization scheme:

$$\big\{(\mathbf{w}_1(0,j_1), \mathbf{w}_2(0,j_1,1))\big\}_{j_1 \in [n_1]} \sim \rho^0 \quad \text{i.i.d.}$$

for suitable probability measure ρ^0 over $\mathbb{R}^d \times \mathbb{R}$. In this case, $\rho^0 = \mathrm{Law}(w_1(0,C_1), w_2(0,C_1,1))$. To measure the training quality, we consider the population loss

$$\mathcal{L}(f_1, f_2) = \mathbb{E}_Z[\mathscr{L}(Y, \hat{y}(X; f_1, f_2))].$$

Assumption 6.1. Consider the MF limit corresponding to the network (6.1), such that they are coupled together by the coupling procedure in Section 4.1. We consider the following assumptions:

(1) *Initialization*: The initialization law ρ^0 satisfies

$$\max\Big(\sup_{m\geq 1}\frac{1}{\sqrt{m}}\mathbb{E}_{C_1}[|w_1(0,C_1)|^m]^{1/m}, \sup_{m\geq 1}\frac{1}{\sqrt{m}}\mathbb{E}_{C_1}[|w_2(0,C_1,1)|^m]^{1/m}\Big)\leq K.$$

(2) *Diversity*: The support of ρ^0 contains the graph of a continuous function $F:\mathbb{R}^d\to\mathbb{R}$ such that $|F(u)|\leq K$ for all $u\in\mathbb{R}^d$.

(3) *Regularity*: φ_1 is K-bounded, φ_1' and φ_2' are K-bounded and K-Lipschitz, φ_2' is non-zero everywhere, $\partial_2\mathcal{L}(\cdot,\cdot)$ is K-Lipschitz in the second variable and K-bounded,[3] and $|X|\leq K$ with probability 1.

(4) *Convergence*: There exist limits $\bar{w}_1$ and $\bar{w}_2$ such that as $t\to\infty$, there exists a coupling π_t of P_1 and itself such that

$$\mathbb{E}_{\pi_t}\big[|\bar{w}_2(C_1)||w_1(t,C_1')-\bar{w}_1(C_1)|+|w_2(t,C_1',1)-\bar{w}_2(C_1)|\big]\to 0$$

for $(C_1,C_1')\sim\pi_t$. Furthermore, ess-sup $|\frac{\partial}{\partial t}w_2(t,C_1,1)|\to 0$.

(5) *Universal approximation*: $\{\varphi_1(\langle u,\cdot\rangle):u\in\mathbb{R}^d\}$ has dense span in $L^2(\mathcal{P}_X)$ (the space of square integrable functions with respect to the measure $\mathcal{P}_X$, which is the distribution of the input X).

Notice that if $(w_1(t,C_1),w_2(t,C_1,1))$ converges to $(\bar{w}_1(C_1),\bar{w}_2(C_1))$ in the Wasserstein-2 distance as $t\to\infty$, then one can prove the first part of the convergence condition in Assumption 6.1 via the initialization and regularity conditions and Lemma 3.2.

We state the main result.

Theorem 6.2. *Consider the MF limit corresponding to the network* (6.1)*, such that they are coupled together by the coupling procedure in Section* 4.1*. Under Assumption* 6.1*, the following hold*:

- Case 1 (convex loss): *If $\mathcal{L}$ is convex in the second variable, then*

$$\lim_{t\to\infty}\mathcal{L}(W(t))=\inf_{f_1,f_2}\mathcal{L}(f_1,f_2)=\inf_{\tilde{y}}\mathbb{E}_Z[\mathcal{L}(Y,\tilde{y}(X))].$$

- Case 2 (generic non-negative loss): *Suppose $\partial_2\mathcal{L}(y,\hat{y})=0$ implies $\mathcal{L}(y,\hat{y})=0$. If $y=y(x)$ is a function of x, then $\mathcal{L}(W(t))=0$ as $t\to\infty$.*

The proof is deferred to Appendix E. We refer the readers to Section 6.2.1 where we present a high-level proof plan for the three-layer case, which is also applicable to

[3] We denote by $\partial_2\mathcal{L}(\cdot,\cdot)$ the partial derivative of $\mathcal{L}$ with respect to the second variable.

the present two-layer case. The following result is straightforward from Theorem 6.2 and Corollary 4.9.

Corollary 6.3. *Consider the neural network* (6.1)*. Under the same setting as Theorem* 6.2*, in Case 1,*

$$\lim_{t\to\infty}\lim_{n_1\to\infty}\lim_{\epsilon\to 0}\mathbb{E}_Z[\mathscr{L}(Y,\hat{\mathbf{y}}(\lfloor t/\epsilon\rfloor,X))]=\inf_{f_1,f_2}\mathcal{L}(f_1,f_2)=\inf_{\tilde{y}}\mathbb{E}_Z[\mathscr{L}(Y,\tilde{y}(X))]$$

in probability, and in Case 2, the same holds with the right-hand side being 0*.*

Let us make a remark on the setting. Examples of suitable φ_1 include sigmoid/tanh activation, sinusoids and the Gaussian pdf, whose universal approximation is known [6, 10] (where we assume the convention that the last entry of the data input x is 1). Examples of suitable φ_2 include smoothened leaky-ReLU, sigmoid/tanh and linear activation. Examples of suitable (and convex) loss $\mathscr{L}$ include Huber loss and exponential loss. Importantly, $\mathscr{L}$ needs not be convex. Assumption 6.1 (4) is technical and does not seem removable. Note that this assumption specifies the mode of convergence and is not an assumption on the limits $\bar{w}_1$ and $\bar{w}_2$. In particular, the first condition (convergence in moment) of Assumption 6.1 (4) is a common assumption in the literature [9]. See also Section 9 where we further this discussion in the context of prior works.

Regarding the uniform convergence condition ess-sup $|\frac{\partial}{\partial t}w_2(t,C_1,1)|\to 0$ in Assumption 6.1 (4), there is a converse relation between global convergence and this condition. Thus, this uniform convergence condition gives a sharp characterization of global convergence.

Proposition 6.4. *Consider the MF limit corresponding to the network* (6.1)*, such that they are coupled together by the coupling procedure in Section* 4.1*. Suppose that the initialization and regularity assumptions* (*i.e., the first and third assumptions*) *of Assumption* 6.1 *hold, and that* $\mathscr{L}(y,\hat{y})\to\infty$ *as* $|\hat{y}|\to\infty$ *for each* y*. Then the following hold*:

- Case 1 (convex loss): *If* $\mathscr{L}$ *is convex in the second variable and* $\mathcal{L}(W(t))\to\inf_F\mathcal{L}(F)$ *as* $t\to\infty$*, then it must be that*

$$\sup_{c_1\in\Omega_1}\left|\frac{\partial}{\partial t}w_2(t,c_1,1)\right|\to 0\quad \text{as } t\to\infty.$$

- Case 2 (generic non-negative loss): *Suppose* $\partial_2\mathscr{L}(y,\hat{y})=0$ *implies* $\mathscr{L}(y,\hat{y})=0$*, and* $y=y(x)$ *is a function of* x*. If* $\mathcal{L}(W(t))\to 0$ *as* $t\to\infty$*, then the same conclusion also holds.*

Such a converse result was shown in the work [39] for two-layer neural networks. It is also a special case of Proposition 7.3, which is a similar converse for multilayer networks; hence, we shall omit the proof of Proposition 6.4.

6.2. The case $L = 3$

We now turn to the case of three-layer networks, $L = 3$. Our development here applies insights already seen in the case $L = 2$, most notably the universal approximation property at the first layer and the topology argument. Our present case is complicated by the presence of a third layer, which requires extra conditions to ensure that the same proof technique is applicable. We again stress that, similar to the case $L = 2$, here we do not rely critically on any convexity property, and the same proof of global convergence should extend beyond the specific network architecture to be considered here (the network (6.2) below).

Before we proceed, we specify the three-layer network under consideration and its training. We also follow the development for i.i.d. initialization in Section 5; in particular, we work with the infinite-M MF limit.

Three-layer network. For $\mathbf{n} = \{n_1, n_2\}$, we consider the following neural network:

$$\begin{aligned}\hat{\mathbf{y}}(k, x) &= \varphi_3(\mathbf{H}_3(k, x, 1)), \qquad (6.2)\\ \mathbf{H}_3(k, x, 1) &= \frac{1}{n_2}\sum_{j_2=1}^{n_2} \mathbf{w}_3(k, j_2, 1)\,\varphi_2(\mathbf{H}_2(k, x, j_2)),\\ \mathbf{H}_2(k, x, j_2) &= \frac{1}{n_1}\sum_{j_1=1}^{n_1} \mathbf{w}_2(k, j_1, j_2)\,\varphi_1(\langle \mathbf{w}_1(k, j_1), x\rangle),\end{aligned}$$

in which $x \in \mathbb{R}^d$, $\mathbf{w}_1(k, j_1) \in \mathbb{R}^d$, $\mathbf{w}_2(k, j_1, j_2) \in \mathbb{R}$, $\mathbf{w}_3(k, j_2, 1) \in \mathbb{R}$, $\varphi_1 \colon \mathbb{R} \to \mathbb{R}$, $\varphi_2 \colon \mathbb{R} \to \mathbb{R}$, $\varphi_3 \colon \mathbb{R} \to \mathbb{R}$ and $k \in \mathbb{N}_{\geq 0}$ indicating the discrete time. We train the network with SGD with respect to the loss $\mathscr{L} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ and the data $z(k) = (x(k), y(k))$ drawn independently at time k:

$$\begin{aligned}&\mathbf{w}_3(k+1, j_2, 1) - \mathbf{w}_3(k, j_2, 1)\\ &\quad = -\epsilon\xi_3(k\epsilon)\partial_2\mathscr{L}(y(k), \hat{\mathbf{y}}(k, x(k)))\,\varphi_3'(\mathbf{H}_3(k, x(k), 1))\,\varphi_2(\mathbf{H}_2(k, x(k), j_2)),\\ &\mathbf{w}_2(k+1, j_1, j_2) - \mathbf{w}_2(k, j_1, j_2) = -\epsilon\Delta_2^{\mathbf{H}}(k, z(k), j_2)\,\varphi_1(\langle \mathbf{w}_1(k, j_1), x\rangle),\\ &\mathbf{w}_1(k+1, j_1) - \mathbf{w}_1(k, j_1)\\ &\quad = -\epsilon\Big(\frac{1}{n_2}\sum_{j_2=1}^{n_2} \Delta_2^{\mathbf{H}}(k, z(k), j_2)\mathbf{w}_2(k, j_1, j_2)\Big)\varphi_1'(\langle \mathbf{w}_1(k, j_1), x(k)\rangle)x(k),\end{aligned}$$

in which

$$\Delta_2^{\mathbf{H}}(k, z, j_2) = \partial_2\mathscr{L}(y, \hat{\mathbf{y}}(k, x))\,\varphi_3'(\mathbf{H}_3(k, x, 1))\mathbf{w}_3(k, j_2, 1)\,\varphi_2'(\mathbf{H}_2(k, x, j_2)).$$

Here $\epsilon \in \mathbb{R}_{>0}$ is the learning rate and $\xi_3 \colon \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ is the learning rate schedule for the third layer. Note that here we only consider non-negative ξ_3. We consider the

following (ρ^1, ρ^2, ρ^3)-i.i.d. initialization:

$$\{\mathbf{w}_1(0, j_1)\}_{j_1 \in [n_1]} \sim \rho^1 \text{ i.i.d.}, \quad \{\mathbf{w}_2(0, j_1, j_2)\}_{j_1 \in [n_1], j_2 \in [n_2]} \sim \rho^2 \text{ i.i.d.},$$
$$\{\mathbf{w}_3(0, j_2, 1)\}_{j_2 \in [n_2]} \sim \rho^3 \text{ i.i.d.}$$

all independently.

Infinite-M MF limit. Following Section 5.1.2, in the current context of the three-layer neural network (6.2), we define the dynamics w_1^*, w_2^* and w_3^* as follows:

$$w_3^*(t, u_3) = u_3 - \int_0^t \xi_3(s) \mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}^*(s, X)) \varphi_3'(H_3^*(s, X)) \varphi_2(H_2^*(s, X, u_3))] \, ds,$$
$$w_2^*(t, u_1, u_2, u_3) = u_2 - \int_0^t \mathbb{E}_Z[\Delta_2^{H*}(s, Z, u_3) \varphi_1(\langle w_1^*(s, u_1), X \rangle)] \, ds,$$
$$w_1^*(t, u_1) = u_1 - \int_0^t \mathbb{E}_Z\Big[\Big(\int \Delta_2^{H*}(s, Z, u_3) w_2^*(s, u_1, u_2, u_3) \rho^2(du_2) \rho^3(du_3) \Big)$$
$$\times \varphi_1'(\langle w_1^*(s, u_1), X \rangle) X \Big] \, ds,$$

for all $u_i \in \text{supp}(\rho^i)$, for $i = 1, 2, 3$, in which

$$\hat{y}^*(x; f_1, f_2, f_3) = \varphi_3(H_3^*(x; f_1, f_2, f_3)),$$
$$H_3^*(x; f_1, f_2, f_3) = \int f_3(u_3) \varphi_2(H_2^*(x, u_3; f_1, f_2)) \rho^3(du_3),$$
$$H_2^*(x, u_3; f_1, f_2) = \int f_2(u_1, u_2, u_3) \varphi_1(\langle f_1(u_1), x \rangle) \rho^1(du_1) \rho^2(du_2),$$
$$\Delta_2^{H*}(z, u_3; f_1, f_2, f_3) = \partial_2 \mathcal{L}(y, \hat{y}^*(x; f_1, f_2, f_3)) \varphi_3'(H_3^*(x; f_1, f_2, f_3)) f_3(u_3)$$
$$\times \varphi_2'(H_2^*(x, u_3; f_1, f_2)),$$

and $\hat{y}^*(t, x)$, $H_3^*(t, x)$, $H_2^*(t, x, u_3)$ and $\Delta_2^{H*}(t, z, u_3)$ are their short-hand notations when $f_1 = w_1^*(t, \cdot)$, $f_2 = w_2^*(t, \cdot, \cdot, \cdot)$ and $f_3 = w_3^*(t, \cdot)$. Let us also define $W^*(t) = \{w_1^*(t, \cdot), w_2^*(t, \cdot, \cdot, \cdot), w_3^*(t, \cdot)\}$. To measure the training quality for $W^*(t)$, we consider the population loss $\mathcal{L}(W^*(t))$ in which

$$\mathcal{L}(f_1, f_2, f_3) = \mathbb{E}_Z[\mathcal{L}(Y, \hat{y}^*(X; f_1, f_2, f_3))].$$

This gives the infinite-M MF limit for the neural network (6.2).

Assumption 6.5. Consider the infinite-M MF limit $W^*(t)$ corresponding to the network (6.2). We consider the following assumptions:

(1) *Initialization*: The initialization distributions satisfy

$$\sup_{m \geq 1} \frac{1}{\sqrt{m}} \Big(\int |u_1|^m \rho^1(du_1) \Big)^{1/m} \leq K, \quad \sup_{m \geq 1} \frac{1}{\sqrt{m}} \Big(\int |u_2|^m \rho^2(du_2) \Big)^{1/m} \leq K,$$

$$\sup_{m\geq 1}\frac{1}{\sqrt{m}}\Big(\int |u_3|^m \rho^3(du_3)\Big)^{1/m} \leq K.$$

(2) *Diversity*: The support of ρ^1 is $\mathbb{R}^d$.

(3) *Regularity*: φ_1 and φ_2 are K-bounded, φ_1', φ_2' and φ_3' are K-bounded and K-Lipschitz, φ_2' and φ_3' are non-zero everywhere, $\partial_2\mathcal{L}(\cdot,\cdot)$ is K-Lipschitz in the second variable and K-bounded, and $|X| \leq K$ with probability 1.

(4) *Convergence*: There exist functions $\bar{w}_1$, $\bar{w}_2$ and $\bar{w}_3$ such that as $t \to \infty$, there exists a coupling π_t of $\rho^1 \times \rho^2 \times \rho^3$ and itself such that

$$\int (1+|\bar{w}_3(u_3)|)|\bar{w}_3(u_3)||\bar{w}_2(u_1,u_2,u_3)||w_1^*(t,u_1') - \bar{w}_1(u_1)|\, d\pi_t(u,u') \to 0,$$

$$\int (1+|\bar{w}_3(u_3)|)|\bar{w}_3(u_3)||w_2^*(t,u_1',u_2',u_3') - \bar{w}_2(u_1,u_2,u_3)|\, d\pi_t(u,u') \to 0,$$

$$\int (1+|\bar{w}_3(u_3)|)|w_3^*(t,u_3') - \bar{w}_3(u_3)|\, d\pi_t(u,u') \to 0,$$

for $u = (u_1,u_2,u_3)$ and $u' = (u_1',u_2',u_3')$. Furthermore,

$$\operatorname*{ess\text{-}sup}_{U_i\sim\rho^i} \Big|\frac{\partial}{\partial t} w_2^*(t,U_1,U_2,U_3)\Big| \to 0.$$

(5) *Universal approximation*: $\{\varphi_1(\langle u,\cdot\rangle) : u \in \mathbb{R}^d\}$ has dense span in $L^2(\mathcal{P}_X)$ (the space of square integrable functions with respect to the measure $\mathcal{P}_X$, which is the distribution of the input X).

As a remark, the first part of the convergence assumption follows from the convergence of the tuple $(w_1^*(t,\cdot), w_2^*(t,\cdot,\cdot,\cdot), w_3^*(t,\cdot))$ to $(\bar{w}_1,\bar{w}_2,\bar{w}_3)$ in the Wasserstein-4 distance, i.e.,

$$\inf_{\pi}\int \big(|w_1^*(t,u_1') - \bar{w}_1(u_1)|^4 + |w_2^*(t,u_1',u_2',u_3') - \bar{w}_2(u_1,u_2,u_3)|^4 + |w_3^*(t,u_3') - \bar{w}_3(u_3)|^4\big)\, d\pi(u_1,u_2,u_3,u_1',u_2',u_3') \to 0,$$

where the infimum is over all couplings π of $\rho^1 \times \rho^2 \times \rho^3$ and itself. In particular, one can prove so with the initialization and regularity conditions and Lemma 3.2.

Theorem 6.6. *Consider the infinite-M MF limit $W^*(t)$ corresponding to the network (6.2), under Assumption 6.5. Further assume either*:

- (untrained third layer) $\int \mathbb{I}(u_3 \neq 0)\rho^3(du_3) > 0$ *and* $\xi_3(\cdot) = 0$ (*and therefore* $w_3^*(t,u_3) = u_3$ *unchanged at all* $t \geq 0$), *or*
- (trained third layer) $\xi_3(\cdot) = 1$ *and the initialization satisfies that* $\mathcal{L}(w_1^0,w_2^0,w_3^0) < \mathbb{E}_Z[\mathcal{L}(Y,\varphi_3(0))]$.

Then the following hold:

- Case 1 (convex loss): *If $\mathcal{L}$ is convex in the second variable, then*

$$\lim_{t\to\infty} \mathcal{L}(W^*(t)) = \inf_{\tilde{y}} \mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))].$$

- Case 2 (generic non-negative loss): *Suppose $\partial_2\mathcal{L}(y, \hat{y}) = 0$ implies $\mathcal{L}(y, \hat{y}) = 0$. If $y = y(x)$ is a function of x, then $\mathcal{L}(W^*(t)) \to 0$ as $t \to \infty$.*

The proof is deferred to Section 6.3. While global convergence is proven via the infinite-M MF limit W^*, it is easy to adapt the proof to prove the same for the canonical MF limits that are described in Section 5.1.1, giving a statement similar to the two-layer case (Theorem 6.2). By working with the infinite-M limit, specifically combining Theorem 6.6 with Corollary 5.4, we immediately obtain the following result.

Corollary 6.7. *Consider the neural network* (6.2)*. Under the same setting as Theorem 6.6, in Case 1,*

$$\lim_{t\to\infty} \lim_{n_1,n_2\to\infty} \lim_{\epsilon\to 0} \mathbb{E}_Z[\mathcal{L}(Y, \hat{\mathbf{y}}(\lfloor t/\epsilon \rfloor, X))] = \inf_{\tilde{y}} \mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))].$$

in probability, and in Case 2, the same holds with the right-hand side being 0*. Here $n_1, n_2 \to \infty$ in such a way that $n_{\min} \to \infty$ and $n_{\min}^{-c} \log n_{\max} \to 0$ for any $c > 0$, with $n_{\min} = \min\{n_1, n_2\}$ and $n_{\max} = \max\{n_1, n_2\}$.*

Similar to Section 6.1, here we also have a converse relation between global convergence and the essential supremum condition in Assumption 6.5 (4).

Proposition 6.8. *Consider the infinite-M MF limit $W^*(t)$ corresponding to the network* (6.2)*. Suppose that the initialization and regularity assumptions (i.e., the first and third assumptions) of Assumption 6.5 hold, and that $\mathcal{L}(y, \hat{y}) \to \infty$ as $|\hat{y}| \to \infty$ for each y. Further assume that there exists $\bar{w}_3$ such that as $t \to \infty$, there is a coupling π_t^3 of ρ^3 and itself such that*

$$\int |w_3^*(t, u_3') - \bar{w}_3(u_3)| \, d\pi_t^3(u_3, u_3') \to 0.$$

Then the following hold:

- Case 1 (convex loss): *If $\mathcal{L}$ is convex in the second variable and*

$$\lim_{t\to\infty} \mathcal{L}(W^*(t)) = \inf_{\tilde{y}} \mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))],$$

then it must be that

$$\sup_{u_1\in\mathbb{R}^d, u_2\in\mathrm{supp}(\rho^2)} \mathbb{E}_{U_3\sim\rho^3}\left[\left|\frac{\partial}{\partial t} w_2^*(t, u_1, u_2, U_3)\right|\right] \to 0 \quad \text{as } t \to \infty.$$

- Case 2 (generic non-negative loss): *Suppose* $\partial_2 \mathcal{L}(y, \hat{y}) = 0$ *implies* $\mathcal{L}(y, \hat{y}) = 0$, *and* $y = y(x)$ *is a function of* x. *If* $\mathcal{L}(W^*(t)) \to 0$ *as* $t \to \infty$, *then the same conclusion also holds.*

  We prove Proposition 6.8 in Appendix E.

6.2.1. High-level idea of the proof. Before we proceed, we give a high-level discussion of the proof of Theorem 6.6. This is meant to provide intuitions and explain the technical crux, so our discussion may simplify and deviate from the actual proof. Our first insight is to look at the second layer's weight w_2^*. Recall that

$$\frac{\partial}{\partial t} w_2^*(t, u_1, u_2, u_3) = -\mathbb{E}_Z[\Delta_2^{H*}(Z, u_3; W^*(t))\varphi_1(\langle w_1^*(t, u_1), X\rangle)].$$

At convergence time $t = \infty$, we expect to have zero movement and hence, denoting $\bar{W} = \{\bar{w}_1, \bar{w}_2, \bar{w}_3\}$:

$$\mathbb{E}_Z[\Delta_2^{H*}(Z, u_3; \bar{W})\varphi_1(\langle \bar{w}_1(u_1), X\rangle)] = 0$$

for $u_1 \in \mathrm{supp}(\rho^1)$, $u_3 \in \mathrm{supp}(\rho^3)$. Suppose for the moment that we are allowed to make an additional (strong) assumption on the limit $\bar{w}_1$, that is, $\mathrm{supp}(\bar{w}_1(U_1)) = \mathbb{R}^d$ for $U_1 \sim \rho^1$. This implies that the universal approximation property, described in Assumption 6.5 (5), holds at $t = \infty$; more specifically, it implies $\{\varphi_1(\langle \bar{w}_1(u_1), \cdot\rangle) : u_1 \in \mathrm{supp}(\rho^1)\}$ has dense span in $L^2(\mathcal{P}_X)$. This yields

$$\mathbb{E}_Z[\Delta_2^{H*}(Z, u_3; \bar{W}) \mid X = x] = 0$$

for $\mathcal{P}$-almost every x. Recalling the definition of Δ_2^{H*}, one can then easily show that

$$\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}^*(x; \bar{W})) \mid X = x] = 0.$$

Global convergence follows immediately; for example, in Case 2 of Theorem 6.6, this is equivalent to that $\partial_2 \mathcal{L}(y(x), \hat{y}^*(x; \bar{W})) = 0$ and hence $\mathcal{L}(y(x), \hat{y}^*(x; \bar{W})) = 0$ for $\mathcal{P}$-almost every x. In short, the gradient flow structure of the dynamics of w_2^* provides a seamless way to obtain global convergence. Furthermore, there is no critical reliance on convexity.

However, this plan of attack has a potential flaw in the strong assumption that $\mathrm{supp}(\bar{w}_1(U_1)) = \mathbb{R}^d$, i.e., the universal approximation property holds at convergence time. Indeed, there are setups where it is desirable that $\mathrm{supp}(\bar{w}_1(U_1)) \neq \mathbb{R}^d$, see [7, 22]; for instance, in the case where the neural network is to learn some "sparse and spiky" solution, and hence the weight distribution at convergence time, if successfully trained, cannot have full support. On the other hand, one can entirely expect that if $\mathrm{supp}(w_1^*(0, U_1)) = \mathbb{R}^d$ initially at $t = 0$, then $\mathrm{supp}(w_1^*(t, U_1)) = \mathbb{R}^d$ at any finite $t \geq 0$. The crux of our proof is to show the latter without assuming $\mathrm{supp}(\bar{w}_1(U_1)) = \mathbb{R}^d$.

This is done via an algebraic topology argument, in which the mapping $(t,u)\mapsto M(t,u)$ that maps from $(t,w_1^*(0,u_1))=(t,u_1)$ to $w_1^*(t,u_1)$ is shown to preserve a homotopic structure through time.

6.3. Proof of Theorem 6.6

First, using an algebraic topology argument, we show that if $w_1^*(0,U_1)=U_1\sim\rho^1$ has full support, then so does $w_1^*(t,U_1)$ at any time t. Note that the following result holds beyond the setting of Theorem 6.6.

Lemma 6.9. *Assume $L=3$ and $\mathbb{W}_1=\mathbb{R}^d$ (for some positive integer d), along with Assumptions 2.4–2.6 and 4.6. Under a $(\rho_{\mathbf{w}}^1,\rho_{\mathbf{w}}^2,\rho_{\mathbf{w}}^3)$-i.i.d. initialization, consider the infinite-M limit of the canonical MF limits as described in Section 5.1.2, and in particular the dynamics $\{w_i^*\}_{i=1,2,3}$. Here we disregard the biases by considering $\xi_2^{\mathbf{b}}(\cdot)=\xi_3^{\mathbf{b}}(\cdot)=0$ and $b_2(0,\cdot)=b_3(0,\cdot)=0$. Assume that*

$$\sigma_2^{\mathbf{w}}(\Delta,w,b,g,h)=\bar\sigma_2^{\mathbf{w}}(\Delta,g,h),$$

for some function $\bar\sigma_2^{\mathbf{w}}$, i.e., $\sigma_2^{\mathbf{w}}$ does not depend on the second and third variables. Suppose that the support of $\rho_{\mathbf{w}}^1$ is $\mathbb{W}_1$. Then for all finite time t, the support of $\mathrm{Law}_{U_1\sim\rho_{\mathbf{w}}^1}(w_1^*(t,U_1))$ is $\mathbb{W}_1$.

Proof. Specialized to the current setting, w_1^* and w_2^* satisfy

$$\begin{aligned}
&\frac{\partial}{\partial t}w_1^*(t,u_1)\\
&\quad=-\xi_1^{\mathbf{w}}(t)\mathbb{E}_Z\Big[\sigma_1^{\mathbf{w}}\Big(\int h(t,Z,u_1,u_2,u_3)\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^3(du_3),w_1^*(t,u_1),X\Big)\Big],\\
&\frac{\partial}{\partial t}w_2^*(t,u_1,u_2,u_3)=-\xi_2^{\mathbf{w}}(t)\mathbb{E}_Z[g(t,Z,u_1,u_3)],
\end{aligned}$$

for all $u_1\in\mathrm{supp}(\rho_{\mathbf{w}}^1)$, $u_2\in\mathrm{supp}(\rho_{\mathbf{w}}^2)$ and $u_3\in\mathrm{supp}(\rho_{\mathbf{w}}^3)$, where

$$\begin{aligned}
g(t,z,u_1,u_3)&=\bar\sigma_2^{\mathbf{w}}\big(\Delta_2^{H*}(t,z,u_3),H_2^*(t,x,u_3),H_1^*(t,x,u_1)\big),\\
h(t,z,u_1,u_2,u_3)&=\sigma_1^{\mathbf{H}}\big(\Delta_2^{H*}(t,z,u_3),w_2^*(t,u_1,u_2,u_3),0,\\
&\qquad H_2^*(t,x,u_3),H_1^*(t,x,u_1)\big)\\
&=\sigma_1^{\mathbf{H}}\Big(\Delta_2^{H*}(t,z,u_3),u_2-\int_0^t\xi_2^{\mathbf{w}}(s)\mathbb{E}_Z[g(s,Z,u_1,u_3)]\,ds,0,\\
&\qquad H_2^*(t,x,u_3),H_1^*(t,x,u_1)\Big).
\end{aligned}$$

Here we have shortened the notations to remove dependency on the biases:

$$\begin{gathered}
w_2^*(t,u_1,u_2,u_3)\equiv w_2^*(t,u_1,u_2,u_3,0),\quad \Delta_2^{H*}(t,z,u_3)\equiv\Delta_2^{H*}(t,z,u_3,0),\\
H_2^*(t,x,u_3)\equiv H_2^*(t,x,u_3,0).
\end{gathered}$$

We recall the initialization

$$w_1^*(0,u_1)=u_1,\quad w_2^*(0,\cdot,u_2,\cdot)=u_2,\quad w_2^*(0,u_3,\cdot)=u_3.$$

In the following, we define K_t to be a generic constant that changes with t and which is finite with finite t. We proceed in several steps.

Step 1. We study the function h. We have, from Assumptions 2.6 and 2.4,

$$\begin{aligned}
|\Delta_3^{w*}(t,z,\cdot,\cdot)| &\le K(1+|\Delta_3^{H*}(t,z)|)\le K,\\
|w_3^*(t,u_3,\cdot)| &\le |w_3^*(0,u_3,\cdot)|+K_t=|u_3|+K_t,\\
|\Delta_2^{H*}(t,z,u_3)| &\le K(1+|\Delta_3^{H*}(t,z)|)(1+|w_3^*(t,u_3,\cdot)|)\le K_t(1+|u_3|).
\end{aligned}$$

Consequently, by Assumption 2.6,

$$\begin{aligned}
&|g(t,z,u_1,u_3)|\le K(1+|\Delta_2^{H*}(t,z,u_3)|)\le K_t(1+|u_3|),\\
&|g(t,z,u_1,u_3)-g(t,z,u_1',u_3)|\\
&\quad\le K(1+|\Delta_2^{H*}(t,z,u_3)|)|H_1^*(t,x,u_1)-H_1^*(t,x,u_1')|\\
&\quad\le K_t(1+|u_3|)|H_1^*(t,x,u_1)-H_1^*(t,x,u_1')|,
\end{aligned}$$

for all $u_1\in\mathrm{supp}(\rho_{\mathrm{w}}^1)$ and $u_3\in\mathrm{supp}(\rho_{\mathrm{w}}^3)$. Using these bounds and Assumption 2.6, we obtain

$$\begin{aligned}
&|h(t,z,u_1,u_2,u_3)-h(t,z,u_1',u_2,u_3)|\\
&\quad\le K(1+|\Delta_2^{H*}(t,z,u_3)|)\int_0^t\mathbb{E}_Z[|g(s,Z,u_1,u_3)-g(s,Z,u_1',u_3)|]\,ds\\
&\qquad+K(1+|\Delta_2^{H*}(t,z,u_3)|)\Big(1+|u_2|+\int_0^t\mathbb{E}_Z[|g(s,Z,u_1,u_3)|]\,ds\Big)\\
&\qquad\times\big|H_1^*(t,x,u_1)-H_1^*(t,x,u_1')\big|\\
&\quad\le K_t(1+|u_2|^2+|u_3|^2)\sup_{s\le t}|H_1^*(s,x,u_1)-H_1^*(s,x,u_1')|,
\end{aligned}$$

as well as that

$$\begin{aligned}
|h(t,z,u_1,u_2,u_3)|&\le K(1+|\Delta_2^{H*}(t,z,u_3)|)\\
&\qquad\times\Big(1+|u_2|+\int_0^t\mathbb{E}_Z[|g(s,Z,u_1,u_3)|]\,ds\Big)\\
&\le K_t(1+|u_2|^2+|u_3|^2).
\end{aligned}$$

These are the desired bounds for h.

Step 2. We show that for an arbitrary $T\ge 0$, $w_1^*\colon[0,T]\times\mathbb{W}_1\to\mathbb{W}_1$ is continuous. Now, by using the bound for the function h in Step 1 and Assumptions 2.5 and 2.6,

for $u_1, u_1' \in \mathbb{W}_1$, we have

$$
\begin{aligned}
&\Big|\frac{d}{dt}(w_1^*(t,u_1) - w_1^*(t,u_1'))\Big| \\
&= \Big|\xi_1^{\mathbf{w}}(t)\mathbb{E}_Z\Big[\sigma_1^{\mathbf{w}}\Big(\int h(t,Z,u_1',u_2,u_3)\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^3(du_3), w_1^*(t,u_1'), X\Big) \\
&\qquad - \sigma_1^{\mathbf{w}}\Big(\int h(t,Z,u_1,u_2,u_3)\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^3(du_3), w_1^*(t,u_1), X\Big)\Big]\Big| \\
&\le K|w_1^*(t,u_1) - w_1^*(t,u_1')| + K_t\Big(1 + \int |u_2|^2\rho_{\mathbf{w}}^2(du_2) + \int |u_3|^2\rho_{\mathbf{w}}^3(du_3)\Big) \\
&\qquad \times \mathbb{E}_Z\big[\sup_{s\le t}|H_1^*(s,x,u_1) - H_1^*(s,x,u_1')|\big] \\
&\le K|w_1^*(t,u_1) - w_1^*(t,u_1')| + K_t\mathbb{E}_Z\big[\sup_{s\le t}|H_1^*(s,x,u_1) - H_1^*(s,x,u_1')|\big] \\
&= K|w_1^*(t,u_1) - w_1^*(t,u_1')| + K_t\mathbb{E}_Z\big[\sup_{s\le t}|\phi_1(w_1^*(s,u_1),X) - \phi_1(w_1^*(s,u_1'),X)|\big] \\
&\le K_t \sup_{s\le t}|w_1^*(s,u_1) - w_1^*(s,u_1')|.
\end{aligned}
$$

By Gronwall's inequality,

$$\sup_{s\le t}|w_1^*(s,u_1) - w_1^*(s,u_1')| \le e^{K_t t}|w_1^*(0,u_1) - w_1^*(0,u_1')| = e^{K_t t}|u_1 - u_1'|.$$

Furthermore, by Assumptions 2.4 and 2.6, for $t' \le t$,

$$
\begin{aligned}
&|w_1^*(t,u_1) - w_1^*(t',u_1)| \\
&\quad \le \int_{t'}^t \Big|\xi_1^{\mathbf{w}}(s)\mathbb{E}_Z\Big[\sigma_1^{\mathbf{w}}\Big(\int h(s,Z,u_1,u_2,u_3)\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^3(du_3), w_1^*(s,u_1), X\Big)\Big]\Big| ds \\
&\quad \le K\int_{t'}^t \mathbb{E}_Z\Big[1 + \int |h(s,Z,u_1,u_2,u_3)|\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^3(du_3)\Big] ds \\
&\quad \le K_t\Big(1 + \int |u_2|^2\rho_{\mathbf{w}}^2(du_2) + \int |u_3|^2\rho_{\mathbf{w}}^3(du_3)\Big)|t - t'| \\
&\quad \le K_t|t - t'|.
\end{aligned}
$$

This shows that w_1^* defines a continuous function $w_1^*\colon [0,T]\times\mathbb{W}_1 \to \mathbb{W}_1$.

Step 3. Consider the sphere $\mathbb{S}^d$ which is a compactification of $\mathbb{R}^d$. We can extend w_1^* to a function $M\colon [0,T]\times\mathbb{S}^d \to \mathbb{S}^d$ fixing the point at infinity, which remains a continuous map since $|M(t,u_1) - u_1| = |M(t,u_1) - M(0,u_1)| \le K_T t$. Let $M_t\colon \mathbb{R}^d \to \mathbb{R}^d$ be defined by $M_t(u_1) = M(t,u_1)$. We claim that M_t is surjective for all finite t. Indeed, if M_t fails to be surjective for some t, then for some $p \in \mathbb{S}^d$, $M_t\colon \mathbb{S}^d \to \mathbb{S}^d \setminus \{p\} \to \mathbb{S}^d$ is homotopic to the constant map, but M then gives a homotopy from the identity map M_0 on the sphere to a constant map, which is a contradiction as the sphere $\mathbb{S}^d$ is not contractible. Hence, $w_1^*(t,\cdot)$ is surjective for all finite t.

Now let $U_1 \sim \rho_{\mathbf{w}}^1$, which has full support, and consider $w_1^*(t, U_1)$. Let us assume that $w_1^*(t, U_1)$ does not have full support at some time t, which implies there is an open ball B in $\mathbb{R}^d$ for which $\mathbb{P}(w_1^*(t, U_1) \in B) = 0$. Due to surjectivity and continuity of $u_1 \mapsto w_1^*(t, u_1)$, there is an open set U such that $w_1^*(t, u_1) \in B$ for all $u_1 \in U$. Then $\mathbb{P}(U_1 \in U) = 0$, contradicting the assumption that U_1 has full support. Therefore, $w_1^*(t, U_1)$ must have full support at all $t \geq 0$. ∎

With this lemma, we are ready to prove Theorem 6.6. We recall the setting of Theorem 6.6, and in particular, the neural network (6.2).

Proof of Theorem 6.6. Let $U_i \sim \rho^i$, $i = 1, 2, 3$ independently. It is easy to check that Assumptions 2.4–2.6, as well as the conditions of Lemma 6.9, hold. Therefore, by Lemma 6.9, the support of $\mathrm{Law}(w_1^*(t, U_1))$ is $\mathbb{R}^d$ at all t. We recall from the convergence assumption the limits $\bar{w}_1$, $\bar{w}_2$ and $\bar{w}_3$, and we shall first prove $(\bar{w}_1, \bar{w}_2, \bar{w}_3)$ is a global minimizer of $\mathcal{L}$ in Case 1 and $\mathcal{L}(\bar{w}_1, \bar{w}_2, \bar{w}_3) = 0$ in Case 2.

By the convergence assumption, we have that for any $\epsilon > 0$, there exists $T(\epsilon)$ such that for all $t \geq T(\epsilon)$ and almost surely,

$$\begin{aligned}\epsilon &\geq |\mathbb{E}_Z[\Delta_2^{H*}(t, Z, U_3)\varphi_1(\langle w_1^*(t, U_1), X\rangle)]| \\ &= |\langle \mathbb{E}_Z[\Delta_2^{H*}(t, Z, U_3) \mid X = x], \varphi_1(\langle w_1^*(t, U_1), x\rangle)\rangle_{L^2(\mathcal{P}_X)}|.\end{aligned}$$

Since $\mathrm{Law}(w_1^*(t, U_1))$ has full support, we obtain that for u in a dense subset of $\mathbb{R}^d$,

$$\text{ess-sup}|\langle \mathbb{E}_Z[\Delta_2^{H*}(t, Z, U_3) \mid X = x], \varphi_1(\langle u, x\rangle)\rangle_{L^2(\mathcal{P}_X)}| \leq \epsilon.$$

By continuity of $u \mapsto \varphi_1(\langle u, \cdot\rangle)$ in $L^2(\mathcal{P}_X)$, we extend the above to all $u \in \mathbb{R}^d$. Recall the couplings π_t in Assumption 6.5 (4), since φ_1 is bounded,

$$\begin{aligned}&\mathbb{E}_{(U_3, U_3')\sim\pi_t}\big[\big|\big\langle \mathbb{E}_Z[|\Delta_2^{H*}(t, Z, U_3) - \Delta_2^{H*}(Z, U_3'; \bar{w}_1, \bar{w}_2, \bar{w}_3)| \mid X = x], \\ &\qquad \varphi_1(\langle u, x\rangle)\big\rangle_{L^2(\mathcal{P}_X)}\big|\big] \\ &\leq K\mathbb{E}_{\pi_t}[|\Delta_2^{H*}(t, Z, U_3) - \Delta_2^{H*}(Z, U_3'; \bar{w}_1, \bar{w}_2, \bar{w}_3)|] \\ &\leq K\mathbb{E}_{\pi_t}\big[(1 + |\bar{w}_3(U_3)|)\big(|w_3^*(t, U_3') - \bar{w}_3(U_3)| + |\bar{w}_3(U_3)||w_2^*(t, U_1', U_2', U_3') \\ &\qquad - \bar{w}_2(U_1, U_2, U_3)| + |\bar{w}_3(U_2)||\bar{w}_2(U_1, U_2, U_3)||w_1^*(t, U_1') - \bar{w}_1(U_1)|\big)\big],\end{aligned}$$

where the last step is by the regularity assumption, similar to the calculation in the proof of Theorem 6.2. Recall that the right-hand side converges to 0 as $t \to \infty$. We thus obtain that for all $u \in \mathbb{R}^d$,

$$\mathbb{E}_{U_3}[|\langle \mathbb{E}_Z[\Delta_2^{H*}(Z, U_3; \bar{w}_1, \bar{w}_2, \bar{w}_3) \mid X = x], \varphi_1(\langle u, x\rangle)\rangle_{L^2(\mathcal{P}_X)}|] = 0,$$

which yields that for all $u \in \mathbb{R}^d$ and almost surely,

$$|\langle \mathbb{E}_Z[\Delta_2^{H*}(Z, U_3; \bar{w}_1, \bar{w}_2, \bar{w}_3) \mid X = x], \varphi_1(\langle u, x\rangle)\rangle_{L^2(\mathcal{P}_X)}| = 0.$$

Here we note that, by the regularity assumption,

$$|\mathbb{E}_Z[\Delta_2^{H*}(Z, U_3; \bar{w}_1, \bar{w}_2, \bar{w}_3) \mid X = x]| \le K|\bar{w}_3(U_3)|,$$

and so $\mathbb{E}_Z[\Delta_2^{H*}(Z, u_3; \bar{w}_1, \bar{w}_2, \bar{w}_3) \mid X = x]$ is in $L^2(\mathcal{P}_X)$ for almost every u_3. Since $\{\varphi_1(\langle u, \cdot\rangle) : u \in \mathbb{R}^d\}$ has dense span in $L^2(\mathcal{P}_X)$, we have

$$\mathbb{E}_Z[\Delta_2^{H*}(Z, u_3; \bar{w}_1, \bar{w}_2, \bar{w}_3) \mid X = x] = 0$$

for $\mathcal{P}_X$-almost every x and almost every u_3, and hence

$$\begin{aligned}&\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}^*(X; \bar{w}_1, \bar{w}_2, \bar{w}_3)) \mid X = x] \\ &\quad \times \varphi_3'(H_3^*(x; \bar{w}_1, \bar{w}_2, \bar{w}_3))\bar{w}_3(u_3)\varphi_2'(H_2^*(x, u_3; \bar{w}_1, \bar{w}_2)) = 0.\end{aligned}$$

We note that our assumptions guarantee that $\mathbb{P}(\bar{w}_3(U_3) \neq 0)$ is positive:

- In the case $\int \mathbb{I}(u_3 \neq 0)\rho^3(du_3) > 0$ and $\xi_3(\cdot) = 0$, it is obvious that

$$\mathbb{P}(\bar{w}_3(U_3) \neq 0) > 0.$$

- In the case $\mathcal{L}(w_1^0, w_2^0, w_3^0) < \mathbb{E}_Z[\mathcal{L}(Y, \varphi_3(0))]$, it can be easily checked that

$$\mathcal{L}(w_1^*(t, \cdot), w_2^*(t, \cdot, \cdot, \cdot), w_3^*(t, \cdot)) \le \mathcal{L}(w_1^*(t', \cdot), w_2^*(t', \cdot, \cdot, \cdot), w_3^*(t', \cdot)),$$

  for $t \ge t'$. This is in fact a standard property of gradient flows. In particular, setting $t' = 0$ and taking $t \to \infty$, it is easy to see that

$$\mathcal{L}(\bar{w}_1, \bar{w}_2, \bar{w}_3) \le \mathcal{L}(w_1^0, w_2^0, w_3^0) < \mathbb{E}_Z[\mathcal{L}(Y, \varphi_3(0))].$$

If $\mathbb{P}(\bar{w}_3(U_3) = 0) = 1$, then $\mathcal{L}(\bar{w}_1, \bar{w}_2, \bar{w}_3) = \mathbb{E}_Z[\mathcal{L}(Y, \varphi_3(0))]$, a contradiction. Then since φ_2' and φ_3' are strictly non-zero, we have

$$\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}^*(X; \bar{w}_1, \bar{w}_2, \bar{w}_3)) \mid X = x] = 0$$

for $\mathcal{P}_X$-almost every x.

In Case 1, since $\mathcal{L}$ convex in the second variable, it follows that for any measurable function $\tilde{y}(x)$,

$$\begin{aligned}&\mathcal{L}(y, \tilde{y}(x)) - \mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2, \bar{w}_3)) \\ &\quad \ge \partial_2 \mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2, \bar{w}_3))(\tilde{y}(x) - \hat{y}(x; \bar{w}_1, \bar{w}_2, \bar{w}_3)).\end{aligned}$$

Taking the expectation, we get $\mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))] \ge \mathcal{L}(\bar{w}_1, \bar{w}_2, \bar{w}_3)$, i.e., $(\bar{w}_1, \bar{w}_2, \bar{w}_3)$ is a global minimizer of $\mathcal{L}$.

In Case 2, since y is a function of x, we obtain $\partial_2 \mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2, \bar{w}_3)) = 0$ and hence $\mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2, \bar{w}_3)) = 0$ for $\mathcal{P}_X$-almost every x.

Finally, to connect $\mathcal{L}(\bar{w}_1, \bar{w}_2, \bar{w}_3)$ with $\mathcal{L}(W^*(t))$ in the limit $t \to \infty$, we have

$$\begin{aligned}
&|\mathcal{L}(W^*(t)) - \mathcal{L}(\bar{w}_1, \bar{w}_2, \bar{w}_3)| \\
&\quad = |\mathbb{E}_Z[\mathcal{L}(Y, \hat{y}^*(t, X)) - \mathcal{L}(Y, \hat{y}^*(X; \bar{w}_1, \bar{w}_2, \bar{w}_3))]| \\
&\quad \leq K\mathbb{E}_Z[|\hat{y}^*(t, X) - \hat{y}^*(X; \bar{w}_1, \bar{w}_2, \bar{w}_3)|] \\
&\quad \leq K\mathbb{E}_{\pi_t}\big[|w_3^*(t, U_3') - \bar{w}_3(U_3)| + |\bar{w}_3(U_3)||w_2^*(t, U_1', U_2', U_3') - \bar{w}_2(U_1, U_2, U_3)| \\
&\qquad + |\bar{w}_3(U_3)||\bar{w}_2(U_1, U_2, U_3)||w_1^*(t, U_1') - \bar{w}_1(U_1)|\big],
\end{aligned}$$

which tends to 0 as $t \to \infty$. This completes the proof. ∎

7. Convergence to global optimum: multilayer networks with correlated initializations

In Section 6, we proved that global convergence is guaranteed for networks with $L \leq 3$ and i.i.d. initializations. Underlying these results is a universal approximation property that holds throughout the course of training, and this is shown for quite general data distributions. Recall from Section 5 that i.i.d. initializations cause a certain degenerate behavior in the network with $L \geq 4$. In particular, by Corollary 5.5, neurons at intermediate layers collapse to the same function of the input and therefore are not expected to span the space of functions of the input $L^2(\mathcal{P}_X)$. In other words, these intermediate layers become a bottleneck that hinders universal approximation in the context of more than three layers and general data distributions.

To attain meaningful training, this suggests a departure from i.i.d. initializations. In particular, we propose a correlated initialization scheme that resolves the aforementioned bottleneck problem. To be precise, the key idea lies in the new concept of bidirectional diversity. A similar concept has been encountered in Section 6; for instance, diversity in the two-layer case in Section 6.1 refers to the full support condition of the first layer's weight distribution in the Euclidean space, implied at initialization $t = 0$ by Assumption 6.1 (2) and shown to hold at any finite time t by Lemma E.1. Here bidirectional diversity furthers this idea to the multilayer case with arbitrary depths. Firstly, it is realized in function spaces that are naturally described by our neuronal embedding framework. Secondly, it is bidirectional: roughly speaking, for intermediate layers, diversity holds in both the forward and backward passes. The effect of bidirectional diversity is that a certain universal approximation property, at any finite training time t, is propagated from the first layer to the second to last one. Importantly, the proposed correlated initialization only ensures bidirectional diversity at initialization $t = 0$, but it is the learning dynamics that automatically maintains bidirectional diversity at any finite t. This fact is again shown by a topological invariance argument.

In the following, we first describe the multilayer fully-connected neural network under consideration and its corresponding MF limit. We then describe the proposed correlated initialization, and state and prove the global convergence guarantee.

7.1. Multilayer fully-connected neural network

We consider the following L-layer fully-connected network:

$$\begin{aligned}\hat{\mathbf{y}}(x;\mathbf{W}(k)) &= \varphi_L(\mathbf{H}_L(x,1;\mathbf{W}(k))), \qquad (7.1)\\ \mathbf{H}_i(x,j_i;\mathbf{W}(k)) &= \frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}} \mathbf{w}_i(k,j_{i-1},j_i)\varphi_{i-1}(\mathbf{H}_{i-1}(x,j_{i-1};\mathbf{W}(k))),\\ \mathbf{H}_1(x,j_1;\mathbf{W}(k)) &= \langle \mathbf{w}_1(k,j_1),x\rangle,\end{aligned}$$

for $i = L,\dots,2$, in which $x \in \mathbb{R}^d$ is the input, $\mathbf{W}(k) = \{\mathbf{w}_1(k,\cdot), \mathbf{w}_i(k,\cdot,\cdot) : i = 2,\dots,L\}$ is the weight with $\mathbf{w}_1(k,j_1) \in \mathbb{R}^d$, $\mathbf{w}_i(k,j_{i-1},j_i) \in \mathbb{R}$, and $\varphi_i : \mathbb{R} \to \mathbb{R}$ is the activation. Here the network has widths $\{n_i\}_{i\le L}$ with $n_L = 1$. We train the network with stochastic gradient descent (SGD) with respect to the loss $\mathcal{L} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_{\ge 0}$ and the data $z(k) = (x(k), y(k)) \in \mathbb{R}^d \times \mathbb{R}$ drawn independently at time k from a training distribution $\mathcal{P}$. Given an initialization $\mathbf{W}(0)$, we update $\mathbf{W}(k)$ according to

$$\begin{aligned}\mathbf{w}_i(k+1,j_{i-1},j_i) &= \mathbf{w}_i(k,j_{i-1},j_i) - \epsilon\xi_i(t\epsilon)\Delta_i^{\mathbf{w}}(z(k),j_{i-1},j_i;\mathbf{W}(k)),\\ \mathbf{w}_1(k+1,j_1) &= \mathbf{w}_1(k,j_1) - \epsilon\xi_1(t\epsilon)\Delta_1^{\mathbf{w}}(z(k),j_1;\mathbf{W}(k)),\end{aligned}$$

for $i = L,\dots,2$, in which $j_i \in [n_i]$, $\epsilon \in \mathbb{R}_{>0}$ is the learning rate, $\xi_i : \mathbb{R}_{\ge 0} \mapsto \mathbb{R}_{\ge 0}$ is the learning rate schedule for $\mathbf{w}_i$, and for $z = (x,y)$ and $i = L,\dots,2$, we define

$$\begin{aligned}\Delta_L^{\mathbf{H}}(z,1;\mathbf{W}(k)) &= \partial_2\mathcal{L}(y,\hat{\mathbf{y}}(x;\mathbf{W}(k)))\varphi_L'(\mathbf{H}_L(x,1;\mathbf{W}(k))),\\ \Delta_{i-1}^{\mathbf{H}}(z,j_{i-1};\mathbf{W}(k)) &= \frac{1}{n_i}\sum_{j_i=1}^{n_i}\Delta_i^{\mathbf{H}}(z,j_i;\mathbf{W}(k))\mathbf{w}_i(k,j_{i-1},j_i)\\ &\qquad\times\varphi_{i-1}'(\mathbf{H}_{i-1}(x,j_{i-1};\mathbf{W}(k))),\\ \Delta_i^{\mathbf{w}}(z,j_{i-1},j_i;\mathbf{W}(k)) &= \Delta_i^{\mathbf{H}}(z,j_i;\mathbf{W}(k))\varphi_{i-1}(\mathbf{H}_{i-1}(x,j_{i-1};\mathbf{W}(k))),\\ \Delta_1^{\mathbf{w}}(z,j_1;\mathbf{W}(k)) &= \Delta_1^{\mathbf{H}}(z,j_1;\mathbf{W}(k))x.\end{aligned}$$

In short, for an initialization $\mathbf{W}(0)$, we obtain a SGD trajectory $\mathbf{W}(k)$ of an L-layer network with size $\{n_i\}_{i\le L}$. We also note that this neural network fits into the framework in Section 2.

7.2. Mean field limit

Given a neuronal ensemble $(\Omega, P) = \prod_{i=1}^{L}(\Omega_i, P_i)$ (in which $\Omega_L = \{1\}$), the MF limit that is associated with the network (7.1) is described by the continuous-time

evolution of $W(t) = \{w_1(t,\cdot), w_i(t,\cdot,\cdot) : i = 2,\dots,L\}$, given by the following MF ODEs:

$$\frac{\partial}{\partial t} w_i(t, c_{i-1}, c_i) = -\xi_i(t)\mathbb{E}_Z[\Delta_i^w(Z, c_{i-1}, c_i; W(t))], \quad i = 2,\dots,L,$$
$$\frac{\partial}{\partial t} w_1(t, c_1) = -\xi_1(t)\mathbb{E}_Z[\Delta_1^w(Z, c_1; W(t))],$$

where $w_1 \colon \mathbb{R}_{\geq 0} \times \Omega_1 \to \mathbb{R}^d$ and $w_i \colon \mathbb{R}_{\geq 0} \times \Omega_{i-1} \times \Omega_i \to \mathbb{R}$. Here we define, for $i = L,\dots,2$, the forward quantities

$$\begin{aligned}
\hat{y}(x; W(t)) &= \varphi_L(H_L(x, 1; W(t))),\\
H_i(x, c_i; W(t)) &= \mathbb{E}_{C_{i-1}}[w_i(t, C_{i-1}, c_i)\varphi_{i-1}(H_{i-1}(x, C_{i-1}; W(t)))],\\
H_1(x, c_1; W(t)) &= \langle w_1(t, c_1), x\rangle,
\end{aligned}$$

and the backward quantities

$$\begin{aligned}
\Delta_L^H(z, 1; W(t)) &= \partial_2 \mathcal{L}(y, \hat{y}(x; W(t)))\varphi_L'(H_L(x, 1; W(t))),\\
\Delta_{i-1}^H(z, c_{i-1}; W(t)) &= \mathbb{E}_{C_i}[\Delta_i^H(z, C_i; W(t))w_i(t, c_{i-1}, C_i)\\
&\qquad \times \varphi_{i-1}'(H_{i-1}(x, c_{i-1}; W(t)))],\\
\Delta_i^w(z, c_{i-1}, c_i; W(t)) &= \Delta_i^H(z, c_i; W(t))\varphi_{i-1}(H_{i-1}(x, c_{i-1}; W(t))),\\
\Delta_1^w(z, c_1; W(t)) &= \Delta_1^H(z, c_1; W(t))x.
\end{aligned}$$

As a reminder, the data $Z = (X, Y) \sim \mathcal{P}$ and $C_i \sim P_i$. To recap, given a neuronal ensemble (Ω, P), for each initialization $W(0)$, we have defined a MF limit $W(t)$.

7.3. Global convergence and bidirectional diversity

We begin the study of global convergence of the network (7.1) with an analysis of its MF limit, which is the focus of this subsection. To measure the learning quality, we consider the loss averaged over the data $Z \sim \mathcal{P}$:

$$\mathcal{L}(F) = \mathbb{E}_Z[\mathcal{L}(Y, \hat{y}(X; F))],$$

where $F = \{f_i : i = 1,\dots,L\}$ a set of measurable functions $f_1 \colon \Omega_1 \to \mathbb{R}^d$, $f_i \colon \Omega_{i-1} \times \Omega_i \to \mathbb{R}$ for $i = 2,\dots,L$.

Recall that in our framework, the finite-sized neural network is formally connected with its MF limit via a neuronal embedding. Here without making explicit this connection, one can study the MF limit that is defined on the basis of a given neuronal embedding $(\Omega, P, \{w_i^0\}_{i\leq L})$, where $w_1^0 \colon \Omega_1 \to \mathbb{R}^d$, $w_i^0 \colon \Omega_{i-1} \times \Omega_i \to \mathbb{R}$ for $i = 2,\dots,L$. In particular, we make the following assumptions.

Assumption 7.1. Consider a neuronal embedding $(\Omega, P, \{w_i^0\}_{i\le L})$, recalling $\Omega = \prod_{i=1}^L \Omega_i$ and $P = \prod_{i=1}^L P_i$ with $\Omega_L = \{1\}$. Consider the MF limit associated with the neuronal ensemble (Ω, P) with initialization $W(0)$ such that $w_1(0,\cdot) = w_1^0(\cdot)$ and $w_i(0,\cdot,\cdot) = w_i^0(\cdot,\cdot)$. We make the following assumptions:

(1) *Initialization*: The functions $\{w_i^0\}_{i\le L}$ satisfy

$$\sup_{m\ge 1} \frac{1}{\sqrt{m}} \mathbb{E}[|w_1^0(C_1)|^m]^{1/m} \le K, \quad \sup_{m\ge 1} \frac{1}{\sqrt{m}} \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^m]^{1/m} \le K,$$

for $i = 2, \dots, L$.

(2) *Diversity*: The functions $\{w_i^0\}_{i\le L}$ satisfy

$$\begin{aligned} \operatorname{supp}(w_1^0(C_1), w_2^0(C_1,\cdot)) &= \mathbb{R}^d \times L^2(P_2), \\ \operatorname{supp}(w_i^0(\cdot, C_i), w_{i+1}^0(C_i,\cdot)) &= L^2(P_{i-1}) \times L^2(P_{i+1}) \end{aligned}$$

for $i = 2, \dots, L-1$. (Remark: we write $w_i^0(\cdot, C_i)$ to denote the random mapping $c \mapsto w_i^0(c, C_i)$, and similar for $w_{i+1}^0(C_i,\cdot)$.)

(3) *Regularity*: We assume the following:

- φ_i is K-bounded for $1 \le i \le L-1$, φ_i' is K-bounded and K-Lipschitz for $1 \le i \le L$, and φ_L' is non-zero everywhere,
- $\partial_2 \mathscr{L}(\cdot,\cdot)$ is K-Lipschitz in the second variable and K-bounded,
- $|X| \le K$ with probability 1,
- the learning rate schedule ξ_i is K-bounded and K-Lipschitz for $1 \le i \le L$.

(4) *Convergence*: There exist a coupling π_t of $\prod_{i=1}^L P_i$ and itself such that, for $i = 2, \dots, L$,

$$\begin{aligned} \mathbb{E}_{\pi_t}\Big[|w_i(t, C_{i-1}', C_i') - \bar{w}_i(C_{i-1}, C_i)| \prod\nolimits_{j=i+1}^L |\bar{w}_j(C_{j-1}, C_j)|\Big] \to 0, \\ \mathbb{E}_{\pi_t}\Big[|w_1(t, C_1') - \bar{w}_1(C_1)| \prod\nolimits_{j=2}^L |\bar{w}_j(C_{j-1}, C_j)|\Big] \to 0, \end{aligned}$$

as $t \to \infty$, where $(C_1, \dots, C_L, C_1', \dots, C_L') \sim \pi_t$. Furthermore,

$$\operatorname{ess-sup}\Big|\frac{\partial}{\partial t} w_L(t, C_{L-1}, 1)\Big| \to 0.$$

(Here we take $\prod_{j=i+1}^L = 1$ for $i = L$.)

(5) *Universal approximation*: The set $\{\varphi_1(\langle u,\cdot\rangle) : u \in \mathbb{R}^d\}$ has dense span in $L^2(\mathcal{P}_X)$ (the space of square integrable functions with respect to the measure $\mathcal{P}_X$, which is the distribution of the input X). Furthermore, for each $i = 2, \dots, L-1$, φ_i is non-obstructive in the sense that the set $\{\varphi_i \circ f : f \in L^2(\mathcal{P}_X)\}$ has dense span in $L^2(\mathcal{P}_X)$.

It is easy to see that this set of assumptions satisfies Assumptions 2.4–2.6 and 4.6. As a consequence, by Theorem 3.1, there exists a unique solution W to the MF ODEs on $t \in [0, \infty)$.

Theorem 7.2. *Consider a neuronal embedding* $(\Omega, P, \{w_i^0\}_{i \le L})$ *and the MF limit as in Assumption* 7.1. *Assume* $\xi_L(\cdot) = 1$.

- Case 1 (convex loss): *If* $\mathcal{L}$ *is convex in the second variable, then*

$$\lim_{t \to \infty} \mathcal{L}(W(t)) = \inf_F \mathcal{L}(F) = \inf_{\tilde{y}: \mathbb{R}^d \to \mathbb{R}} \mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))].$$

- Case 2 (generic non-negative loss): *Suppose* $\partial_2 \mathcal{L}(y, \hat{y}) = 0$ *implies* $\mathcal{L}(y, \hat{y}) = 0$. *If* $y = y(x)$ *is a function of* x, *then* $\mathcal{L}(W(t)) \to 0$ *as* $t \to \infty$.

The assumptions here are similar to those made in Theorems 6.2 and 6.6 of Section 6. Similar to the settings of Section 6, the regularity assumption can be satisfied for several common setups and loss functions; for example, this holds when φ_i is sigmoid or tanh for $i \le L - 1$, φ_L is the identity, and $\mathcal{L}$ is the Huber loss. The convergence assumption here is also similar to the convergence assumption in Assumption 6.1 or Assumption 6.5. In particular, the first part of the convergence assumption is essentially a Wasserstein-type convergence; it follows from the convergence of $(w_i(t))_{i=1}^L$ to $(\bar{w}_i)_{i=1}^L$ in an appropriate Wasserstein distance. The fifth assumption is again natural and can be satisfied by common activations. For example, φ_i can be tanh for $i = 1, \ldots, L - 1$. Indeed, whenever $\{\varphi_i(\langle u, \cdot \rangle) : u \in \mathbb{R}^d\}$ has dense span in $L^2(\mathcal{P}_X)$, φ_i is non-obstructive since

$$\text{span}(\{\varphi_i(\langle u, \cdot \rangle) : u \in \mathbb{R}^d\}) \subseteq \text{span}(\{\varphi_i \circ f : f \in L^2(\mathcal{P}_X)\}).$$

The diversity assumption is new. It refers to an initialization scheme that introduces correlation among the weights. In particular, i.i.d. initializations do not satisfy this assumption for $L \ge 3$.

The second assumption is the counterpart of the diversity assumption made in Theorems 6.2 and 6.6, but there is a special difference. In Section 6, the diversity assumption refers to a full support condition of only the first layer's initial weight, which is in the Euclidean space. Here our diversity assumption refers to a particular full support condition for all layers. At a closer look, the condition is in the function space and reflects certain *bidirectional diversity*. In particular, this assumption implies both $w_i^0(\cdot, C_i)$ and $w_i^0(C_{i-1}, \cdot)$ have full supports in $L^2(P_{i-1})$ and $L^2(P_i)$, respectively (which we shall refer to as *forward diversity* and *backward diversity*, respectively).

High-level idea of the proof. The proof proceeds with several insights that have already appeared in Section 6. The novelty of our present analysis lies in the use of the

aforementioned bidirectional diversity. To clarify the point, let us give a brief high-level idea of the proof. At time t sufficiently large, we expect to have

$$\begin{aligned}&\left|\frac{\partial}{\partial t} w_L(t, c_{L-1}, 1)\right| \\ &= |\mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \hat{y}(X; W(t)))\varphi_L'(H_L(X, 1; W(t)))\varphi_{L-1}(H_{L-1}(X, c_{L-1}; W(t)))]| \approx 0\end{aligned}$$

for P_{L-1}-almost every c_{L-1}. If the set of mappings $x \mapsto H_{L-1}(x, c_{L-1}; W(t))$, indexed by c_{L-1}, is diverse in the sense that $\operatorname{supp}(H_{L-1}(\cdot, C_{L-1}; W(t))) = L^2(\mathcal{P}_X)$, then, since φ_{L-1} is non-obstructive, we obtain

$$\mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \hat{y}(X; W(t))) \mid X = x]\varphi_L'(H_L(x, 1; W(t))) \approx 0,$$

and consequently

$$\mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \hat{y}(X; W(t))) \mid X = x] \approx 0$$

for $\mathcal{P}_X$-almost every x. The desired conclusion then follows.

Hence, the crux of the proof is to show that $\operatorname{supp}(H_{L-1}(\cdot, C_{L-1}; W(t))) = L^2(\mathcal{P}_X)$. In fact, we show that this holds for any finite time $t \geq 0$. This follows if we can prove the forward diversity property of the weights, in which $w_i(t, \cdot, C_i)$ has full support in $L^2(P_{i-1})$ for any $t \geq 0$ and $2 \leq i \leq L-1$, and a similar property for $w_1(t, C_1)$. Interestingly, to that end, we actually show that bidirectional diversity, and hence both forward diversity and backward diversity, holds at any time $t \geq 0$, even though we only need forward diversity for our purpose. The full proof is deferred to Section 7.5.

A converse for global convergence. Similar to Section 6, we also have a converse relation between global convergence and the essential supremum condition in Assumption 7.1 (4). The proof is presented in Appendix F.

Proposition 7.3. *Consider the MF limit corresponding to the network (7.1), such that they are coupled together by the coupling procedure in Section 4.1 with a neuronal embedding $(\Omega, P, \{w_i^0\}_{i \leq L})$. Suppose that the initialization and regularity assumptions (i.e., the first and third assumptions) of Assumption 7.1 hold, and that $\mathscr{L}(y, \hat{y}) \to \infty$ as $|\hat{y}| \to \infty$ for each y. Further assume $\xi_L(\cdot) = 1$. Then the following hold*:

- Case 1 (convex loss): *If $\mathscr{L}$ is convex in the second variable and $\mathcal{L}(W(t)) \to \inf_F \mathcal{L}(F)$ as $t \to \infty$, then it must be that*

$$\sup_{c_{L-1} \in \Omega_{L-1}} |\frac{\partial}{\partial t} w_L(t, C_{L-1}, 1)| \to 0 \quad \text{as } t \to \infty.$$

- Case 2 (generic non-negative loss): *Suppose $\partial_2 \mathscr{L}(y, \hat{y}) = 0$ implies $\mathscr{L}(y, \hat{y}) = 0$, and $y = y(x)$ is a function of x. If $\mathcal{L}(W(t)) \to 0$ as $t \to \infty$, then the same conclusion also holds.*

7.4. Connection to the network (7.1)

Theorem 7.2 concerns with the global convergence of the MF limit. To make the connection with a finite-width neural network (7.1), we recall the neuronal embedding $(\Omega, P, \{w_i^0\}_{i\le L})$, as well as the coupling procedure in Section 4.1. We, however, present a twist to the procedure, that is, we choose first the neuronal embedding $(\Omega, P, \{w_i^0\}_{i\le L})$ and then perform the following two steps:

(1) We form the MF limit $W(t)$ (for $t \in \mathbb{R}_{\ge 0}$) associated with the neuronal ensemble (Ω, P) by setting the initialization $W(0)$ to $w_1(0,\cdot) = w_1^0(\cdot)$, $w_i(0,\cdot,\cdot) = w_i^0(\cdot,\cdot)$ and running the MF ODEs.

(2) We independently sample $C_i(j_i) \sim P_i$ for $i = 1,\dots,L$ and $j_i = 1,\dots,n_i$. We then form the neural network initialization $\mathbf{W}(0)$ with $\mathbf{w}_1(0, j_1) = w_1^0(C_1(j_1))$ and $\mathbf{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))$ for $j_i \in [n_i]$. We obtain the network's trajectory $\mathbf{W}(k)$ for $k \in \mathbb{N}_{\ge 0}$ for the network (7.1), with the data $z(k)$ generated independently of $\{C_i(j_i)\}_{i\le L}$ and hence $\mathbf{W}(0)$.

That is, instead of starting with a given initialization law of $\mathbf{W}(0)$ as done in Section 4.1, here we first start with a chosen neuronal embedding. We then form the MF limit $W(t)$ and the neural network initialization $\mathbf{W}(0)$, and hence the dynamics $\mathbf{W}(k)$, based on this neuronal embedding. In other words, the initialization law of $\mathbf{W}(0)$ is deduced from the chosen neuronal embedding. Obviously this procedure ensures that $\bar{\eta}$-independence is satisfied (Assumption 4.4).

In summary, in the present context, the neuronal embedding forms the basis on which the finite-width neural network is realized. Furthermore, the neural network and its MF limit are coupled. Then, using Theorem 7.2 and Corollary 4.9, one can obtain the following result on the optimization efficiency of the neural network with SGD.

Corollary 7.4. *Consider the neural network (6.2) as described by the coupling procedure with the aforementioned twist. Under the same setting as Theorem 7.2, in Case 1,*

$$\lim_{t\to\infty} \lim_{\{n_i\}_{i\le L}} \lim_{\epsilon\to 0} \mathbb{E}_Z[\mathscr{L}(Y, \hat{\mathbf{y}}(X; \mathbf{W}(\lfloor t/\epsilon \rfloor)))] = \inf_F \mathcal{L}(F) = \inf_{\tilde{y}} \mathbb{E}_Z[\mathscr{L}(Y, \tilde{y}(X))]$$

in probability, where the limit of the widths is such that $n_{\min} \to \infty$ and $n_{\min}^{-c} \log n_{\max} \to 0$ for any $c > 0$, with $n_{\min} = \min\{n_i : 1 \le i \le L-1\}$ and $n_{\max} = \max\{n_i : 1 \le i \le L-1\}$. In Case 2, the same holds with the right-hand side being 0.

7.5. Proof of Theorem 7.2

Proof of Theorem 7.2. We divide the proof into several steps.

Step 1: Diversity of the weights. We will first show that $\mathrm{supp}(w_1(t, C_1)) = \mathbb{R}^d$ and $\mathrm{supp}(w_i(t,\cdot, C_i)) = L^2(P_{i-1})$ for $i = 2,\dots,L-1$ and for any $t \ge 0$. We do so

by showing a stronger statement, that the following bidirectional diversity condition holds at any finite training time:

$$\begin{aligned}\operatorname{supp}(w_1(t,C_1), w_2(t,C_1,\cdot)) &= \mathbb{R}^d \times L^2(P_2),\\ \operatorname{supp}(w_i(t,\cdot,C_i), w_{i+1}(t,C_i,\cdot)) &= L^2(P_{i-1}) \times L^2(P_{i+1}), \quad i = 2,\dots,L-1,\end{aligned}$$

for any $t \geq 0$.

We prove the first statement. Given a MF trajectory $(W(t))_{t\geq 0}$ and $u_1 \in \mathbb{R}^d$, $u_2 \in L^2(P_2)$, we consider the following flow on $\mathbb{R}^d \times L^2(P_2)$:

$$\begin{aligned}\frac{\partial}{\partial t} a_2^+(t,c_2;u) &= -\xi_2(t)\mathbb{E}_Z[\Delta_2^H(Z,c_2;W(t))\varphi_1(\langle a_1^+(t;u),X\rangle)],\\ \frac{\partial}{\partial t} a_1^+(t;u) &= -\xi_1(t)\mathbb{E}_Z\big[\mathbb{E}_{C_2}[\Delta_2^H(Z,C_2;W(t))a_2^+(t,C_2;u)]\\ &\qquad \times \varphi_1'(\langle a_1^+(t;u),X\rangle)X\big],\end{aligned} \tag{7.2}$$

for $u = (u_1,u_2)$, with the initialization $a_1^+(0;u) = u_1$ and $a_2^+(0,c_2;u) = u_2(c_2)$. Existence and uniqueness of (a_1^+, a_2^+) follows similarly to Theorem 3.1. We next prove, for all finite $T > 0$ and $u^+ = (u_1^+, u_2^+) \in \mathbb{R}^d \times L^2(P_2)$, that there exists $u^- = (u_1^-, u_2^-) \in \mathbb{R}^d \times L^2(P_2)$ such that

$$a_1^+(T;u^-) = u_1^+, \quad a_2^+(T,\cdot;u^-) = u_2^+.$$

We consider the following auxiliary dynamics on $\mathbb{R}^d \times L^2(P_2)$:

$$\begin{aligned}\frac{\partial}{\partial t} a_2^-(t,c_2;u) &= \xi_2(T-t)\mathbb{E}_Z[\Delta_2^H(Z,c_2;W(T-t))\varphi_1(\langle a_1^-(t;u),X\rangle)],\\ \frac{\partial}{\partial t} a_1^-(t;u) &= \xi_1(T-t)\mathbb{E}_Z\big[\mathbb{E}_{C_2}[\Delta_2^H(Z,C_2;W(T-t))a_2^-(t,C_2;u)]\\ &\qquad \times \varphi_1'(\langle a_1^-(t;u),X\rangle)X\big],\end{aligned}$$

initialized at $a_1^-(0;u) = u_1$ and $a_2^-(0,c_2;u) = u_2(c_2)$ for $u = (u_1,u_2) \in \mathbb{R}^d \times L^2(P_2)$. Existence and uniqueness of (a_1^-, a_2^-) follow similarly to Theorem 3.1. Observe that the pair

$$\tilde{a}_1^-(t) = a_1^-(T-t;u^+), \quad \tilde{a}_2^-(t,c_2) = a_2^-(T-t,c_2;u^+)$$

solves the system

$$\begin{aligned}\frac{\partial}{\partial t}\tilde{a}_2^-(t,c_2) &= -\frac{\partial}{\partial t} a_2^-(T-t,c_2;u^+)\\ &= -\xi_2(t)\mathbb{E}_Z[\Delta_2^H(Z,c_2;W(t))\varphi_1(\langle \tilde{a}_1^-(t),X\rangle)],\\ \frac{\partial}{\partial t}\tilde{a}_1^-(t) &= -\frac{\partial}{\partial t} a_1^-(T-t;u^+)\\ &= -\xi_1(t)\mathbb{E}_Z[\mathbb{E}_{C_2}[\Delta_2^H(Z,C_2;W(t))\tilde{a}_2^-(t,C_2)]\varphi_1'(\langle \tilde{a}_1^-(t),X\rangle)X],\end{aligned}$$

initialized at $\tilde{a}_2^-(0,c_2) = a_2^-(T,c_2;u^+)$ and $\tilde{a}_1^-(0) = a_1^-(T;u^+)$. Thus, by uniqueness of the solution to the ODE (7.2), $(\tilde{a}_1^-, \tilde{a}_2^-)$ forms a solution of the ODE (7.2) initialized at

$$\tilde{a}_1^-(0) = a_1^-(T;u^+), \quad \tilde{a}_2^-(0,c_2) = a_2^-(T,c_2;u^+).$$

In particular, the solution $(\tilde{a}_1^-, \tilde{a}_2^-)$ of the ODE (7.2) with this initialization satisfies

$$\tilde{a}_1^-(T) = a_1^-(0;u^+) = u_1^+, \quad \tilde{a}_2^-(T,\cdot) = a_2^-(0,\cdot;u^+) = u_2^+.$$

Let $u_1^- = a_1^-(T;u^+)$ and $u_2^- = a_2^-(T,\cdot;u^+)$. Then we have $a_1^+(T;u^-) = u_1^+$ and $a_2^+(T,\cdot;u^-) = u_2^+$, as desired.

Using this, by continuity of the map $u \mapsto (a_1^+(T;u), a_2^+(T,\cdot;u))$, for every $\epsilon > 0$, there exists a neighborhood U of u^- such that for any $u \in U$, we have that $|(a_1^+(T;u), a_2^+(T,\cdot;u)) - u^+| \le \epsilon$. Notice that the MF trajectory $W(t)$ satisfies

$$\begin{aligned} w_1(t,c_1) &= a_1^+(t; w_1(0,c_1), w_2(0,c_1,\cdot)), \\ w_2(t,c_1,\cdot) &= a_2^+(t,\cdot; w_1(0,c_1), w_2(0,c_1,\cdot)). \end{aligned}$$

Then, since $(w_1(0,C_1), w_2(0,C_1,\cdot))$ has full support in $\mathbb{R}^d \times L^2(P_2)$, for any finite $T > 0$, we have that $(w_1(T,C_1), w_2(T,C_1,\cdot))$ has full support in $\mathbb{R}^d \times L^2(P_2)$, proving the first statement.

The other statements can be proven similarly by considering the following pairs of flows on $L^2(P_{i-1}) \times L^2(P_{i+1})$, for $u = (u_1,u_2) \in L^2(P_{i-1}) \times L^2(P_{i+1})$:

$$\begin{aligned} &\frac{\partial}{\partial t} a_i^+(t,c_{i-1};u) \\ &\quad = -\xi_i(t)\mathbb{E}_Z[\Delta_i^a(Z, a_i^+(t,\cdot;u), a_{i+1}^+(t,\cdot;u); W(t))\varphi_{i-1}(H_{i-1}(X,c_{i-1};W(t)))], \\ &\frac{\partial}{\partial t} a_{i+1}^+(t,c_{i+1};u) \\ &\quad = -\xi_{i+1}(t)\mathbb{E}_Z[\Delta_{i+1}^H(Z,c_{i+1};W(t))\varphi_i(H_i^a(Z,a_i^+(t,\cdot;u);W(t)))], \end{aligned}$$

initialized at $a_i^+(0,c_{i-1};u) = u_1(c_{i-1})$ and $a_{i+1}^+(0,c_{i+1};u) = u_2(c_{i+1})$, and

$$\begin{aligned} \frac{\partial}{\partial t} a_i^-(t,c_{i-1};u) &= \xi_i(T-t)\mathbb{E}_Z\big[\Delta_i^a(Z, a_i^-(t,\cdot;u), a_{i+1}^-(t,\cdot;u); W(T-t)) \\ &\qquad \times \varphi_{i-1}(H_{i-1}(X,c_{i-1};W(T-t)))\big], \\ \frac{\partial}{\partial t} a_{i+1}^-(t,c_{i+1};u) &= \xi_{i+1}(T-t)\mathbb{E}_Z\big[\Delta_{i+1}^H(Z,c_{i+1};W(T-t)) \\ &\qquad \times \varphi_i(H_i^a(Z,a_i^-(t,\cdot;u);W(T-t)))\big], \end{aligned}$$

initialized at $a_i^-(0,c_{i-1};u) = u_1(c_{i-1})$ and $a_{i+1}^-(0,c_{i+1};u) = u_2(c_{i+1})$, in which we define, for $f \in L^2(P_{i-1})$ and $g \in L^2(P_{i+1})$,

$$\begin{aligned} \Delta_i^a(z,f,g;W(t)) &= \mathbb{E}_{C_{i+1}}[\Delta_{i+1}^H(z,C_{i+1};W(t))g(C_{i+1})\varphi_i'(H_i^a(z,f;W(t)))], \\ H_i^a(z,f;W(t)) &= \mathbb{E}_{C_{i-1}}[f(C_{i-1})\varphi_{i-1}(H_{i-1}(x,C_{i-1};W(t)))]. \end{aligned}$$

Step 2: Diversity of the pre-activations. We will show that $\operatorname{supp}(H_i(\cdot, C_i; W(t))) = L^2(\mathcal{P}_X)$ for any $t \geq 0$, for $i = 2, \ldots, L-1$ by induction.

Firstly, consider the base case $i = 2$. Recall that

$$H_2(x, c_2; W(t)) = \mathbb{E}_{C_1}[w_2(t, C_1, c_2)\varphi_1(\langle w_1(t, C_1), x\rangle)] \equiv \mathcal{H}_2(t, x, w_2(t, \cdot, c_2)).$$

Observe that the set $\operatorname{cl}(\{\mathcal{H}_2(t, \cdot, f) : f \in L^2(P_1)\})$ is a closed linear subspace of $L^2(\mathcal{P}_X)$. Hence, this set is equal to $L^2(\mathcal{P}_X)$ if it has dense span in $L^2(\mathcal{P}_X)$, which we show now. Indeed, suppose that for some $g \in L^2(\mathcal{P}_X)$ such that $|g| \neq 0$, we have $\mathbb{E}_Z[g(X)\mathcal{H}_2(t, X, f)] = 0$ for all $f \in L^2(P_1)$. Equivalently,

$$\mathbb{E}_{C_1}[f(C_1)\mathbb{E}_Z[g(X)\varphi_1(\langle w_1(t, C_1), X\rangle)]] = 0,$$

for all $f \in L^2(P_1)$. As such, for P_1-almost every c_1,

$$\mathbb{E}_Z[g(X)\varphi_1(\langle w_1(t, c_1), X\rangle)] = 0.$$

Since $\operatorname{supp}(w_1(t, C_1)) = \mathbb{R}^d$ and the mapping $u \mapsto \varphi_1(\langle u, x\rangle)$ is continuous, by the universal approximation assumption for φ_1, we obtain $g(x) = 0$ for P_X-almost every x, which is a contradiction. We have therefore proved that $\operatorname{cl}(\{\mathcal{H}_2(t, \cdot, f) : f \in L^2(P_1)\}) = L^2(\mathcal{P}_X)$. Note that, since $f \mapsto \mathcal{H}_2(t, x, f)$ is continuous, and $\operatorname{supp}(w_2(t, \cdot, C_2)) = L^2(P_1)$, we have $\operatorname{supp}(H_2(\cdot, C_2; W(t))) = L^2(\mathcal{P}_X)$, as desired.

Now let us assume that $\operatorname{supp}(H_{i-1}(\cdot, C_{i-1}; W(t))) = L^2(\mathcal{P}_X)$ for some $i \geq 3$ (the induction hypothesis). We would like to show $\operatorname{supp}(H_i(\cdot, C_i; W(t))) = L^2(\mathcal{P}_X)$. This is similar to the base case. In particular, recall that

$$\begin{aligned} H_i(x, c_i; W(t)) &= \mathbb{E}_{C_{i-1}}[w_i(t, C_{i-1}, c_i)\varphi_{i-1}(H_{i-1}(x, C_{i-1}; W(t)))] \\ &\equiv \mathcal{H}_i(t, x, w_i(t, \cdot, c_i)). \end{aligned}$$

Now suppose that for some $g \in L^2(\mathcal{P}_X)$, $|g| \neq 0$, we have $\mathbb{E}_Z[g(X)\mathcal{H}_i(t, X, f)] = 0$ for all $f \in L^2(P_{i-1})$. Then, for P_{i-1}-almost every c_{i-1},

$$\mathbb{E}_Z[g(X)\varphi_{i-1}(H_{i-1}(X, c_{i-1}; W(t)))] = 0.$$

Recall the induction hypothesis $\operatorname{supp}(H_{i-1}(\cdot, C_{i-1}; W(t))) = L^2(\mathcal{P}_X)$. Since φ_{i-1} is non-obstructive and continuous, we obtain $g(x) = 0$ for P_X-almost every x, which is a contradiction. Therefore, the set $\operatorname{cl}(\{\mathcal{H}_i(t, \cdot, f)\colon f \in L^2(P_{i-1})\})$ has dense span in $L^2(\mathcal{P}_X)$, and again, this implies it is equal to $L^2(\mathcal{P}_X)$. Since $f \mapsto \mathcal{H}_i(t, x, f)$ is continuous and $\operatorname{supp}(w_i(t, \cdot, C_i)) = L^2(P_{i-1})$, we have $\operatorname{supp}(H_i(\cdot, C_i; W(t))) = L^2(\mathcal{P}_X)$.

Step 3: Concluding. Let

$$\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(X; W(t))) \mid X = x]\varphi_L'(H_L(x, 1; W(t))) = \mathcal{H}(x, W(t)).$$

From the last step, we have $\text{supp}(H_{L-1}(\cdot, C_{L-1}; W(t))) = L^2(\mathcal{P}_X)$ for any $t \geq 0$. Recall that

$$\frac{\partial}{\partial t} w_L(t, c_{L-1}, 1) = -\mathbb{E}_Z[\mathcal{H}(X, W(t))\varphi_{L-1}(H_{L-1}(X, c_{L-1}; W(t)))].$$

By the convergence assumption, for any $\epsilon > 0$, there exists $T(\epsilon) > 0$ such that for any $t \geq T(\epsilon)$, and for P_{L-1}-almost every c_{L-1},

$$|\mathbb{E}_Z[\mathcal{H}(X, W(t))\varphi_{L-1}(H_{L-1}(X, c_{L-1}; W(t)))]| \leq \epsilon.$$

We claim that $\mathcal{H}(x, W(t)) \to \mathcal{H}(x, \{\bar{w}_i\}_{i \leq L})$ in $L^1(\mathcal{P}_X)$ as $t \to \infty$. Assuming this claim and recalling that φ_{L-1} is K-bounded by the regularity assumption, we then have that for some $T'(\epsilon) \geq T(\epsilon)$, and for any $t \geq T'(\epsilon)$,

$$\begin{aligned}
&\text{ess-sup}|\mathbb{E}_Z[\mathcal{H}(X, \{\bar{w}_i\}_{i \leq L})\varphi_{L-1}(H_{L-1}(X, C_{L-1}; W(t)))]| \\
&\quad \leq K\mathbb{E}_Z[|\mathcal{H}(X, \{\bar{w}_i\}_{i \leq L}) - \mathcal{H}(X, W(t))|] \\
&\qquad + \text{ess-sup}|\mathbb{E}_Z[\mathcal{H}(X, W(t))\varphi_{L-1}(H_{L-1}(X, C_{L-1}; W(t)))]| \leq K\epsilon.
\end{aligned}$$

Since $\text{supp}(H_{L-1}(\cdot, C_{L-1}; W(t))) = L^2(\mathcal{P}_X)$ and φ_{L-1} is continuous,

$$|\mathbb{E}_Z[\mathcal{H}(X, \{\bar{w}_i\}_{i \leq L})f(X)]| \leq K\epsilon \quad \text{for all } f \in S,$$

for $S = \{\varphi_{L-1} \circ g : g \in L^2(\mathcal{P}_X)\}$. Since $\epsilon > 0$ is arbitrary,

$$|\mathbb{E}_Z[\mathcal{H}(X, \{\bar{w}_i\}_{i \leq L})f(X)]| = 0 \quad \text{for all } f \in S.$$

Furthermore, since φ_{L-1} is non-obstructive, S has dense span in $L^2(\mathcal{P}_X)$. Therefore, $\mathcal{H}(x, \{\bar{w}_i\}_{i \leq L}) = 0$ for $\mathcal{P}_X$-almost every x. Since φ_L' is non-zero everywhere,

$$\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(X; \{\bar{w}_i\}_{i \leq L})) \mid X = x] = 0$$

for $\mathcal{P}_X$-almost every x.

In Case 1, due to convexity of $\mathcal{L}$, for any measurable function $\tilde{y}$,

$$\mathcal{L}(y, \tilde{y}(x)) - \mathcal{L}(y, \hat{y}(x; \{\bar{w}_i\}_{i \leq L})) \geq \partial_2 \mathcal{L}(y, \hat{y}(x, \{\bar{w}_i\}_{i \leq L}))(\tilde{y}(x) - \hat{y}(x, \{\bar{w}_i\}_{i \leq L})).$$

Taking expectation, we get $\mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))] \geq \mathcal{L}(\{\bar{w}_i\}_{i \leq L})$.

In Case 2, we have that $\partial_2 \mathcal{L}(y(x), \hat{y}(x; \{\bar{w}_i\}_{i \leq L})) = 0$, and therefore it follows that $\mathcal{L}(y(x), \hat{y}(x; \{\bar{w}_i\}_{i \leq L})) = 0$, for $\mathcal{P}_X$-almost every x, since y is a function of x.

This gives a result on $\mathcal{L}(\{\bar{w}_i\}_{i \leq L})$, conditional on the claim that $\mathcal{H}(x, W(t)) \to \mathcal{H}(x, \{\bar{w}_i\}_{i \leq L})$ in $L^1(\mathcal{P}_X)$ as $t \to \infty$. We now prove the claim. Recall the coupling π_t in Assumption 7.1.4. In the following, we let $(C_1, \ldots, C_L, C_1', \ldots, C_L') \sim \pi_t$. For brevity, we denote

$$\delta_i(t, x, c_i, c_i') = |H_i(x, c_i'; W(t)) - H_i(x, c_i; \{\bar{w}_i\}_{i \leq L})|.$$

First observe that by the regularity assumption, for $2 \le i \le L$,

$$\begin{aligned}\delta_i(t,x,c_i,c_i') \le K\mathbb{E}_{C_{i-1},C_{i-1}'}\big[&|w_i(t,C_{i-1}',c_i') - \bar{w}_i(C_{i-1},c_i)| \\ &+ |\bar{w}_i(C_{i-1},c_i)|\delta_{i-1}(t,x,C_{i-1},C_{i-1}')\big],\\ \delta_1(t,x,c_1,c_1') \le K|w_1(t,c_1') - \bar{w}_1(c_1)|.&\end{aligned}$$

This gives

$$\begin{aligned}\mathbb{E}_Z[|\mathcal{H}(X,W(t)) - \mathcal{H}(X,\{\bar{w}_i\}_{i\le L})|] &\le K\mathbb{E}_Z[\delta_L(t,X,1,1)]\\ &\le K^L\sum_{i=2}^{L}\mathbb{E}\Big[|w_i(t,C_{i-1}',C_i') - \bar{w}_i(C_{i-1},C_i)|\prod_{j=i+1}^{L}|\bar{w}_j(C_{j-1},C_j)|\Big]\\ &\quad + K^L\mathbb{E}\Big[|w_1(t,C_1') - \bar{w}_1(C_1)|\prod_{j=2}^{L}|\bar{w}_j(C_{j-1},C_j)|\Big].\end{aligned}$$

By the convergence assumption, the right-hand side tends to 0 as $t \to \infty$. This proves the claim.

Finally, let us connect $\mathcal{L}(W(t))$ with $\mathcal{L}(\{\bar{w}_i\}_{i\le L})$:

$$\begin{aligned}|\mathcal{L}(W(t)) - \mathcal{L}(\{\bar{w}_i\}_{i\le L})| &= |\mathbb{E}_Z[\mathcal{L}(Y,\hat{y}(X;W(t))) - \mathcal{L}(Y,\hat{y}(X;\{\bar{w}_i\}_{i\le L}))]|\\ &\le K\mathbb{E}_Z[|\hat{y}(X;W(t)) - \hat{y}(X;\{\bar{w}_i\}_{i\le L})|]\\ &\le K\mathbb{E}_Z[\delta_L(t,X,1,1)],\end{aligned}$$

which again tends to 0 as $t \to \infty$. This concludes the proof. ∎

8. Convergence to global optimum under Morse–Sard assumptions

In this section, we show that global convergence is guaranteed under a different set of convergence assumptions, namely, convergence in moments of the MF limit and certain Morse–Sard assumptions. This generalizes the global convergence mechanism of [9] for two-layer networks to settings where the loss $\mathcal{L}$ is not necessarily convex and the depth $L \ge 2$.

8.1. The two-layer case

Consider the two-layer setting of Section 6.1. We make the following assumption.

Assumption 8.1. There exist limits $\bar{w}_1$ and $\bar{w}_2$ such that the following hold:

(1) (*Wasserstein-type convergence*.) There exists a coupling π_t of P_1 and itself such that

$$\mathbb{E}_{\pi_t}[|\bar{w}_2(C_1)|\,|w_1(t,C_1') - \bar{w}_1(C_1)| + |w_2(t,C_1',1) - \bar{w}_2(C_1)|] \to 0$$

as $t \to \infty$, where $(C_1, C_1') \sim \pi_t$.

(2) (*Morse–Sard in the limit.*) With $\bar{W} = \{\bar{w}_1, \bar{w}_2\}$, the mapping

$$u_1 \mapsto \bar{\mathcal{F}}(u_1) := |\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(X; \bar{W}))\varphi_2'(H_2(X, 1; \bar{W}))\varphi_1(\langle u_1, X\rangle)]|^2$$

satisfies the following property. As $r \to \infty$, $\tilde{u}_1 \mapsto \bar{\mathcal{F}}(r\tilde{u}_1)$ converges uniformly in $C^1(\mathbb{S}^{d-1})$ to a function $\bar{\mathcal{F}}^\infty \colon \mathbb{S}^{d-1} \to \mathbb{R}$, where $\mathbb{S}^{d-1} = \{\tilde{u}_1 \in \mathbb{R}^d : |\tilde{u}_1| = 1\}$. Furthermore, for any stationary point u_1^* of $\bar{\mathcal{F}}$ with $\bar{\mathcal{F}}(u_1^*) > 0$, and for any $\delta_0 > 0$, there exists $\delta \in (0, \delta_0)$ so that for S_δ the connected component of the set $\{u : \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}(u_1^*) - \delta\}$ that contains u_1^*, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$, the boundary of the closure of S_δ. Similarly, for any stationary point $\tilde{u}_1^*$ of $\bar{\mathcal{F}}^\infty$ with $\bar{\mathcal{F}}^\infty(\tilde{u}_1^*) > 0$ and for any $\delta_0 > 0$, there exists $\delta \in (0, \delta_0)$ so that for $\tilde{S}_\delta$ the connected component of the set $\{\tilde{u} \in \mathbb{S}^{d-1} : \bar{\mathcal{F}}^\infty(\tilde{u}) > \bar{\mathcal{F}}^\infty(\tilde{u}_1^*) - \delta\}$ that contains $\tilde{u}_1^*$, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}^\infty(\tilde{u})| > \xi$ for all $\tilde{u} \in \partial \mathrm{cl}(\tilde{S}_\delta)$.

The convergence condition in the above assumption is actually the first part of Assumption 6.1 (4), and hence the remark for Assumption 6.1 (4) applies; i.e., one can deduce this condition from the convergence of $(w_1(t, \cdot), w_2(t, \cdot, 1))$ to $(\bar{w}_1, \bar{w}_2)$ as $t \to \infty$ in the Wasserstein-2 distance.

Theorem 8.2. *Consider the MF limit corresponding to the network* (6.1)*, such that they are coupled together by the coupling procedure in Section* 4.1*. Under Assumptions* 6.1 (1)–(3) *and* (5)*, and* 8.1*, the following hold*:

- Case 1 (convex loss): *If $\mathcal{L}$ is convex in the second variable, then*

$$\lim_{t\to\infty} \mathcal{L}(W(t)) = \inf_{f_1, f_2} \mathcal{L}(f_1, f_2) = \inf_{\tilde{y}} \mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))].$$

- Case 2 (generic non-negative loss): *Suppose $\partial_2 \mathcal{L}(y, \hat{y}) = 0$ implies $\mathcal{L}(y, \hat{y}) = 0$. If $y = y(x)$ is a function of x, then $\mathcal{L}(W(t)) = 0$ as $t \to \infty$.*

Let us make a comparison with the two-layer setting in Section 6.1, and in particular, Assumption 6.1. We see that the convergence assumption in Assumption 6.1 is replaced by Assumption 8.1. More specifically the uniform convergence condition of $\frac{\partial}{\partial t} w_2(t, C_1, 1)$ in Assumption 6.1 is replaced by the Morse–Sard condition of Assumption 8.1.

Similar to the proof of Theorem 6.2 in Section 6.1, the role of the Morse–Sard condition is – together with the full support property $\mathrm{supp}(\mathrm{Law}(w_1(t, C_1))) = \mathbb{R}^d$ by Lemma E.1 – to affirm that

$$\bar{\mathcal{F}}(u_1) = |\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(X; \bar{W}))\varphi_2'(H_2(X, 1; \bar{W}))\varphi_1(\langle u_1, X\rangle)]|^2 = 0$$

for all $u_1 \in \mathbb{R}^d$, after which universal approximation is invoked to yield the desired global convergence. The main idea is the following: should the above not hold, there

exists a region of u_1 where $\bar{\mathcal{F}}(u_1) > 0$. Since $\mathrm{supp}(\mathrm{Law}(w_1(t, C_1))) = \mathbb{R}^d$, at any time t, for a non-negligible mass on C_1, $w_1(t, C_1)$ fully occupies the region. The Morse–Sard condition ensures that the interaction over time between the two layers $w_1(t, C_1)$ and $w_2(t, C_1, 1)$ in this region would however force the dynamics to diverge so long as $\bar{\mathcal{F}}(u_1) > 0$.

The proof of Theorem 8.2 is deferred to Appendix G. We also refer to Section 9 for further discussions. In the following, we extend this result to the multilayer case and present its proof.

8.2. The multilayer case

One can obtain a multilayer analogue of Theorem 8.2. The key idea behind the Morse–Sard condition is similar to the two-layer case. Here the advantage of our framework becomes clearer since it easily accommodates the idea in the multilayer setup.

Recall the setting of Section 7. We make the following assumption which is a direct analogue of Assumption 8.1 in the two-layer case.

Assumption 8.3. There exist limits $\{\bar{w}_i\}_{i \le L}$ such that the following hold:

(1) (*Wasserstein-type convergence.*) There exist couplings π_t of $\prod_{i=1}^L P_i$ and itself such that

$$\mathbb{E}_{\pi_t}\Big[|w_1(t, C_1') - \bar{w}_1(C_1)|^2 \prod_{j=2}^L |\bar{w}_j(C_{j-1}, C_j)|^2\Big] \to 0,$$
$$\mathbb{E}_{\pi_t}\Big[|w_i(t, C_{i-1}', C_i') - \bar{w}_i(C_{i-1}, C_i)|^2 \prod_{j=i+1}^L |\bar{w}_j(C_{j-1}, C_j)|^2\Big] \to 0,$$

for $i = 2, \dots, L$ as $t \to \infty$, where $(C_1, \dots, C_L, C_1', \dots, C_L') \sim \pi_t$.

(2) (*Morse–Sard in the limit.*) With $\bar{W} = \{\bar{w}_i\}_{i \le L}$, the mapping

$$u_{L-1} \in L^2(P_{L-2}) \mapsto \bar{\mathcal{F}}(u_{L-1}),$$

defined by

$$\bar{\mathcal{F}}(u_{L-1}) := \Big|\mathbb{E}_Z\big[\partial_2 \mathcal{L}(Y, y(X; \bar{W}))\varphi_L'(H_L(X, 1; \bar{W})) \times \varphi_{L-1}(\langle u_{L-1}, H_{L-2}(X, \cdot; \bar{W})\rangle_{L^2(P_{L-2})})\big]\Big|^2,$$

satisfies the following property. As $r \to \infty$, $\tilde{u}_{L-1} \mapsto \bar{\mathcal{F}}(r\tilde{u}_{L-1})$ converges uniformly in $C^1(\mathbb{S}(L^2(P_{L-2})))$ to a function $\bar{\mathcal{F}}^\infty : \mathbb{S}(L^2(P_{L-2})) \to \mathbb{R}$, where $\mathbb{S}(L^2(P_{L-2})) = \{u \in L^2(P_{L-2}) : |u|_{L^2(P_{L-2})} = 1\}$. Furthermore, for any stationary point u_{L-1}^* of $\bar{\mathcal{F}}$ and for any $\delta_0 > 0$, there exists $\delta \in (0, \delta_0)$ so that for S_δ the connected component of the set $\{u : \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}(u_{L-1}^*) - \delta\}$ that contains u_{L-1}^*, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$, the boundary of the closure of S_δ. Similarly, for any stationary point $\tilde{u}_{L-1}^*$ of $\bar{\mathcal{F}}^\infty$

and for any $\delta_0 > 0$, there exists $\delta \in (0, \delta_0)$ so that for S_δ the connected component of the set $\{u : \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}^\infty(\tilde{u}^*_{L-1}) - \delta\}$ which contains $r\tilde{u}^*_{L-1}$ for all r sufficiently large, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$.

The convergence condition in the above assumption is actually the first part of Assumption 7.1 (4), and hence the remark for Assumption 7.1 (4) applies; i.e., one can deduce this condition from the convergence of $(w_i(t))_{i=1}^L$ to $(\bar{w}_i)_{i=1}^L$ in an appropriate Wasserstein distance. We now state the theorem. The proof is deferred to Section 8.3.

Theorem 8.4. *Consider a neuronal embedding* $(\Omega, P, \{w_i^0\}_{i \le L})$ *and the MF limit described in Section 7, and in particular, under Assumptions* 7.1 (1)–(3) *and* (5)*, and* 8.3*. Assume* $\xi_L(\cdot) = \xi_{L-1}(\cdot) = 1$.

- Case 1 (convex loss): *If* $\mathcal{L}$ *is convex in the second variable, then*

$$\lim_{t\to\infty} \mathcal{L}(W(t)) = \inf_F \mathcal{L}(F) = \inf_{\tilde{y}:\mathbb{R}^d\to\mathbb{R}} \mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))].$$

- Case 2 (generic non-negative loss): *Suppose* $\partial_2\mathcal{L}(y, \hat{y}) = 0$ *implies* $\mathcal{L}(y, \hat{y}) = 0$. *If* $y = y(x)$ *is a function of* x*, then* $\mathcal{L}(W(t)) \to 0$ *as* $t \to \infty$.

8.3. Proof of Theorem 8.4

Proof of Theorem 8.4. In the following, for $f, g \in L^2(P_{L-2})$, let us write $\langle f, g\rangle$ in place of $\langle f, g\rangle_{L^2(P_{L-2})}$ for brevity. Let us define

$$\bar{H}_i(x, C_i) = H_i(x, C_i; \bar{W}), \quad \bar{H}_L(x) = \bar{H}_L(x, 1), \quad \bar{y}(x) = \hat{y}(x; \bar{W}),$$

and for $u_{L-1} \in L^2(P_{L-2})$,

$$\begin{aligned}
\bar{G}_L(u_{L-1}) &= \mathbb{E}_Z[\partial_2\mathcal{L}(Y, \bar{y}(X))\varphi_L'(\bar{H}_L(X))\varphi_{L-1}(\langle u_{L-1}, \bar{H}_{L-2}(X, \cdot)\rangle)], \\
\bar{G}_{L-1}(u_{L-1}) &= \mathbb{E}_Z\big[\partial_2\mathcal{L}(Y, \bar{y}(X))\varphi_L'(\bar{H}_L(X)) \\
&\qquad \times \varphi_{L-1}'(\langle u_{L-1}, \bar{H}_{L-2}(X, \cdot)\rangle)\bar{H}_{L-2}(X, \cdot)\big], \\
G_L(t, u_{L-1}) &= \mathbb{E}_Z\big[\partial_2\mathcal{L}(Y, \hat{y}(t, X))\varphi_L'(H_L(t, X, 1)) \\
&\qquad \times \varphi_{L-1}(\langle u_{L-1}, H_{L-2}(t, X, \cdot)\rangle)\big], \\
G_{L-1}(t, u_{L-1}) &= \mathbb{E}_Z\big[\partial_2\mathcal{L}(Y, \hat{y}(t, X))\varphi_L'(H_L(t, X, 1)) \\
&\qquad \times \varphi_{L-1}'(\langle u_{L-1}, H_{L-2}(t, X, \cdot)\rangle)H_{L-2}(t, X, \cdot)\big].
\end{aligned}$$

Notice that $\bar{G}_{L-1}(u_{L-1}), G_{L-1}(t, u_{L-1}) \in L^2(P_{L-2})$, which follows easily from Assumptions 7.1 (3) and 8.3 (1) and Lemma 3.2. Similar to the proof of Theorem 8.2, with Assumptions 7.1 (3) and 8.3 (1), we obtain that, as $t \to \infty$,

$$\mathbb{E}[|\bar{H}_L(X) - H_L(t, X, 1)|^2] \to 0, \quad \mathbb{E}[|\bar{y}(X) - \hat{y}(t, X)|^2] \to 0,$$
$$\mathbb{E}[|\partial_2\mathcal{L}(Y, \hat{y}(t, X)) - \partial_2\mathcal{L}(Y, \bar{y}(X))|^2] \to 0,$$

and uniformly in u_{L-1},

$$|G_{L-1}(t,u_{L-1})-\bar{G}_{L-1}(u_{L-1})|^2\to 0,\quad |G_L(t,u_{L-1})-\bar{G}_L(u_{L-1})|^2\to 0.$$

Consider the limit potential given by

$$\bar{\mathcal{F}}(u_{L-1})=\frac{1}{2}|\bar{G}_L(u_{L-1})|^2.$$

By Assumption 7.1 (3), $u_{L-1}\mapsto\bar{\mathcal{F}}(u_{L-1})$ is continuous. Notice that

$$\nabla\bar{\mathcal{F}}(u_{L-1})=\frac{1}{2}\cdot 2\bar{G}_L(u_{L-1})\nabla(\bar{G}_L(u_{L-1}))=\bar{G}_L(u_{L-1})\bar{G}_{L-1}(u_{L-1}).$$

Let $\bar{\mathcal{F}}^\infty:\mathbb{S}(L^2(P_{L-2}))\to\mathbb{R}$ be defined by $\bar{\mathcal{F}}^\infty(\tilde{u}_{L-1})=\lim_{r\to\infty}\bar{\mathcal{F}}(r\tilde{u}_{L-1})$, which exists by Assumption 8.3. We shall argue that $\bar{\mathcal{F}}(u_{L-1})=0$ for all $u_{L-1}\in L^2(P_{L-2})$ by contradiction. To that end, let us assume that $\bar{\mathcal{F}}(u_{L-1})\neq 0$ for some u_{L-1}. Note that $\bar{\mathcal{F}}$ is bounded by a constant, by Assumption 7.1 (3). Thus, either there is a local maximizer u^*_{L-1} of $\bar{\mathcal{F}}$ with $\bar{\mathcal{F}}(u^*_{L-1})>0$ or there is a local maximizer $\tilde{u}^*_{L-1}$ of $\bar{\mathcal{F}}^\infty$ with $\bar{\mathcal{F}}^\infty(\tilde{u}^*_{L-1})>0$.

First consider the case that $\bar{\mathcal{F}}$ has a local maximizer u^*_{L-1} with $\bar{\mathcal{F}}(u^*_{L-1})>0$. Under Assumption 8.3, there exists $\delta\in(0,\bar{\mathcal{F}}(u^*_{L-1}))$ arbitrarily small so that for S_δ the connected component of the set $\{u:\bar{\mathcal{F}}(u)>\bar{\mathcal{F}}(u^*_{L-1})-\delta\}$ that contains u^*_{L-1}, there is $\xi>0$ such that $|\nabla\bar{\mathcal{F}}(u_{L-1})|>\xi$ for all $u_{L-1}\in\partial\mathrm{cl}(S_\delta)$. Let T_0 be sufficiently large so that for $t\geq T_0$, we have if $u_{L-1}\in\partial\mathrm{cl}(S_\delta)$, $|\bar{G}_{L-1}(u_{L-1})-G_{L-1}(t,u_{L-1})|\leq\xi/\sqrt{8\bar{\mathcal{F}}(u^*_{L-1})}$, which, similar to the proof of Theorem 8.2, implies

$$\langle\bar{G}_{L-1}(u_{L-1}),G_{L-1}(t,u_{L-1})\rangle>\frac{\xi^2}{4\bar{\mathcal{F}}(u^*_{L-1})}.\tag{8.1}$$

Also, we further enlarge T_0 so that $|\bar{G}_L(u_{L-1})-G_L(t,u_{L-1})|\leq\frac{1}{2}\sqrt{\bar{\mathcal{F}}(u^*_{L-1})-\delta}$ for $t\geq T_0$ and any $u_{L-1}\in\mathrm{cl}(S_\delta)$, and hence

$$\begin{aligned}G_L(t,u_{L-1})&\geq\bar{G}_L(u_{L-1})-\frac{1}{2}\sqrt{\bar{\mathcal{F}}(u^*_{L-1})-\delta}>\bar{G}_L(u_{L-1})-\frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_{L-1})}\\&=\bar{G}_L(u_{L-1})-\frac{1}{2}|\bar{G}_L(u_{L-1})|,\end{aligned}\tag{8.2}$$

$$\begin{aligned}G_L(t,u_{L-1})&\leq\bar{G}_L(u_{L-1})+\frac{1}{2}\sqrt{\bar{\mathcal{F}}(u^*_{L-1})-\delta}<\bar{G}_L(u_{L-1})+\frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_{L-1})}\\&=\bar{G}_L(u_{L-1})+\frac{1}{2}|\bar{G}_L(u_{L-1})|.\end{aligned}\tag{8.3}$$

Furthermore, notice that

$$\begin{aligned}&\frac{\partial}{\partial t}\bar{G}_L(w_{L-1}(t,\cdot,C_{L-1}))\\&\quad=-w_L(t,C_{L-1},1)\langle\bar{G}_{L-1}(w_{L-1}(t,\cdot,C_{L-1})),G_{L-1}(t,w_{L-1}(t,\cdot,C_{L-1}))\rangle.\end{aligned}\tag{8.4}$$

Let $\tilde{\Omega}_{L-1}$ be the subset of Ω_{L-1} consisting of c_{L-1}, where $|w_L(0, c_L, 1)| < 1$. As shown in Step 1 of the proof of Theorem 7.2,for any $t \geq 0$, we have

$$\text{supp}(w_{L-1}(t, \cdot, C_{L-1}), C_{L-1} \in \tilde{\Omega}_{L-1}) = L^2(P_{L-2}),$$

and hence for any open subset B of $L^2(P_{L-2})$, there exists a positive mass of $C_{L-1} \in \tilde{\Omega}_{L-1}$ such that $w_{L-1}(t, \cdot, C_{L-1}) \in B$. In the following, we consider $C_{L-1} \in \tilde{\Omega}_{L-1}$. We further divide the argument into two cases: $\bar{G}_L(u^*_{L-1}) > 0$ and $\bar{G}_L(u^*_{L-1}) < 0$.

Let us consider the case $\bar{G}_L(u^*_{L-1}) > 0$. Then we can choose sufficiently small δ such that $\bar{G}_L(u_{L-1}) > 0$ for all $u_{L-1} \in \text{cl}(S_\delta)$. Furthermore, consider the scenario that there exists $T \geq T_0$ such that a positive mass of $(w_{L-1}(T, \cdot, C_{L-1}), w_L(T, C_{L-1}, 1))$ on $C_{L-1} \in \tilde{\Omega}_{L-1}$ has $w_{L-1}(T, \cdot, C_{L-1}) \in S_\delta$ and $w_L(T, C_{L-1}, 1) < 0$. Note that, by equation (8.2), if $w_{L-1}(t, \cdot, C_{L-1}) \in S_\delta$, then

$$\begin{aligned}\frac{\partial}{\partial t} w_L(t, C_{L-1}, 1) &= -G_L(t, w_{L-1}(T, \cdot, C_{L-1})) \\ &\leq -\Big(\bar{G}_L(w_{L-1}(T, \cdot, C_{L-1})) - \frac{1}{2}|\bar{G}_L(w_{L-1}(T, \cdot, C_{L-1}))|\Big) < 0.\end{aligned}$$

Define $T_1 = \inf\{t \geq T : w_{L-1}(t, \cdot, C_{L-1}) \notin S_\delta\}$. Then $t \mapsto w_L(t, C_{L-1}, 1)$ is decreasing on $t \in [T, T_1)$. Let us argue that $T_1 = \infty$. Indeed, suppose T_1 is finite. We then have, by continuity, that $w_{L-1}(T_1, \cdot, C_{L-1}) \in \partial \text{cl}(S_\delta)$ and $w_L(T_1, C_{L-1}, 1) \leq w_L(T, C_{L-1}, 1) < 0$. As such, $\frac{\partial}{\partial t}\bar{G}_L(w_{L-1}(T_1, \cdot, C_{L-1})) > 0$, by equations (8.1) and (8.4). By continuity, for some $\gamma > 0$, we have $\frac{\partial}{\partial t}\bar{G}_L(w_{L-1}(T_1 + t, \cdot, C_{L-1})) > 0$ for all $t \in [0, \gamma]$. But then

$$\bar{G}_L(w_{L-1}(T_1 + t, \cdot, C_{L-1})) \geq \bar{G}_L(w_{L-1}(T_1, \cdot, C_{L-1})) \geq \sqrt{2(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)},$$

and hence $w_{L-1}(T_1 + t, \cdot, C_{L-1}) \in S_\delta$ for all $t \leq \gamma$, contradicting the definition of T_1. Therefore, $T_1 = \infty$, i.e., for $t \geq T$ and $C_1 \in \tilde{\Omega}_1$ with $w_{L-1}(T, \cdot, C_{L-1}) \in S_\delta$ and $w_L(T, C_{L-1}, 1) < 0$, we have $w_{L-1}(t, \cdot, C_{L-1}) \in S_\delta$, which implies

$$\begin{aligned}G_L(t, w_{L-1}(t, \cdot, C_{L-1})) &\geq \frac{1}{2}\bar{G}_L(w_{L-1}(t, \cdot, C_{L-1})) \\ &= \sqrt{\frac{1}{2}\bar{\mathcal{F}}(w_{L-1}(t, \cdot, C_{L-1}))} \geq \sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)},\end{aligned}$$

where the first inequality is by equation (8.2) and the fact $\bar{G}_L(u_{L-1}) > 0$ for all $u_{L-1} \in \text{cl}(S_\delta)$. In particular, there is a positive mass of $(w_{L-1}(T, \cdot, C_{L-1}), w_L(T, C_{L-1}, 1))$ with $G_L(t, w_{L-1}(t, \cdot, C_{L-1})) \geq \sqrt{(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)/2}$ for all $t \geq T$. Noting that

$$\frac{d}{dt}\mathbb{E}[\mathcal{L}(Y, \hat{y}(t, X))] \leq -\mathbb{E}[|G_L(t, w_{L-1}(t, \cdot, C_{L-1}))|^2],$$

we obtain $\frac{d}{dt}\mathbb{E}[\mathcal{L}(Y, \hat{y}(t, X))]$ being bounded above by a strictly negative constant for all $t \geq T$, which is a contradiction since $\mathcal{L}$ is bounded below.

Next we are considering the scenario that for all $t \geq T_0$, the probability that $w_{L-1}(t,\cdot,C_{L-1}) \in S_\delta$ and $w_L(t,C_{L-1},1) < 0$ on $C_{L-1} \in \tilde{\Omega}_{L-1}$ is zero. Let us argue that for any $t \geq T_0$ and for a.e. $C_{L-1} \in \tilde{\Omega}_{L-1}$ with $w_{L-1}(t,\cdot,C_{L-1}) \in S_\delta$, we have $w_{L-1}(s,\cdot,C_{L-1}) \in S_\delta$ for all $s \in [T_0,t]$. Indeed, consider t and $C_{L-1} \in \tilde{\Omega}_{L-1}$ such that $w_{L-1}(t,\cdot,C_{L-1}) \in S_\delta$ and $w_{L-1}(T',\cdot,C_{L-1}) \notin S_\delta$ for some $T' \in [T_0,t)$. Let $t' = \sup\{s \in [T',t] : w_{L-1}(s,\cdot,C_{L-1}) \notin S_\delta\} < t$. By continuity, $w_{L-1}(t',\cdot,C_{L-1}) \in \partial \mathrm{cl}(S_\delta)$ and so, by equation (8.1),

$$\langle \bar{G}_{L-1}(w_{L-1}(t',\cdot,C_{L-1})), G_{L-1}(t', w_{L-1}(t',\cdot,C_{L-1}))\rangle > \frac{\xi^2}{4\bar{\mathcal{F}}(u^*_{L-1})}.$$

By continuity, there exists $t'' \in (t',t)$ such that for all $s \in [t',t'']$,

$$\langle \bar{G}_{L-1}(w_{L-1}(s,\cdot,C_{L-1})), G_{L-1}(s, w_{L-1}(s,\cdot,C_{L-1}))\rangle \geq \frac{\xi^2}{100\bar{\mathcal{F}}(u^*_{L-1})}.$$

By definition of t', we also have $w_{L-1}(s,\cdot,C_{L-1}) \in S_\delta$ and thus $w_L(s,C_{L-1},1) \geq 0$ for any $s \in (t',t]$. Then, by equation (G.4), $\frac{\partial}{\partial t}\bar{G}_L(w_{L-1}(s,\cdot,C_{L-1})) \leq 0$ for all $s \in (t',t'']$ and therefore

$$\bar{G}_L(w_{L-1}(t'',\cdot,C_{L-1})) \leq \bar{G}_L(w_{L-1}(t',\cdot,C_{L-1})) = \sqrt{2(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)},$$

where the equality follows from $w_{L-1}(t',\cdot,C_{L-1}) \in \partial \mathrm{cl}(S_\delta)$. However, this contradicts with $w_{L-1}(t'',\cdot,C_{L-1}) \in S_\delta$. Therefore, it holds that for any $t \geq T_0$, for a.e. $C_{L-1} \in \tilde{\Omega}_{L-1}$ with $w_{L-1}(t,\cdot,C_{L-1}) \in S_\delta$, we have $w_{L-1}(s,\cdot,C_{L-1}) \in S_\delta$ and hence $w_L(s,C_{L-1},1) \geq 0$ for all $s \in [T_0,t]$. Since $w_{L-1}(t,\cdot,C_{L-1})$ on $C_{L-1} \in \tilde{\Omega}_{L-1}$ has full support at any $t \geq 0$, we have, for any $t_0 \geq T_0$, that there is a positive mass on $C_{L-1} \in \tilde{\Omega}_{L-1}$ such that $w_{L-1}(t_0,\cdot,C_{L-1}) \in S_\delta$ and hence, as shown, $w_{L-1}(s,\cdot,C_{L-1}) \in S_\delta$ and $w_L(s,C_{L-1},1) \geq 0$ for all $s \in [T_0,t_0]$. Note that we have $w_L(T_0,C_{L-1},1) \leq M(T_0)$ for some finite $M(T_0) > 0$ for $C_{L-1} \in \tilde{\Omega}_{L-1}$ (which follows from the fact that $|\frac{\partial}{\partial t} w_L(t,\cdot,1)| \leq K$, by Assumption 6.1 (3) and $w_L(0,C_{L-1},1) < 1$). Also note that for $w_{L-1}(s,\cdot,C_{L-1}) \in S_\delta$ and $s \geq T_0$,

$$\begin{aligned}\frac{\partial}{\partial t} w_L(s,C_{L-1},1) &= -G_L(s, w_{L-1}(s,\cdot,C_{L-1})) \leq -\frac{1}{2}\bar{G}_L(w_{L-1}(s,\cdot,C_{L-1})) \\ &= -\sqrt{\frac{1}{2}\bar{\mathcal{F}}(w_{L-1}(s,\cdot,C_{L-1}))} \leq -\sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)}\end{aligned}$$

a strictly negative constant, where the first inequality is by equation (8.2) and the fact that $\bar{G}_L(u_{L-1}) > 0$ for all $u_{L-1} \in \mathrm{cl}(S_\delta)$. As such, for any $t_0 \geq T_0$ such that

$$M(T_0) - (t_0 - T_0)\sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)} < 0,$$

there is a positive mass on $C_{L-1} \in \tilde{\Omega}_{L-1}$ such that firstly $w_L(s, C_{L-1}, 1) \geq 0$ for all $s \in [T_0, t_0]$ and secondly there exists $t \in [T_0, t_0]$ in which

$$w_L(t, C_{L-1}, 1) \leq M(T_0) - (t - T_0)\sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u^*_{L-1}) - \delta)} < 0.$$

We again obtain a contradiction.

The case $\bar{G}_L(u^*_{L-1}) < 0$ can be treated similarly, with the use of equation (8.2) replaced by equation (8.3). Both cases lead to a contradiction, ruling out the possibility that there is a local maximizer u^*_{L-1} of $\bar{\mathcal{F}}$ with $\bar{\mathcal{F}}(u^*_{L-1}) > 0$.

Next consider the case where $\bar{\mathcal{F}}$ does not have any local maximizer in $L^2(P_{L-2})$ but $\bar{\mathcal{F}}^\infty$ has a local maximizer $\tilde{u}^*_{L-1}$ with $\bar{\mathcal{F}}^\infty(\tilde{u}^*_{L-1}) > 0$. Under Assumption 8.3, there exists $\delta \in (0, \bar{\mathcal{F}}^\infty(\tilde{u}^*_{L-1}))$ arbitrarily small so that for S_δ the connected component of the set $\{u : \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}^\infty(\tilde{u}^*_{L-1}) - \delta\}$, which contains $r\tilde{u}^*_{L-1}$ for all r sufficiently large, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$. The rest of the argument can be repeated as before to yield a contradiction.

In short, we have shown that $\bar{\mathcal{F}}(u_{L-1}) = \frac{1}{2}|\bar{G}_L(u_{L-1})|^2 = 0$, and equivalently,

$$\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \bar{y}(X))\varphi'_L(\bar{H}_L(X))\varphi_{L-1}(\langle u_{L-1}, \bar{H}_{L-2}(X, \cdot)\rangle)] = 0$$

for all $u_{L-1} \in L^2(P_{L-2})$. The proof can now be completed similar to the proof of Theorem 7.2. ∎

9. Further discussions

Having presented our neuronal embedding framework for multilayer MF neural networks and proven several results concerning i.i.d. initializations and global convergence under various settings, we now place the discussion of our work in the context of related works.

9.1. Two-layer neural networks

The MF view on the training dynamics of neural networks has gathered significant interests in the recent literature, starting with the two-layer case [9, 22, 26, 32, 34]. In this case, it is known that convergence to global optimum is possible for gradient descent or SGD [9, 22, 32], with a potentially exponential rate [18] and a dimension-independent width [21]. This line of works has also inspired research into new training algorithms [27, 31, 38], stability properties of the trained networks [33], other architectures which are compositions of multiple MF neural networks [12, 20] and MF neural networks in other machine learning contexts [1, 25]. Most works focus on fully-connected networks on the Euclidean space and utilize certain convexity properties

to study optimization efficiency. The MF formulation of the two-layer case in these works enjoys the wealth of the mathematics of optimal transport and gradient flows in measure spaces [3].

Our work, on the other hand, considers general Hilbert spaces which can be infinite-dimensional (Section 2) and does not rely critically on convexity (see Theorems 6.2, 6.6, 7.2, 8.2 and 8.4). Our framework departs from the Wasserstein gradient flow viewpoint, and while being in the early stage of technical foundations, it is demonstrated to give useful results including and beyond the two-layer case.

9.2. Multilayer neural networks

As mentioned in the introduction, the multilayer case poses a major conceptual challenge. Prior to our work, several ideas have been proposed independently. The work in [23] puts forth the idea that a neuron is represented by a stochastic (Markov) kernel and gives a heuristic derivation, where the MF limit is described by a certain evolution of measures over the space of stochastic kernels. The work in [4] rigorously derives the MF limit as an evolution of a measure on paths through layers. In [35], the network is viewed as a time-dependent function of its initialization and this function simplifies upon concentrations over the randomness of the initialization. All three works employ scalings with respect to the widths, in which normalizations are applied at every layers, not just the last layer, together with compensating learning rates. This thereby ensures nonlinear evolution at all layers.

Working under the same scalings, our framework gives a new perspective via a central question: how does one describe an ensemble of an arbitrary number of neurons? Answering this question, our idea of a neuronal embedding allows one to describe the MF limit in a clean and rigorous manner. In particular, it avoids extra assumptions made in [4,35]: unlike our work, [4] assumes untrained first and last layers and requires non-trivial technical tools; [35] takes an unnatural sequential limit of the widths and proves a non-quantitative result, whereas we prove a quantitative bound that essentially requires only the minimum of the widths to be large. An advantage of our framework comes from the fact that while MF formulations in [4,35] are specific to and exploit i.i.d. initializations, our formulation does not, and thereby allows to study i.i.d. initializations as well as interesting non-i.i.d. initialization schemes. Compared to [23], while a certain step of our analysis takes an inspiration from the idea of stochastic kernels in [23], our framework circumvents its technical cumbersomeness and gives a rigorous and clean mathematical treatment.

After our first preprint, the work [14] takes another view on this challenge. In particular, considering a finite set of training data, [14] encodes each neuron by its pre-activation values, computed over the entire training data, at initialization. As a specific interpretation by [14], the pre-activation values at initialization capture a

certain sense of "features" seen by the neurons. Meanwhile our framework identifies neuron j_i at layer i via the sample $C_i(j_i)$ drawn from the space (Ω_i, P_i) (Section 4.1) and remains general about this space. One may observe the following connection: a specific choice of the neuronal embedding can be built over random variables that are defined by the pre-activation values at initialization. The generality of our framework maintains freedom over choices of the neuronal embedding, including this specific choice. For example, when the training data size is infinite, an idealized situation commonly assumed in theoretical studies, then if one follows [14], each pre-activation becomes a function over an infinite domain, instead of a finite-dimensional vector. This potentially poses technical complications, which can be avoided simply by a different choice of the neuronal embedding in our framework.

9.3. Degeneracy with i.i.d. initializations

As shown in Section 5, i.i.d. initializations cause strong degeneracy for a network depth at least four. The work [4] is the first to realize and take advantage of this phenomenon to formulate the MF limit; in particular, the measure on the paths in [4] admits a product structure, signifying the mutually independent nature of the evolutions of weights at different layers in the infinite width limit. Note, however, that [4] explicitly exploits this degeneracy phenomenon to formulate the MF limit. In contrast, our framework is general and upon specializing to the case of i.i.d. initializations, it allows to derive this phenomenon in greater details and simultaneously remove certain technical assumptions in [4]. In particular, we remove the technical conditions of random input and output features and no biases of [4]. In addition, one can use Corollary 5.4 to immediately verify that in the setting of no biases, $L \geq 5$ and untrained first and last layers ($\xi_1^{\mathbf{w}}(\cdot) = \xi_L^{\mathbf{w}}(\cdot) = 0$), the weights and activations in the limit satisfy the McKean–Vlasov equation in [4].

Such degeneracy is generally undesirable. The fact that our framework is not specific to i.i.d. initializations allows for an escape from this situation. In this aspect, our framework follows closely the spirit of the work [23], whose MF formulation is also not specific to i.i.d. initializations. Through the language of stochastic kernels, [23] envisions a scenario in which evolutions of the weights at different layers are stochastically coupled. The usefulness of such scenario is realized by our global convergence guarantee for multilayer networks with arbitrary depths in Sections 7 and 8 (Theorems 7.2 and 8.4), with the novel idea of bidirectional diversity for non-i.i.d. initialization.

9.4. Global convergence

Optimization efficacy has been one major question that sets the MF literature apart from other theoretical studies of neural networks, where one witnesses new involve-

ment of sophisticated mathematical tools and insights. As mentioned, the two-layer case has enjoyed numerous efforts to establish global convergence (see, e.g., [7,9,18,22,31,32,38,39]). Our work is the first to obtain global convergence guarantees in the MF regime for the multilayer case.

Two-layer networks: *comparison to* [9]. Closely relevant to our thread of results is the work [9]. This work treats the two-layer case under certain convergence and Morse–Sard assumptions and convex losses. To make a direct comparison with [9], let us first focus on the two-layer case, and in particular, Theorem 6.2 together with its accompanying Assumption 6.1. Several elements in our analysis are inspired by this work; we also differ in crucial ways. Similar to [9], our proof also hinges on the insight that a certain diversity property is held throughout the course of training. We assume a universal approximation property (Assumption 6.1 (5)), which is natural in neural network learning, and dispense with convexity of the loss, whereas [9] does not utilize universal approximation and requires convex losses. In our convergence assumption (Assumption 6.1 (4)) the moment convergence condition is similar to the convergence assumption in [9]. We differ from [9] fundamentally in the uniform convergence condition ess-sup $|\frac{\partial}{\partial t} w_2(t, C_1, 1)| \to 0$ of the second layer's weight. On one hand, this condition replaces the Morse–Sard condition in [9], which is difficult to verify in general. On the other hand, it is a natural assumption to make: as shown in Proposition 6.4, if this uniform convergence condition fails, global convergence cannot be attained. In short, using the insight on diversity, together with universal approximation, we uncover a new mechanism for global convergence without the need for convex losses.

Multilayer networks. While [9] is specific to two-layer networks, we further the insight on diversity to the multilayer case, where we introduce the new notion of bidirectional diversity. In the context of two-layer networks, diversity refers to that the first layer's weight distribution has full support in the Euclidean space. In the multilayer case, this notion no longer resides in the Euclidean space, but it is realized in function spaces that are naturally described by the neuronal embedding framework. Moreover, as noted in Section 7.3, it highlights an interesting dynamical mechanism, in which adjacent layers interact with each other over time in such a way that diversity is preserved through the depth of the network and at any time, roughly speaking.

Similar to the two-layer case, in place of the Morse–Sard assumption in [9], we show global convergence under uniform convergence of the gradient update at a certain layer (Theorems 6.6 and 7.2). Again we note per Propositions 6.8 and 7.3, there is a converse relation between this uniform convergence and global convergence; if the former fails, so does the latter.

Several of these insights are utilized in the recent work [14], which proves a global convergence guarantee for a residual MF neural architecture under the uniform

convergence assumption of the gradient update. In this architecture, a skip connection is introduced to route the first layer directly to the second last one. Thanks to this skip connection, diversity is essentially transferred directly from the first layer to the second last layer. In short, in [14], diversity is maintained with the help of architectural imposition. In contrast, in our global convergence result for the multilayer case, diversity is maintained automatically by the training dynamics.

The work [20], which studies a type of composition of many two-layer MF networks, and a recent update of [35], which studies the three-layer case, establish conditions of stationary points to be global optima with certain overlapping ideas. However, they require essentially a certain diversity assumption on the limit point (i.e., at convergence $t = \infty$). We do not need to make this assumption: the remark in Section 6.2.1 highlights the dynamical nature of the proof where diversity is assumed at initialization $t = 0$ only and proven to hold at any finite training time $t < \infty$. As explained in Section 6.2.1, diversity may not hold at $t = \infty$ and global convergence can still be attained regardless.

Let us mention again that global convergence results in those works are proven under the convex loss assumption. On the other hand, our results allow for removal of this assumption and our proofs do not make use of convexity in any crucial way.

Convergence under Morse–Sard assumptions. Our framework is able to give a self-contained proof of global convergence under the Morse–Sard assumption, without the aforementioned uniform convergence assumption (Theorems 8.2 and 8.4).

Let us place this discussion in the two-layer context, particularly Theorem 8.2 and its accompanying Assumption 8.1. Observe that Assumption 8.1 (2) follows immediately if $\bar{\mathcal{F}}$ and $\bar{\mathcal{F}}^\infty$ satisfy Morse–Sard type regularity, i.e., the sets of regular values of $\bar{\mathcal{F}}$ and $\bar{\mathcal{F}}^\infty$ are dense (hence the name "Morse–Sard"). Indeed, assume that $\bar{\mathcal{F}}$ and $\bar{\mathcal{F}}^\infty$ satisfy Morse–Sard type regularity. Let $\hat{S}_\delta = \{u : \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}(u_1^*) - \delta\}$. In that case, for any stationary point u_1^* of $\bar{\mathcal{F}}$ with $\bar{\mathcal{F}}(u_1^*) > 0$, and for any $\delta_0 > 0$, there exists $\delta \in (0, \delta_0)$ so that any $u \in \partial \mathrm{cl}(\hat{S}_\delta)$ satisfies $\nabla \bar{\mathcal{F}}(u) \neq 0$. Over a bounded connected component S_δ of $\hat{S}_\delta$, this immediately implies the existence of $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$. Over an unbounded connected component S_δ of $\hat{S}_\delta$, whenever $\bar{\mathcal{F}}(u_1^*) - \delta$ is a regular value of $\bar{\mathcal{F}}^\infty$, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for $u \in \partial \mathrm{cl}(S_\delta) \setminus \mathbb{B}(0, r)$ for some r sufficiently large, where $\mathbb{B}(0, r)$ is the ball around 0 with radius r. Since $\bar{\mathcal{F}}(u_1^*) - \delta$ is a regular value of $\bar{\mathcal{F}}$, by making $\xi > 0$ smaller if needed, we can guarantee that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for $u \in \partial \mathrm{cl}(S_\delta) \cap \mathbb{B}(0, r)$, and hence $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$.

Thus our Morse–Sard condition is similar to (and slightly weaker than) the Morse–Sard assumption of [9]. As stated, it is sufficient for this condition to hold with respect to the limit $\bar{W} = \{\bar{w}_1, \bar{w}_2\}$. A counterpart statement of the Morse–Sard assumption

of [9] would impose the condition on a generic class of pairs of functions $\tilde{W} = \{\tilde{w}_1, \tilde{w}_2\}$ that contains $\bar{W}$ and as such trivially imply our assumption.

As explained in Section 8.1, the Morse–Sard condition forces the interaction over time between the weights of the two layers in a specific way that guarantees global convergence. This idea was realized by [9] in the language of Wasserstein gradient flows for a convex loss function $\mathcal{L}$ and two-layer neural networks. Here in the two-layer case, firstly Theorem 8.2 extends the result to generic losses; secondly and more importantly, it demonstrates that the same idea could be naturally executed in our framework without the use of Wasserstein gradient flows.

Theorem 8.4 demonstrates further the applicability of our argument to the multilayer case, which the Wasserstein gradient flow formulation has difficulty with.

9.5. Empirical findings and other infinite-width scalings

Mathematical ideas aside, one important aspect is how well one can observe the MF limiting behavior in multilayer networks with finite but large widths, normalized under the MF scaling. This has been demonstrated positively in the work [23]. In particular, [23] performs experiments on several real-life machine learning tasks and finds that the evolution curves of certain performance metrics, such as the training loss and the classification accuracy, are almost insensitive to the widths – provided sufficiently large – and hence they exhibit a limiting behavior. As [23] shows, this occurs as soon as the widths are on the order of just a few hundreds, which is common in practice.

The MF scaling is not the only infinite-width scaling with interesting properties. Another popular scaling regime is the neural tangent kernel (NTK) scaling [2, 8, 11, 17, 19, 41]. In the NTK scaling, the weights do not move and the learning dynamics becomes linearized, although several interesting properties such as convergence to the global optimum are attainable. For this reason, it is often said that the NTK-scaled infinite-width neural networks do not perform feature learning. This NTK-like behavior is not what is observed in practical neural networks with finite but large widths. In contrast, the MF-scaled networks have nonlinear dynamics and weights moving away from initialization, and thus said to perform feature learning in the literature.

The MF scaling is not necessarily the only scaling with feature learning (see, e.g., [16, 40]). It is known that in the standard scaling that matches with the usual practice, the networks are NTK-like in the infinite-width limit [21, 40]. Consequently, any infinite-width scalings with feature learning are only proxies of practical finite-width neural networks. Despite this fact, we note that [23] demonstrates on several real-life machine learning tasks that a MF multilayer network, without heavy hyperparameter tuning, can achieve realistic performances, comparable to practical neural networks that are similar in architectural designs and training procedures; [20] demonstrates an

improved performance over strong and well-tuned practical neural networks by using the MF scaling. In other words, the MF scaling offers a good proxy, with potentially no loss in practical performances.

A few alternative scalings, accompanied by suitable initializations and learning rates, are proposed in [16,40] to avoid the NTK-like behavior. Theoretical understanding of feature learning in these scalings is currently limited to just a single SGD step, unlike our work which studies the full learning trajectory of MF networks and proves the presence of meaningful learning via global convergence. As said, all these scalings are proxies of practical finite-width networks. Furthermore, it is argued in [21] that for two-layer infinite-width networks that are close to practical networks, the behavior near initialization is more NTK-like, while that in the long-time horizon is more MF-like. We expect a similar situation for the multilayer case, in which case it is insufficient to understand neural networks by analyzing only a few initial SGD steps. Our work also demonstrates the goodness of well-designed non-i.i.d. initializations, which thus far have been under-explored in the literature.

In a later follow-up work [29], our neuronal embedding framework is extended to study a finite-width correction to the infinite-width MF limit and the implicit bias of gradient descent training in this finite-width regime, hence paving the path to address the aforementioned limitation of the infinite-width viewpoint.

A. Useful tools

We state a martingale concentration result, which is a special case of [30, Theorem 3.2] and applies to a more general Banach space.

Theorem A.1 (Concentration of martingales in separable Hilbert spaces.). *Consider a martingale $Z_n \in \mathbb{Z}$, a separable Hilbert space, adapted to $\mathcal{F}_n$, such that $|Z_n - Z_{n-1}| \le R$ and $Z_0 = 0$. Then, for any $t > 0$,*

$$\mathbb{P}\big(\max_{k\le n} |Z_k| \ge t\big) \le 2 \inf_{\lambda>0} \exp\Big(-\lambda t + \operatorname{ess\text{-}sup} \sum_{k=1}^{n} \mathbb{E}[e^{\lambda|Z_k - Z_{k-1}|} - 1 - \lambda|Z_k - Z_{k-1}| \mid \mathcal{F}_{k-1}]\Big).$$

In particular, for $t < nR$,

$$\mathbb{P}\big(\max_{k\le n} |Z_k| \ge t\big) \le 2\exp\Big(-\frac{t^2}{16nR^2}\Big).$$

Proof. The first part follows from [30, Theorem 3.2]. The second part follows from the observation that for $\lambda < 1/(2R)$,

$$\mathbb{E}[e^{\lambda|Z_k - Z_{k-1}|} - 1 - \lambda|Z_k - Z_{k-1}| \mid \mathcal{F}_{k-1}] < 4\lambda^2 R^2,$$

and as such we have for $t < nR$,

$$\mathbb{P}\Big(\max_{k\le n}|Z_k| \ge t\Big) \le 2 \inf_{0<\lambda<1/(2R)} \exp(-\lambda t + 4n\lambda^2 R^2) \le 2\exp\Big(-\frac{t^2}{16nR^2}\Big). \quad \blacksquare$$

Next we state two results for η-independent random variables in separable Hilbert spaces.

Theorem A.2 (Concentration of η-independent bounded sum in separable Hilbert spaces). *Consider n η-independent random variables $X_1, \dots, X_n$ in a separable Hilbert space, where $\eta \in [0,1]$. Suppose that $|X_i - \mathbb{E}[X_i]| \le R$ almost surely. Then, for $\delta > 2\eta R$, we have*

$$\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i - \mathbb{E}[X_i]\Big| \ge \delta\Big) \le 2\exp\Big(-\frac{n\delta^2}{64R^2}\Big).$$

Proof. Since $|X_i - \mathbb{E}[X_i]| \le R$, the claims are immediate for $\delta \ge R$. Let

$$Z_i = X_1 - \mathbb{E}[X_1] + X_2 - \mathbb{E}[X_2 \mid X_1] + \cdots + X_i - \mathbb{E}[X_i \mid X_1, \dots, X_{i-1}].$$

Then Z_i is a martingale adapted to $\mathcal{F}_i = \sigma(X_1, \dots, X_{i-1})$. By Theorem A.1,

$$\mathbb{P}(|Z_n| \ge t) \le 2\exp\Big(-\frac{t^2}{16nR^2}\Big),$$

assuming that $t < nR$. Using the η-independence property, we have that

$$\begin{aligned}\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i - \mathbb{E}[X_i]\Big| \ge \delta n\Big) &\le \mathbb{P}(|Z_n| \ge \delta n - n\eta R) \\ &\le 2\exp\Big(-\frac{(\delta-\eta R)^2 n}{16R^2}\Big) \le 2\exp\Big(-\frac{\delta^2 n}{64R^2}\Big).\end{aligned}$$

for $\delta \in (2\eta R, (1+\eta)R)$. $\blacksquare$

Theorem A.3 (Moments of η-independent heavy-tailed sum in separable Hilbert spaces). *Consider $(X_1, X_2, \dots)$ being η-independent random variables in a separable Hilbert space. Suppose that for some constant $k \ge 1$ (with $k \le K$), for any $i \in \mathbb{N}_{>0}$,*

$$\sup_{m\ge 1} m^{-k/2}\mathbb{E}[|X_i|^m]^{1/m} \le K.$$

Then, for $m \ge 1$,

$$\mathbb{E}\Big[\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i]\Big|^m\Big]^{1/m} \le Km^{1+k/2}\max(\eta n^{0.01}, n^{-1/2}).$$

Proof. It is easy to see that it suffices to prove the claim for $m \ge 2$. Let us define $\mathcal{F}_i = \sigma(X_1, \dots, X_{i-1})$ and

$$Z_i = X_1 - \mathbb{E}[X_1] + X_2 - \mathbb{E}[X_2 \mid \mathcal{F}_2] + \cdots + X_i - \mathbb{E}[X_i \mid \mathcal{F}_i].$$

Then Z_i is a martingale adapted to $\mathcal{F}_i$. Note that for any $m \geq 1$ and $B > 0$:

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{i=1}^n X_i - \mathbb{E}[X_i]\Big|^m\Big]^{1/m} \\
&\quad \leq \mathbb{E}[|Z_n|^m]^{1/m} + \mathbb{E}\Big[\Big|\sum_{i=1}^n \mathbb{E}[X_i \mid \mathcal{F}_i] - \mathbb{E}[X_i]\Big|^m\Big]^{1/m} \\
&\quad \leq \mathbb{E}[|Z_n|^m]^{1/m} + \mathbb{E}\Big[\Big|\sum_{i=1}^n \mathbb{E}[X_i \mathbb{I}_{|X_i|\geq B} \mid \mathcal{F}_i] - \mathbb{E}[X_i \mathbb{I}_{|X_i|\geq B}]\Big|^m\Big]^{1/m} + n\eta B \\
&\quad \leq \mathbb{E}[|Z_n|^m]^{1/m} + n^{1-1/m}\Big(\sum_{i=1}^n \mathbb{E}[|X_i|^m \mathbb{I}_{|X_i|\geq B}]\Big)^{1/m} + n\eta B.
\end{aligned}$$

By [30, Theorem 4.1], for $m \geq 2$,

$$\begin{aligned}
&\mathbb{E}[|Z_n|^m]^{1/m} \\
&\quad \leq Km\Big(\sum_{i=1}^n \mathbb{E}[|X_i|^m]\Big)^{1/m} + K\sqrt{m}\,\mathbb{E}\Big[\Big(\sum_{i=1}^n \mathbb{E}[|X_i - \mathbb{E}[X_i \mid \mathcal{F}_i]|^2 \mid \mathcal{F}_i]\Big)^{m/2}\Big]^{1/m} \\
&\quad \leq Km\Big(\sum_{i=1}^n \mathbb{E}[|X_i|^m]\Big)^{1/m} + K\sqrt{m}\,\mathbb{E}\Big[\Big(\sum_{i=1}^n \mathbb{E}[|X_i|^2 \mid \mathcal{F}_i]\Big)^{m/2}\Big]^{1/m} \\
&\quad \leq K(m + \sqrt{m}n^{1/2-1/m})\Big(\sum_{i=1}^n \mathbb{E}[|X_i|^m]\Big)^{1/m} \\
&\quad \leq K(mn^{1/m} + \sqrt{m}n^{1/2})m^{k/2} \leq Km^{1+k/2}n^{1/2}.
\end{aligned}$$

We also have

$$\mathbb{E}[|X_i|^m \mathbb{I}_{|X_i|\geq B}] \leq \mathbb{E}[|X_i|^{2m}]^{1/2}\mathbb{P}(|X_i| \geq B)^{1/2} \leq K^m m^{mk/2}\exp(-KB^{2/k}),$$

since $|X_i|^{1/k}$ is K-sub-Gaussian. Therefore,

$$\mathbb{E}\Big[\Big|\sum_{i=1}^n X_i - \mathbb{E}[X_i]\Big|^m\Big]^{1/m} \leq Km^{1+k/2}n^{1/2} + Km^{k/2}\exp\Big(-\frac{KB^{2/k}}{m}\Big)n + n\eta B.$$

The claim is satisfied for $B^{2/k} = cm\log n$ with a suitable constant c. ∎

B. Remaining proofs for Section 3

B.1. Proof of Lemma 3.2

Proof of Lemma 3.2. Let $\eta(i) = 2^{L-i}$ and $\bar{\eta}(i) = 2^{i-1}$. Let us define

$$[\![w_i]\!]_{m,t} = \sqrt{\frac{50}{m}}\,\mathbb{E}\big[\sup_{s\leq t}|w_i(s, C_{i-1}, C_i)|^m\big]^{1/m}, \quad i = 2, \ldots, L.$$

We prove the following by backward induction, for $i = 2, \dots, L$ and any $m \geq 50$:

$$\sqrt{\frac{50}{m}}\,\mathbb{E}\Big[\sup_{s\leq t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(s,Z,C_i)|^m\Big]^{1/m} \leq K^{3\eta(i-1)-2}(1+t^{\eta(i)-1})\prod_{j=i+1}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i)}+|\!|\!|b_j|\!|\!|_0^{\bar\eta(j-i)}),$$

$$[\![w_i]\!]_{m,t} \leq K^{3\eta(i-1)-1}(1+[\![w_i]\!]_{m,0})(1+t^{\eta(i)})\prod_{j=i+1}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i)}+|\!|\!|b_j|\!|\!|_0^{\bar\eta(j-i)}),$$

$$|\!|\!|b_i|\!|\!|_t \leq K^{3\eta(i-1)-1}(1+|\!|\!|b_i|\!|\!|_0)(1+t^{\eta(i)})\prod_{j=i+1}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i)}+|\!|\!|b_j|\!|\!|_0^{\bar\eta(j-i)}),$$

for some immaterial constant $K \geq 1$, where, by standard convention, $\prod_{j=i+1}^{L} = 1$ if $i = L$.

Let us start with $i = L$. By Assumption 2.6, for $\mathcal{P}$-almost every z,

$$\sup_{t\geq 0}|\Delta_L^H(t,z,1)| \leq K.$$

Consequently, for $\mathcal{P}$-almost every z,

$$\max\Big(\sup_{t\geq 0}\sup_{c_{L-1}\in\Omega_{L-1}}|\Delta_L^w(t,z,c_{L-1},1)|,\ \sup_{t\geq 0}|\Delta_L^b(t,z,1)|\Big) \leq K\big(1+\sup_{t\geq 0}|\Delta_L^H(t,z,1)|\big) \leq K^2.$$

Together with Assumption 2.4 and the fact that w_L and b_L satisfy the MF ODEs, this implies

$$[\![w_L]\!]_{m,t} \leq [\![w_L]\!]_{m,0} + K^2 t, \quad |\!|\!|b_L|\!|\!|_t \leq |\!|\!|b_L|\!|\!|_0 + K^2 t.$$

These prove the statement for $i = L$.

Next, assuming the statement for $i+1$, we prove the statement for i, where $1 < i < L$. Using Cauchy–Schwarz's inequality, we have from Assumption 2.6, for $m \geq 50$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{s\leq t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(s,Z,C_i)|^m\Big] \\
&\leq K^m\mathbb{E}\Big[\sup_{s\leq t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\big|\mathbb{E}_{C_{i+1}}\big[(1+|\Delta_{i+1}^H(s,Z,C_{i+1})|) \\
&\quad\times(1+|w_{i+1}(s,C_i,C_{i+1})|+|b_{i+1}(s,C_{i+1})|)\big]\big|^m\Big] \\
&\leq K^m\mathbb{E}\Big[\mathbb{E}_{C_{i+1}}\big[1+\sup_{s\leq t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(s,Z,C_{i+1})|^2\big]^{m/2} \\
&\quad\times\mathbb{E}_{C_{i+1}}\big[1+\sup_{s\leq t}|w_{i+1}(s,C_i,C_{i+1})|^2+\sup_{s\leq t}|b_{i+1}(s,C_{i+1})|^2\big]^{m/2}\Big]
\end{aligned}$$

$$
\begin{aligned}
&\le K^m\big(1+\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(s,Z,C_{i+1})|^2\big]^{m/2}\big)\\
&\quad\times\mathbb{E}\Big[1+\mathbb{E}_{C_{i+1}}\big[\sup_{s\le t}|w_{i+1}(s,C_i,C_{i+1})|^m\big]+\mathbb{E}\big[\sup_{s\le t}|b_{i+1}(s,C_{i+1})|^2\big]^{m/2}\Big]\\
&\le K^m\big(1+\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(s,Z,C_{i+1})|^2\big]^{m/2}\big)\\
&\quad\times(1+m^{m/2}[\![w_{i+1}]\!]_{m,t}^m+|||b_{i+1}|||_t^m),
\end{aligned}
$$

which implies, by the induction hypothesis,

$$
\begin{aligned}
&\sqrt{\frac{50}{m}}\,\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(s,Z,C_i)|^m\big]^{1/m}\\
&\le K\big(1+\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(s,Z,C_{i+1})|^2\big]^{1/2}\big)(1+[\![w_{i+1}]\!]_{m,t}+|||b_{i+1}|||_t)\\
&\le K\Big[1+K^{3\eta(i)-2}(1+t^{\eta(i+1)-1})\prod_{j=i+2}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i-1)}+|||b_j|||_0^{\bar\eta(j-i-1)})\Big]\\
&\quad\times\Big[1+K^{3\eta(i)-1}(1+[\![w_{i+1}]\!]_{m,0}+|||b_{i+1}|||_0)(1+t^{\eta(i+1)})\\
&\quad\times\prod_{j=i+2}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i-1)}+|||b_j|||_0^{\bar\eta(j-i-1)})\Big]\\
&\le K^{3\eta(i-1)-2}(1+t^{\eta(i)-1})\prod_{j=i+1}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i)}+|||b_j|||_0^{\bar\eta(j-i)}).
\end{aligned}
$$

Therefore, by Assumptions 2.6 and 2.4, with the fact that w_i satisfies the MF ODEs, we have

$$
\begin{aligned}
[\![w_i]\!]_{m,t}&=\sqrt{\frac{50}{m}}\,\mathbb{E}\Big[\sup_{s\le t}\Big|w_i(0,C_{i-1},C_i)\\
&\qquad-\int_0^s\xi_i^{\mathbf{w}}(s')\mathbb{E}_Z[\Delta_i^w(s',Z,C_{i-1},C_i)]\,ds'\Big|^m\Big]^{1/m}\\
&\le[\![w_i]\!]_{m,0}+\frac{K}{\sqrt{m}}\,\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^w(s,Z,C_{i-1},C_i)|^m\big]^{1/m}t\\
&\le[\![w_i]\!]_{m,0}+K\Big(1+\frac{1}{\sqrt{m}}\,\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(s,Z,C_i)|^m\big]^{1/m}\Big)t\\
&\le K^{3\eta(i-1)-1}(1+[\![w_i]\!]_{m,0})(1+t^{\eta(i)})\prod_{j=i+1}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-i)}+|||b_j|||_0^{\bar\eta(j-i)}).
\end{aligned}
$$

We obtain a similar bound for $|||b_i|||_t$. This completes the backward induction. With

the same argument, one can obtain a similar bound for $i = 1$:

$$\sqrt{\frac{50}{m}}\,\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_1^H(s,Z,C_1)|^m\Big]^{1/m} \le K^{3\eta(0)-2}(1+t^{\eta(1)-1})\prod_{j=2}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-1)}+|\!|\!|b_j|\!|\!|_0^{\bar\eta(j-1)}),$$

$$|\!|\!|w_1|\!|\!|_t \le K^{3\eta(0)-1}(1+|\!|\!|w_1|\!|\!|_0)(1+t^{\eta(1)})\prod_{j=2}^{L}(1+[\![w_j]\!]_{m,0}^{\bar\eta(j-1)}+|\!|\!|b_j|\!|\!|_0^{\bar\eta(j-1)}).$$

By taking the supremum on m or setting $m = 50$, these bounds imply the claimed bound on $[\![W]\!]_{\psi,t}$ and $|\!|\!|W|\!|\!|_t$. In addition, from the bounds on $[\![w_i]\!]_{m,t}$, it follows that $\sup_{s\le t}|w_i(s,C_{i-1},C_i)|$ is $K_0(t)$-sub-Gaussian for $2 \le i \le L$. Together with the union bound, we then get the claimed probability bound. ∎

B.2. Proof of Lemma 3.4

We state the following two useful auxiliary lemmas.

Lemma B.1. *Consider two collections of MF parameters $W', W'' \in \mathcal{W}_T$. Under Assumption 2.5, for any $t \le T$ and $1 \le i \le L$, the following hold*:

$$\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}}|H_i(X,C_i;W'(s))-H_i(X,C_i;W''(s))|^2\Big]^{1/2} \le K^L K_0^L(T)\|W'-W''\|_t,$$

$$\sup_{s\le t}\operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}}|\hat y(X;W'(s))-\hat y(X;W''(s))| \le K^L K_0^L(T)\|W'-W''\|_t.$$

Lemma B.2. *Given $B \ge 0$, consider two collections of MF parameters $W', W'' \in \mathcal{W}_T$ such that*

$$\mathbb{P}(\mathsf{max}_T^w(W') \ge K_0(T)B) \le 2Le^{1-K_1B^2},$$

$$\mathbb{P}(\mathsf{max}_T^w(W'') \ge K_0(T)B) \le 2Le^{1-K_1B^2}.$$

Under Assumptions 2.5 and 2.6, for any $t \le T$ and $2 \le i \le L$, the following hold:

$$\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^w(Z,C_{i-1},C_i;W'(s))-\Delta_i^w(Z,C_{i-1},C_i;W''(s))|^2\Big]^{1/2} \le D(t,W',W''),$$

$$\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^b(Z,C_i;W'(s))-\Delta_i^b(Z,C_i;W''(s))|^2\Big]^{1/2} \le D(t,W',W''),$$

$$\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_1^w(Z,C_1;W'(s))-\Delta_1^w(Z,C_1;W''(s))|^2\Big]^{1/2} \le D(t,W',W''),$$

in which $D(t,W',W'') := (KK_0(T))^{2L+2}((1+B)\|W'-W''\|_t + \sqrt{L}e^{-K_1B^2/2})$.

These lemmas lay the foundation to prove Lemma 3.4.

Proof of Lemma 3.4. Let us recall the quantity $D(t, W', W'')$ from Lemma B.2. Let us note a simple identity:

$$\begin{aligned}\mathbb{E}_C\Big[\Big(\int_0^t f(s,C)\,ds\Big)^2\Big] &= \int_0^t\int_0^t \mathbb{E}_C[f(s_1,C)f(s_2,C)]\,ds_1\,ds_2\\ &\le \int_0^t\int_0^t \mathbb{E}_C[|f(s_1,C)|^2]^{1/2}\mathbb{E}_C[|f(s_2,C)|^2]^{1/2}\,ds_1\,ds_2\\ &= \Big(\int_0^t \mathbb{E}_C[|f(s,C)|^2]^{1/2}\,ds\Big)^2.\end{aligned}$$

Now, for any $i \ge 2$,

$$\begin{aligned}&\mathbb{E}\Big[\Big(\int_0^t \Big|\frac{\partial}{\partial t}F_i^w(W')(s,C_{i-1},C_i) - \frac{\partial}{\partial t}F_i^w(W'')(s,C_{i-1},C_i)\Big|\,ds\Big)^2\Big]^{1/2}\\ &= \mathbb{E}\Big[\Big(\int_0^t \big|\xi_i^{\mathbf{w}}(s)\mathbb{E}_Z[\Delta_i^w(Z,C_{i-1},C_i;W'(s)) - \Delta_i^w(Z,C_{i-1},C_i;W''(s))]\big|\,ds\Big)^2\Big]^{1/2}\\ &\overset{(a)}{\le} K\mathbb{E}\Big[\Big(\int_0^t \big|\mathbb{E}_Z[\Delta_i^w(Z,C_{i-1},C_i;W'(s)) - \Delta_i^w(Z,C_{i-1},C_i;W''(s))]\big|\,ds\Big)^2\Big]^{1/2}\\ &\overset{(b)}{\le} K\int_0^t \mathbb{E}\big[\big|\mathbb{E}_Z[\Delta_i^w(Z,C_{i-1},C_i;W'(s)) - \Delta_i^w(Z,C_{i-1},C_i;W''(s))]\big|^2\big]^{1/2}\,ds\\ &\overset{(c)}{\le} K\int_0^t D(s,W',W'')\,ds,\end{aligned}$$

where (a) is due to Assumption 2.4, (b) is by the aforementioned identity, and (c) is an application of Lemma B.2. Therefore,

$$\begin{aligned}&\mathbb{E}\big[\sup_{s\le t}|F_i^w(W')(s,C_{i-1},C_i) - F_i^w(W'')(s,C_{i-1},C_i)|^2\big]^{1/2}\\ &\quad\le \mathbb{E}\Big[\sup_{s\le t}\Big(\int_0^s \Big|\frac{\partial}{\partial t}F_i^w(W')(s',C_{i-1},C_i) - \frac{\partial}{\partial t}F_i^w(W'')(s',C_{i-1},C_i)\Big|\,ds'\Big)^2\Big]^{1/2}\\ &\quad\le K\int_0^t D(s,W',W'')\,ds.\end{aligned}$$

One can show the same bound for F_i^b and F_1^w. This completes the proof. ∎

Lemmas B.1 and B.2 are in fact special cases of the following lemmas.

Lemma B.3. *Consider two collections of MF parameters $W', W'' \in \mathcal{W}_T$. Suppose that we define $\tilde{C}_1,\dots,\tilde{C}_L$ independent random variables on $\Omega_1,\dots,\Omega_L$, such that $\tilde{C}_i$ is independent of $C_1,\dots,C_{i-1},C_{i+1},\dots,C_L$, and that there exists some $K_*(T) \ge K_0(T)$ such that all following quantities are upper-bounded by $K_*(T)$ for all $t \le T$*

and for $W = W'$ or $W = W''$:

$$\max_{2\le i\le L}\mathbb{E}\big[\sup_{s\le t}|w_i(s,\tilde{C}_{i-1},\tilde{C}_i)|^{50}\big]^{1/50},\ \max_{2\le i\le L}\mathbb{E}\big[\sup_{s\le t}|w_i(s,C_{i-1},\tilde{C}_i)|^{50}\big]^{1/50},$$
$$\max_{2\le i\le L}\mathbb{E}\big[\sup_{s\le t}|w_i(s,\tilde{C}_{i-1},C_i)|^{50}\big]^{1/50},\ \max_{2\le i\le L}\mathbb{E}\big[\sup_{s\le t}|w_i(s,C_{i-1},C_i)|^{50}\big]^{1/50},$$
$$\max_{2\le i\le L}\mathbb{E}\big[\sup_{s\le t}|b_i(s,\tilde{C}_i)|^{50}\big]^{1/50},\ \max_{2\le i\le L}\mathbb{E}\big[\sup_{s\le t}|b_i(s,C_i)|^{50}\big]^{1/50},$$
$$\mathbb{E}\big[\sup_{s\le t}|w_1(s,\tilde{C}_1)|^{50}\big]^{1/50},\mathbb{E}\big[\sup_{s\le t}|w_1(s,C_1)|^{50}\big]^{1/50}.$$

Under Assumption 2.5, for any $t \le T$ and $1 \le i \le L$, we have

$$\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}}|H_i(X,\tilde{C}_i;W'(s))-H_i(X,\tilde{C}_i;W''(s))|^2\big]^{1/2}\le K^L K_*^L(T)\tilde{d}_t(W',W''),$$

and the same holds if we replace $\tilde{C}_i$ with C_i in the left-hand side of the above. Here we have defined the metrics:

$$\begin{aligned}
\tilde{d}_t(W',W'') &= \max\big(\max_{2\le i\le L}\tilde{d}_t(w_i',w_i''),\max_{2\le i\le L}\tilde{d}_t(b_i',b_i''),\tilde{d}_t(w_1',w_1'')\big),\\
\tilde{d}_t(w_i',w_i'') &= \max\Big(\mathbb{E}\big[\sup_{s\le t}|w_i'(s,\tilde{C}_{i-1},\tilde{C}_i)-w_i''(s,\tilde{C}_{i-1},\tilde{C}_i)|^2\big]^{1/2},\\
&\qquad \mathbb{E}\big[\sup_{s\le t}|w_i'(s,C_{i-1},\tilde{C}_i)-w_i''(s,C_{i-1},\tilde{C}_i)|^2\big]^{1/2},\\
&\qquad \mathbb{E}\big[\sup_{s\le t}|w_i'(s,\tilde{C}_{i-1},C_i)-w_i''(s,\tilde{C}_{i-1},C_i)|^2\big]^{1/2},\\
&\qquad \mathbb{E}\big[\sup_{s\le t}|w_i'(s,C_{i-1},C_i)-w_i''(s,C_{i-1},C_i)|^2\big]^{1/2}\Big),\\
\tilde{d}_t(b_i',b_i'') &= \max\Big(\mathbb{E}\big[\sup_{s\le t}|b_i'(s,\tilde{C}_i)-b_i''(s,\tilde{C}_i)|^2\big]^{1/2},\\
&\qquad \mathbb{E}\big[\sup_{s\le t}|b_i'(s,C_i)-b_i''(s,C_i)|^2\big]^{1/2}\Big),\\
\tilde{d}_t(w_1',w_1'') &= \max\Big(\mathbb{E}\big[\sup_{s\le t}|w_1'(s,\tilde{C}_1)-w_1''(s,\tilde{C}_1)|^2\big]^{1/2},\\
&\qquad \mathbb{E}\big[\sup_{s\le t}|w_1'(s,C_1)-w_1''(s,C_1)|^2\big]^{1/2}\Big).
\end{aligned}$$

(*Note that the random variables $\tilde{C}_i$ are general, and may be chosen to be equal to C_i. The space $\mathcal{W}_T$ which contains W' and W'' is defined with respect to the random variables $C_1,\dots,C_L$.*)

Lemma B.4. *Consider two collections of MF parameters $W', W'' \in \mathcal{W}_T$. Suppose we define the random variables $\tilde{C}_1,\dots,\tilde{C}_L$, the bounding constant $K_*(T)$ and the metric $\tilde{d}_t(W',W'')$ as given in the statement of Lemma B.3. Further assume that for*

some non-negative function Ξ *and some* $B \geq 0$,

$$\mathbb{P}(\widetilde{\max}_T^w(W') \geq K_*(T)B) \leq \Xi(B),$$
$$\mathbb{P}(\widetilde{\max}_T^w(W'') \geq K_*(T)B) \leq \Xi(B),$$

in which we define

$$\widetilde{\max}_t^w(W') = \max_{2\leq i\leq L} \sup_{s\leq t} \max\big(|w_i'(s, C_{i-1}, \tilde{C}_i)|, |w_i'(s, \tilde{C}_{i-1}, C_i)|, |w_i'(s, \tilde{C}_{i-1}, \tilde{C}_i)|, |w_i'(s, C_{i-1}, C_i)|\big).$$

Under Assumptions 2.5 and 2.6, for any $t \leq T$ *and* $2 \leq i \leq L$, *we have*

$$\mathbb{E}\big[\sup_{s\leq t} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\Delta_i^w(Z, \tilde{C}_{i-1}, \tilde{C}_i; W'(s)) - \Delta_i^w(Z, \tilde{C}_{i-1}, \tilde{C}_i; W''(s))|^2\big]^{1/2} \leq \tilde{D}(t, W', W''),$$
$$\mathbb{E}\big[\sup_{s\leq t} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\Delta_i^b(Z, \tilde{C}_i; W'(s)) - \Delta_i^b(Z, \tilde{C}_i; W''(s))|^2\big]^{1/2} \leq \tilde{D}(t, W', W''),$$
$$\mathbb{E}\big[\sup_{s\leq t} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\Delta_1^w(Z, \tilde{C}_1; W'(s)) - \Delta_1^w(Z, \tilde{C}_1; W''(s))|^2\big]^{1/2} \leq \tilde{D}(t, W', W''),$$

and the same holds if we replace $\tilde{C}_i$ *or* $\tilde{C}_{i-1}$ *with* C_i *or* C_{i-1}, *respectively, in the left-hand side of each line above. Here*

$$\tilde{D}(t, W', W'') := (KK_*(T))^{3L+2}((1 + B)\tilde{d}_t(W', W'') + \sqrt{\Xi(B)}).$$

Next we prove each of the remaining lemmas.

B.3. Proof of Lemmas B.1 and B.3

Proof of Lemma B.1. The first bound is a direct corollary of Lemma B.3 by setting $\tilde{C}_i = C_i$ for all $i \in [L]$. In addition, by Assumption 2.5,

$$\sup_{s\leq t} \operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}} |\hat{y}(X; W'(s)) - \hat{y}(X; W''(s))| \leq KD_L(t) \leq K^L K_0^L(T)\|W' - W''\|_t,$$

completing the proof. ∎

Proof of Lemma B.3. Let us denote

$$D_i(t) = \mathbb{E}\big[\sup_{s\leq t} \operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}} |H_i(X, C_i; W'(s)) - H_i(X, C_i; W''(s))|^2\big]^{1/2},$$
$$\tilde{D}_i(t) = \mathbb{E}\big[\sup_{s\leq t} \operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}} |H_i(X, \tilde{C}_i; W'(s)) - H_i(X, \tilde{C}_i; W''(s))|^2\big]^{1/2}.$$

By Assumption 2.5,

$$\tilde{D}_1(t) \leq K\mathbb{E}\big[\sup_{s\leq t} |w_1'(s, \tilde{C}_1) - w_1''(s, \tilde{C}_1)|^2\big]^{1/2} \leq K\tilde{d}_t(W', W'').$$

The same bound holds for $D_1(t)$. Next let us consider $\tilde{D}_i(t)$. By Assumption 2.5, using Cauchy–Schwarz's inequality, we obtain

$$\begin{aligned}
\tilde{D}_i(t) &\le K\mathbb{E}\Big[\sup_{s\le t}\mathbb{E}_{C_{i-1}}\big[(1+|w_i'(s,C_{i-1},\tilde{C}_i)|+|w_i''(s,C_{i-1},\tilde{C}_i)|+|b_i'(s,\tilde{C}_i)|\\
&\quad+|b_i''(s,\tilde{C}_i)|)\operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}}|H_{i-1}(X,C_{i-1};W'(s))-H_{i-1}(X,C_{i-1};W''(s))|\big]^2\Big]^{1/2}\\
&\quad+K\mathbb{E}\Big[\sup_{s\le t}\mathbb{E}_{C_{i-1}}[|w_i'(s,C_{i-1},\tilde{C}_i)-w_i''(s,C_{i-1},\tilde{C}_i)|+|b_i'(s,\tilde{C}_i)-b_i''(s,\tilde{C}_i)|]^2\Big]^{1/2}\\
&\le K\mathbb{E}\Big[D_{i-1}^2(t)\sup_{s\le t}\mathbb{E}_{C_{i-1}}\big[1+|w_i'(s,C_{i-1},\tilde{C}_i)|^2+|w_i''(s,C_{i-1},\tilde{C}_i)|^2\\
&\quad+|b_i'(s,\tilde{C}_i)|^2+|b_i''(s,\tilde{C}_i)|^2\big]\Big]^{1/2}\\
&\quad+K\mathbb{E}\Big[\sup_{s\le t}\mathbb{E}_{C_{i-1}}[|w_i'(s,C_{i-1},\tilde{C}_i)-w_i''(s,C_{i-1},\tilde{C}_i)|+|b_i'(s,\tilde{C}_i)-b_i''(s,\tilde{C}_i)|]^2\Big]^{1/2}\\
&\le KK_*(T)D_{i-1}(t)+K\tilde{d}_t(W',W'').
\end{aligned}$$

We have the same bound for $D_i(t)$. Hence,

$$\max(D_i(t),\tilde{D}_i(t))\le KK_*(T)\max(D_{i-1}(t),\tilde{D}_{i-1}(t))+K\tilde{d}_t(W',W'').$$

This, in particular, implies

$$\max_{1\le i\le L}\max(D_i(t),\tilde{D}_i(t))\le K^LK_*^L(T)\,\tilde{d}_t(W',W''),$$

which proves the statement. ∎

B.4. Proof of Lemmas B.2 and B.4

Proof of Lemma B.2. This is a special case of Lemma B.4 with $\tilde{C}_i=C_i$ for all $i\in[L]$, $K_*(T)=K_0(T)$ and $\Xi(B)=2Le^{1-K_1B^2}$. ∎

Proof of Lemma B.4. First of all, by Cauchy–Schwarz's inequality, we have, from Assumption 2.6,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z,\tilde{C}_i;W'(s))|^{50}\Big]^{1/50}\\
&\quad\le K\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\big|\mathbb{E}_{C_{i+1}}\big[(1+|\Delta_{i+1}^H(Z,C_{i+1};W'(s))|)\\
&\qquad\times(1+|w_{i+1}'(s,\tilde{C}_i,C_{i+1})|+|b_{i+1}'(s,C_{i+1})|)\big]\big|^{50}\Big]^{1/50}\\
&\quad\le K\mathbb{E}\Big[\mathbb{E}_{C_{i+1}}\big[1+\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(Z,C_{i+1};W'(s))|^2\big]^{25}\\
&\qquad\times\mathbb{E}_{C_{i+1}}\big[1+\sup_{s\le t}|w_{i+1}'(s,\tilde{C}_i,C_{i+1})|^2+\sup_{s\le t}|b_{i+1}'(s,C_{i+1})|^2\big]^{25}\Big]^{1/50}
\end{aligned}$$

$$\begin{aligned}
&\le K\big(1+\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(Z,C_{i+1};W'(s))|^2\big]^{1/2}\big)\\
&\quad\times\big(1+\mathbb{E}\big[\sup_{s\le t}|w'_{i+1}(s,\tilde{C}_i,C_{i+1})|^{50}\big]^{1/50}+\mathbb{E}\big[\sup_{s\le t}|b'_{i+1}(s,C_{i+1})|^2\big]^{1/2}\big)\\
&\le KK_*(T)\big(1+\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(Z,C_{i+1};W'(s))|^{50}\big]^{1/50}\big).
\end{aligned}$$

We have, similarly,

$$\begin{aligned}
&\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i}^H(Z,C_{i};W'(s))|^{50}\big]^{1/50}\\
&\quad\le KK_*(T)\big(1+\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i+1}^H(Z,C_{i+1};W'(s))|^{50}\big]^{1/50}\big).
\end{aligned}$$

Therefore, for any $i\in[L]$,

$$\begin{aligned}
\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i}^H(Z,C_{i};W'(s))|^{50}\big]^{1/50}&\le K^LK_*^L(T),\\
\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i}^H(Z,\tilde{C}_{i};W'(s))|^{50}\big]^{1/50}&\le K^LK_*^L(T).
\end{aligned}\tag{B.1}$$

The same bound holds for W''. With this, let us proceed with two steps.

Step 1. For brevity, let us define

$$\begin{aligned}
D_i^H(t)&=\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i}^H(Z,C_{i};W'(s))-\Delta_{i}^H(Z,C_{i};W''(s))|^2\big]^{1/2},\\
\tilde{D}_i^H(t)&=\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{i}^H(Z,\tilde{C}_{i};W'(s))-\Delta_{i}^H(Z,\tilde{C}_{i};W''(s))|^2\big]^{1/2}.
\end{aligned}$$

We first have, from Assumption 2.6 and Lemma B.3,

$$\begin{aligned}
D_L^H(t)=\tilde{D}_L^H(t)&\le K\sup_{s\le t}\operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}}|H_L(X,1;W'(s))-H_L(X,1;W''(s))|\\
&\quad+K\sup_{s\le t}\operatorname*{ess\text{-}sup}_{X\sim\mathcal{P}}|\hat{y}(X;W'(s))-\hat{y}(X;W''(s))|\\
&\le K^LK_*^L(T)\tilde{d}_t(W',W'').
\end{aligned}$$

Next we consider $\tilde{D}_{i-1}^H$ and D_{i-1}^H for $i\ge 2$. By Assumption 2.6,

$$\tilde{D}_{i-1}^H(t)\le K\big(\tilde{D}_{i-1}^{H,1}(t)+\tilde{D}_{i-1}^{H,2}(t)+\tilde{D}_{i-1}^{H,3}(t)+\tilde{D}_{i-1}^{H,4}(t)+\tilde{D}_{i-1}^{H,5}(t)\big),$$

in which

$$\begin{aligned}
\tilde{D}_{i-1}^{H,1}(t)=\mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[\big(1+|w'_i(s,\tilde{C}_{i-1},C_i)|+|w''_i(s,\tilde{C}_{i-1},C_i)|\\
+|b'_i(s,C_i)|+|b''_i(s,C_i)|\big)|\Delta_i^H(z,C_i;W'(s))-\Delta_i^H(z,C_i;W''(s))|\big]^2\Big]^{1/2},
\end{aligned}$$

$$\begin{aligned}
\tilde{D}_{i-1}^{H,2}(t) =& \mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[(1+|\Delta_i^H(Z,C_i;W'(s))|+|\Delta_i^H(Z,C_i;W''(s))|)\\
&\times(|w_i'(s,\tilde{C}_{i-1},C_i)-w_i''(s,\tilde{C}_{i-1},C_i)|+|b_i'(s,C_i)-b_i''(s,C_i)|)\big]^2\Big]^{1/2},\\
\tilde{D}_{i-1}^{H,3}(t) =& \mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[(1+|\Delta_i^H(Z,C_i;W'(s))|+|\Delta_i^H(Z,C_i;W''(s))|)\\
&\times(1+|w_i'(s,\tilde{C}_{i-1},C_i)|+|w_i''(s,\tilde{C}_{i-1},C_i)|+|b_i'(s,C_i)|+|b_i''(s,C_i)|)\\
&\times|H_i(X,C_i;W'(s))-H_i(X,C_i;W''(s))|\big]^2\Big]^{1/2},\\
\tilde{D}_{i-1}^{H,4}(t) =& \mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[(1+|\Delta_i^H(Z,C_i;W'(s))|+|\Delta_i^H(Z,C_i;W''(s))|)\\
&\times(1+|b_i'(s,C_i)|+|b_i''(s,C_i)|)\big]^2\\
&\times|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\Big]^{1/2},\\
\tilde{D}_{i-1}^{H,5}(t) =& \mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[(1+|\Delta_i^H(Z,C_i;W'(s))|+|\Delta_i^H(Z,C_i;W''(s))|)\\
&\times(|w_i'(s,\tilde{C}_{i-1},C_i)|+|w_i''(s,\tilde{C}_{i-1},C_i)|)\big]^2\\
&\times|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\Big]^{1/2}.
\end{aligned}$$

We bound each term. For $\tilde{D}_{i-1}^{H,1}$, we use Cauchy–Schwarz's inequality to obtain

$$\begin{aligned}
\tilde{D}_{i-1}^{H,1}(t) &\le K\mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[\big(1+|w_i'(s,\tilde{C}_{i-1},C_i)|^2+|w_i''(s,\tilde{C}_{i-1},C_i)|^2\\
&\quad+|b_i'(s,C_i)|^2+|b_i''(s,C_i)|^2\big)|D_i^H(t)|^2\big]\Big]^{1/2}\\
&\le KK_*(T)D_i^H(t).
\end{aligned}$$

Similarly, using equation (B.1),

$$\begin{aligned}
\tilde{D}_{i-1}^{H,2}(t) &\le K\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[1+|\Delta_i^H(Z,C_i;W'(s))|^2+|\Delta_i^H(Z,C_i;W''(s))|^2\big]^{1/2}\\
&\quad\times\tilde{d}_t(W',W'')\\
&\le K^LK_*^L(T)\,\tilde{d}_t(W',W'').
\end{aligned}$$

To bound $\tilde{D}_{i-1}^{H,3}$, we use Lemma B.3 and equation (B.1):

$$\begin{aligned}
&\tilde{D}_{i-1}^{H,3}(t)\\
&\le K\mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[1+|\Delta_i^H(Z,C_i;W'(s))|^4+|\Delta_i^H(Z,C_i;W''(s))|^4\big]^{1/2}\\
&\quad\times\mathbb{E}_{C_i}\big[1+|w_i'(s,\tilde{C}_{i-1},C_i)|^4+|w_i''(s,\tilde{C}_{i-1},C_i)|^4+|b_i'(s,C_i)|^4+|b_i''(s,C_i)|^4\big]^{1/2}\\
&\quad\times\mathbb{E}_{C_i}[|H_i(X,C_i;W'(s))-H_i(X,C_i;W''(s))|^2]\Big]^{1/2}\\
&\le K^{2L+2}K_*^{2L+2}(T)\,\tilde{d}_t(W',W''),
\end{aligned}$$

and similarly, for $\tilde{D}_{i-1}^{H,4}$,

$$\begin{aligned}
&\tilde{D}_{i-1}^{H,4}(t)\\
&\le K\mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[1+|\Delta_i^H(Z,C_i;W'(s))|^2+|\Delta_i^H(Z,C_i;W''(s))|^2\big]\\
&\quad\times\mathbb{E}_{C_i}[1+|b_i'(s,C_i)|^2+|b_i''(s,C_i)|^2]\\
&\quad\times|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\Big]^{1/2}\\
&\le K^{2L+2}K_*^{2L+2}(T)\,\tilde{d}_t(W',W'').
\end{aligned}$$

The treatment of $\tilde{D}_{i-1}^{H,5}$ requires more care. Cauchy–Schwarz's inequality and equation (B.1) give us

$$\begin{aligned}
&\tilde{D}_{i-1}^{H,5}(t)\\
&\le K\mathbb{E}_{\tilde{C}_{i-1}}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\mathbb{E}_{C_i}\big[1+|\Delta_i^H(Z,C_i;W'(s))|^2+|\Delta_i^H(Z,C_i;W''(s))|^2\big]\\
&\quad\times\mathbb{E}_{C_i}[|w_i'(s,\tilde{C}_{i-1},C_i)|^2+|w_i''(s,\tilde{C}_{i-1},C_i)|^2]\\
&\quad\times|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\Big]^{1/2}\\
&\le K^LK_*^L(T)\mathbb{E}\big[\sup_{s\le t}(|w_i'(s,\tilde{C}_{i-1},C_i)|^2+|w_i''(s,\tilde{C}_{i-1},C_i)|^2)\\
&\quad\times\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\big]^{1/2}.
\end{aligned}$$

Recall our assumption

$$\begin{aligned}
\mathbb{P}(\widetilde{\mathsf{max}}_T^w(W')\ge K_*(T)B)\le\Xi(B),\\
\mathbb{P}(\widetilde{\mathsf{max}}_T^w(W'')\ge K_*(T)B)\le\Xi(B).
\end{aligned}$$

We also have, from Assumption 2.5,

$$\begin{aligned}
&\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_{i-1}(X,\tilde{C}_{i-1};W'(s))|^8\big]^{1/8}\\
&\quad\le K\big(1+\mathbb{E}\big[\sup_{s\le t}|w_{i-1}'(s,C_{i-2},\tilde{C}_{i-1})|^8\big]^{1/8}+\mathbb{E}\big[\sup_{s\le t}|b_{i-1}'(s,\tilde{C}_{i-1})|^8\big]^{1/8}\big)\\
&\quad\le KK_*(T),
\end{aligned}$$

and similarly,

$$\mathbb{E}\big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^8\big]^{1/8}\le KK_*(T).$$

As such, denoting the event

$$E=\big\{\sup_{s\le t}|w_i'(s,\tilde{C}_{i-1},C_i)|\ge K_*(T)B,\ \sup_{s\le t}|w_i''(s,\tilde{C}_{i-1},C_i)|\ge K_*(T)B\big\},$$

we obtain, from Lemma B.3,

$$
\begin{aligned}
&\tilde{D}_{i-1}^{H,5}(t)\\
&\le K^L K_*^L(T)\mathbb{E}\Big[\sup_{s\le t}(|w_i'(s,\tilde{C}_{i-1},C_i)|^2+|w_i''(s,\tilde{C}_{i-1},C_i)|^2)\\
&\quad\times \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2(\mathbb{I}(\neg E)+\mathbb{I}(E))\Big]^{1/2}\\
&\le K^L K_*^{L+1}(T)B\\
&\quad\times\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\Big]^{1/2}\\
&\quad+K^L K_*^L(T)\mathbb{E}\Big[\sup_{s\le t}|w_i'(s,\tilde{C}_{i-1},C_i)|^8+\sup_{s\le t}|w_i''(s,\tilde{C}_{i-1},C_i)|^8\Big]^{1/8}\\
&\quad\times\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(|H_{i-1}(X,\tilde{C}_{i-1};W'(s))|^8+|H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^8)\Big]^{1/8}\mathbb{P}(E)^{1/2}\\
&\le K^{2L+2}K_*^{2L+2}(T)B\tilde{d}_t(W',W'')+K^L K_*^{L+2}(T)\sqrt{\Xi(B)}.
\end{aligned}
$$

Putting all the bounds together yields

$$
\begin{aligned}
\tilde{D}_{i-1}^H(t)\le{}& KK_*(T)D_i^H(t)+(KK_*(T))^{2L+2}(1+B)\,\tilde{d}_t(W',W'')\\
&+K^L K_*^{L+2}(T)\sqrt{\Xi(B)}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
D_{i-1}^H(t)\le{}& KK_*(T)D_i^H(t)+(KK_*(T))^{2L+2}(1+B)\,\tilde{d}_t(W',W'')\\
&+K^L K_*^{L+2}(T)\sqrt{\Xi(B)}.
\end{aligned}
$$

Together with the bound on D_L^H and $\tilde{D}_L^H$, we thus obtain

$$
\begin{aligned}
&\max_{1\le i\le L}\max(\tilde{D}_i^H(t),D_i^H(t))\\
&\qquad\le(KK_*(T))^{3L+2}((1+B)\tilde{d}_t(W',W'')+\sqrt{\Xi(B)}).
\end{aligned}
\tag{B.2}
$$

This completes the first step.

Step 2. We now prove the main claims of the lemma. For brevity, for $i\ge 2$, let us denote

$$
\begin{aligned}
\tilde{D}_i^w(t)&=\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^w(Z,\tilde{C}_{i-1},\tilde{C}_i;W'(s))-\Delta_i^w(Z,\tilde{C}_{i-1},\tilde{C}_i;W''(s))|^2\Big]^{1/2},\\
\tilde{D}_i^b(t)&=\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^b(Z,\tilde{C}_i;W'(s))-\Delta_i^b(Z,\tilde{C}_i;W''(s))|^2\Big]^{1/2},\\
\tilde{D}_1^w(t)&=\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_1^w(Z,\tilde{C}_1;W'(s))-\Delta_1^w(Z,\tilde{C}_1;W''(s))|^2\Big]^{1/2}.
\end{aligned}
$$

By Assumption 2.5,

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

in which

$$\begin{aligned}\tilde{D}_i^{w,1}(t) &= \mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1+|\Delta_i^H(Z,\tilde{C}_i;W'(s))|^2+|\Delta_i^H(Z,\tilde{C}_i;W''(s))|^2)\\ &\quad\times(|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2)\Big]^{1/2},\\ \tilde{D}_i^{w,2}(t) &= \mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\big(|\Delta_i^H(Z,\tilde{C}_i;W'(s))-\Delta_i^H(Z,\tilde{C}_i;W''(s))|^2\\ &\quad+|w_i'(s,\tilde{C}_{i-1},\tilde{C}_i)-w_i''(s,\tilde{C}_{i-1},\tilde{C}_i)|^2+|b_i'(s,\tilde{C}_i)-b_i''(s,\tilde{C}_i)|^2\\ &\quad+|H_i(X,\tilde{C}_i;W'(s))-H_i(X,\tilde{C}_i;W''(s))|^2\big)\Big]^{1/2}.\end{aligned}$$

We bound $\tilde{D}_i^{w,1}$:

$$\begin{aligned}\tilde{D}_i^{w,1}(t) &\le \mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1+|\Delta_i^H(Z,\tilde{C}_i;W'(s))|^2+|\Delta_i^H(Z,\tilde{C}_i;W''(s))|^2)\Big]^{1/2}\\ &\quad\times\mathbb{E}\Big[\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_{i-1}(X,\tilde{C}_{i-1};W'(s))-H_{i-1}(X,\tilde{C}_{i-1};W''(s))|^2\Big]^{1/2}\\ &\le K^{2L}K_*^{2L}(T)\tilde{d}_t(W',W''),\end{aligned}$$

where we have used equation (B.1) and Lemma B.3. We also have the following bound on $\tilde{D}_i^{w,2}$ from Lemma B.3 and equation (B.2):

$$\tilde{D}_i^{w,2}(t)\le(KK_*(T))^{3L+2}((1+B)\tilde{d}_t(W',W'')+\sqrt{\Xi(B)}),$$

which therefore leads to

$$\tilde{D}_i^{w}(t)\le(KK_*(T))^{3L+2}((1+B)\tilde{d}_t(W',W'')+\sqrt{\Xi(B)}).$$

The same bound similarly applies to $\tilde{D}_i^b(t)$ and $\tilde{D}_1^w(t)$. ∎

C. Remaining proofs for Section 4

C.1. Proofs of Propositions 4.14, 4.15 and 4.16

Before delving into the proofs, we introduce some auxiliary results. We first present a useful concentration result. In fact, the tail bound can be improved using the argument in [15], but the following simpler version is sufficient for our purposes.

Lemma C.1. *Consider an integer $n\ge 2$; let $(c_1,c_2,\dots,c_n)$ be η-independent for $\eta\in[0,1/2]$ and let x be another independent random variable. Let $\mathbb{E}_x$ and $\mathbb{E}_c$ denote the expectations with respect to x only and $\{c_i\}_{i\in[n]}$ only, respectively. Consider a collection of mappings $\{f_i\}_{i\in[n]}$, which map to the same separable Hilbert space.*

Let $f_i(x) = \mathbb{E}_c[f_i(c_i, x)]$. *Assume that* $|f_i(c_i, x) - f_i(x)| \leq R$ *for almost every* x *and* c_i, *then for any* $\delta > 2\eta R$,

$$\mathbb{P}\Big(\mathbb{E}_x\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} f_i(c_i, x) - f_i(x)\Big|\Big] \geq \delta\Big) \leq \frac{4R}{\delta}\exp\Big(-\frac{n\delta^2}{512R^2}\Big).$$

Proof. For brevity, let us define

$$Z_n(x) = \sum_{i=1}^{n}(f_i(c_i, x) - f_i(x)).$$

By Theorem A.2, for $\delta > 2\eta R$,

$$\mathbb{P}(|Z_n(x)| \geq n\delta \mid x) \leq 2\exp\Big(-\frac{n\delta^2}{64R^2}\Big),$$

and therefore

$$\mathbb{P}(|Z_n(x)| \geq n\delta) \leq 2\exp\Big(-\frac{n\delta^2}{64R^2}\Big),$$

since the right-hand side is uniform in x. Next note that, with respect to the randomness of x only,

$$\begin{aligned}\mathbb{E}_x[|Z_n(x)|] &= \mathbb{E}_x[|Z_n(x)|\mathbb{I}(|Z_n(x)| \geq n\delta/2)] + \mathbb{E}_x[|Z_n(x)|\mathbb{I}(|Z_n(x)| < n\delta/2)] \\ &\leq \mathbb{E}_x[|Z_n(x)|\mathbb{I}(|Z_n(x)| \geq n\delta/2)] + n\delta/2.\end{aligned}$$

As such, by Markov's inequality and Cauchy–Schwarz's inequality,

$$\begin{aligned}\mathbb{P}(\mathbb{E}_x[|Z_n(x)|] \geq n\delta) &\leq \mathbb{P}\big(\mathbb{E}_x[|Z_n(x)|\mathbb{I}(|Z_n(x)| \geq n\delta/2)] \geq n\delta/2\big) \\ &\leq \frac{2}{n\delta}\,\mathbb{E}[|Z_n(x)|\mathbb{I}(|Z_n(x)| \geq n\delta/2)] \\ &\leq \frac{2}{n\delta}\,\mathbb{E}[|Z_n(x)|^2]^{1/2}\,\mathbb{P}(|Z_n(x)| \geq n\delta/2)^{1/2} \\ &\leq \frac{4}{n\delta}\,\mathbb{E}[|Z_n(x)|^2]^{1/2}\exp\Big(-\frac{n\delta^2}{512R^2}\Big).\end{aligned}$$

Notice that, since $c_1, \ldots, c_n$ are η-independent and $f_i(x) = \mathbb{E}_c[f_i(c_i, x)]$,

$$\mathbb{E}[|Z_n(x)|^2] \leq nR^2 + \eta n^2 R^2.$$

We thus get

$$\mathbb{P}(\mathbb{E}_x[|Z_n(x)|] \geq n\delta) \leq \frac{4\sqrt{1+\eta n}R}{\sqrt{n}\delta}\exp\Big(-\frac{n\delta^2}{512R^2}\Big) \leq \frac{4R}{\delta}\exp\Big(-\frac{n\delta^2}{512R^2}\Big).$$

This proves the claim. ∎

The next useful result concerns with the sampling at initialization.

Lemma C.2. *Under Assumption 4.6, following the coupling procedure, we have, for any $\delta > 0$ and $B > 0$, with probability at least $1 - KLn_{\max}\exp(-K(\delta \wedge \delta^{1/26})n_{\min}^{1/52})$, that the following hold, for $2 \le i \le L$:*

Moment bounds:

$$|||\tilde{W}|||_0 = |||\mathbf{W}|||_0 \le |||W|||_0 + \delta^{1/50},$$

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_{C_{i-1}}[|w_i^0(C_{i-1}, C_i(j_i))|^{50}]\Big)^{1/50} \le \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^{50}]^{1/50} + \delta^{1/50},$$

$$\Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_{C_i}[|w_i^0(C_{i-1}(j_{i-1}), C_i)|^{50}]\Big)^{1/50} \le \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^{50}]^{1/50} + \delta^{1/50}.$$

Excess bounds:

$$\Big|\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\mathbb{I}(|w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))| \ge B) - \mathbb{P}(|w_i^0(C_{i-1}, C_i)| \ge B)\Big| \le \delta,$$

$$\Big|\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{I}(|b_i^0(C_i(j_i))| \ge B) - \mathbb{P}(|b_i^0(C_i)| \ge B)\Big| \le \delta.$$

Here $n_{\max} = \max(n_1, \dots, n_L)$ and $n_{\min} = \min(n_1, \dots, n_{L-1})$.

Proof. We treat the bounds separately.

The moment bounds. We recall that

$$|||\tilde{W}|||_0 = \max\Big(\max_{2\le i\le L}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))|^{50}\Big)^{1/50},$$
$$\max_{2\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}|b_i^0(C_i(j_i))|^{50}\Big)^{1/50}, \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}|w_1^0(C_1(j_1))|^{50}\Big)^{1/50}\Big).$$

Let us first prove the following:

$$\mathbb{P}(Z^{(1)} \ge \delta) \le e \cdot \exp(-K\delta^{1/26}n_1^{1/52}),$$
$$Z^{(1)} = \Big|\frac{1}{n_1}\sum_{j_1=1}^{n_1}|w_1^0(C_1(j_1))|^{50} - \mathbb{E}[|w_1^0(C_1)|^{50}]\Big|.$$

Indeed, we note that for any $m \ge 1$, $\mathbb{E}[|w_1^0(C_1)|^{50m}]^{1/m} \le Km^{25}$. As such, by Theorem A.3,

$$\mathbb{E}[|Z^{(1)}|^m]^{1/m} \le Km^{26}n_1^{-1/2}.$$

This implies $|Z^{(1)}|^{1/52}$ is $Kn_1^{-1/104}$-sub-Gaussian, from which the claim follows. Using the same argument, we get

$$\mathbb{P}(Z^{(L)} \geq \delta) \leq e \cdot \exp(-K\delta^{1/26} n_{L-1}^{1/52}),$$
$$Z^{(L)} = \Big| \frac{1}{n_{L-1}} \sum_{j_{L-1}=1}^{n_{L-1}} |w_L^0(C_{L-1}(j_{L-1}), 1)|^{50} - \mathbb{E}[|w_L^0(C_{L-1}, 1)|^{50}] \Big|,$$

as well as that

$$\mathbb{P}(A^{(i)} \geq \delta) \leq e \cdot \exp(-K\delta^{1/26} n_i^{1/52}),$$
$$A^{(i)} = \Big| \frac{1}{n_i} \sum_{j_i=1}^{n_i} |b_i^0(C_i(j_i))|^{50} - \mathbb{E}[|b_i^0(C_i)|^{50}] \Big|,$$

for $2 \leq i \leq L-1$. In addition, since $\Omega_L = \{1\}$ and $n_L = 1$, it is obvious that we have $|b_L^0(C_L(j_L))| = |b_L^0(C_L)|$.

Next for $2 \leq i \leq L-1$, without loss of generality, suppose $n_i \geq n_{i-1}$. Let us prove the following:

$$\mathbb{P}(Z^{(i)} \geq \delta) \leq e n_i \cdot \exp(-K\delta^{1/26} n_{i-1}^{1/52}),$$
$$Z^{(i)} = \Big| \frac{1}{n_{i-1} n_i} \sum_{j_{i-1}=1}^{n_{i-1}} \sum_{j_i=1}^{n_i} |w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))|^{50} - \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^{50}] \Big|.$$

For fixed $j_i \in [n_i]$, let us first consider

$$\tilde{C}_{j_i}(j_{i-1}) = \big(C_{i-1}(j_{i-1}), C_i((j_{i-1} + j_i) \bmod n_i)\big).$$

For any 1-bounded function f, due to independence between $C_i(j_i)$ and $C_{i-1}(j_{i-1})$ and Assumption 4.4, we have

$$\big|\mathbb{E}[f(\tilde{C}_{j_i}(j_{i-1})) \mid \{\tilde{C}_{j_i}(j'_{i-1})\}_{j'_{i-1} \neq j_{i-1}}, C_i((j_{i-1} + j_i) \bmod n_i)] - \mathbb{E}[f(\tilde{C}_{j_i}(j_{i-1})) \mid C_i((j_{i-1} + j_i) \bmod n_i)]\big| \leq \eta_{i-1},$$

which implies

$$\big|\mathbb{E}[f(\tilde{C}_{j_i}(j_{i-1})) \mid \{\tilde{C}_{j_i}(j'_{i-1})\}_{j'_{i-1} \neq j_{i-1}}] - \mathbb{E}[f(\tilde{C}_{j_i}(j_{i-1}))]\big| \leq \eta_{i-1}.$$

That is, $\{\tilde{C}_{j_i}(j_{i-1})\}_{j_{i-1} \in [n_{i-1}]}$ is η_{i-1}-independent. Hence, by the same argument, by letting

$$Z_{j_i}^{(i)} = \Big| \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} |w_i^0(\tilde{C}_{j_i}(j_{i-1}))|^{50} - \mathbb{E}[|w_i^0(\tilde{C}_{j_i}(j_{i-1}))|^{50}] \Big|,$$

we have

$$\mathbb{P}(Z_{j_i}^{(i)} \geq \delta) \leq e \cdot \exp(-K\delta^{1/26} n_{i-1}^{1/52}).$$

By the union bound,

$$\mathbb{P}(Z^{(i)} \geq \delta) \leq \mathbb{P}\big(\max_{j_i \leq n_i} Z_{j_i}^{(i)} \geq \delta\big) \leq e n_i \cdot \exp(-K\delta^{1/26} n_{i-1}^{1/52}),$$

which is the desired claim.

Upon an application of the union bound, these probability bounds imply the bound on the probability of the event $|||\tilde{W}|||_0 = |||\mathbf{W}|||_0 \leq |||W|||_0 + \delta^{1/50}$. The rest of the bounds are similarly proven.

The excess bounds. Without loss of generality, assume $n_i \geq n_{i-1}$ for $2 \leq i \leq L-1$. Let us denote

$$D_{j_i}^{(i)} = \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \mathbb{I}(|w_i^0(\tilde{C}_{j_i}(j_{i-1}))| \geq B) - \mathbb{P}(|w_i^0(C_{i-1}, C_i)| \geq B).$$

Recall previously that $\{\tilde{C}_{j_i}(j_{i-1})\}_{j_{i-1} \in [n_{i-1}]}$ is η_{i-1}-independent. As such, by Theorem A.3,

$$\mathbb{E}[|D_{j_i}^{(i)}|^m]^{1/m} \leq Kmn_{i-1}^{-1/2}.$$

This implies that $|D_{j_i}^{(i)}|^{1/2}$ is $Kn_{i-1}^{-1/4}$-sub-Gaussian, and hence for any $\delta > 0$,

$$\mathbb{P}(|D_{j_i}^{(i)}| \geq \delta) \leq e \cdot \exp(-K\delta n_{i-1}^{1/2}).$$

The union bound yields

$$\mathbb{P}\Big(\Big|\frac{1}{n_i} \sum_{j_i=1}^{n_i} D_{j_i}^{(i)}\Big| \geq \delta\Big) \leq K n_i \exp(-K\delta n_{i-1}^{1/2}).$$

Note that this holds for any $B > 0$. The rest of the bounds are similarly proven. ∎

Similar to Lemma 3.2, one can prove the following.

Lemma C.3. *Under Assumptions 2.4 and 2.6, for any $t \in [0, \infty)$,*

$$|||\tilde{W}|||_t \leq K^{\kappa_L}(1 + t^{\kappa_L})(1 + |||\tilde{W}|||_0^{\kappa_L}), \quad |||\mathbf{W}|||_{\lfloor t/\epsilon \rfloor} \leq K^{\kappa_L}(1 + t^{\kappa_L})(1 + |||\mathbf{W}|||_0^{\kappa_L}),$$

where $\kappa_L = K^L$ for some constant $K > 1$ sufficiently large. In particular, for any $i \in [L]$,

$$\Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \sup_{s \leq t} \operatorname{ess-sup}_{Z \sim \mathcal{P}} |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(s))|^{50}\Big)^{1/50} \leq K^{\kappa_L}(1 + t^{\kappa_L})(1 + |||\tilde{W}|||_0^{\kappa_L}),$$

$$\Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \sup_{s \leq t} \operatorname{ess-sup}_{Z \sim \mathcal{P}} |\Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(\lfloor s/\epsilon \rfloor))|^{50}\Big)^{1/50} \leq K^{\kappa_L}(1 + t^{\kappa_L})(1 + |||\mathbf{W}|||_0^{\kappa_L}).$$

Furthermore, by defining

$$
\begin{aligned}
|||W|||_{\mathrm{samp},t} = \max\Bigg(& \max_{2\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_{C_{i-1}}\big[\sup_{s\le t}|w_i(s,C_{i-1},C_i(j_i))|^{50}\big]\Big)^{1/50},\\
& \max_{2\le i\le L}\Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_{C_i}\big[\sup_{s\le t}|w_i(s,C_{i-1}(j_{i-1}),C_i)|^{50}\big]\Big)^{1/50},\\
& \max_{2\le i\le L}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\le t}|w_i(s,C_{i-1}(j_{i-1}),C_i(j_i))|^{50}\Big)^{1/50},\\
& \max_{2\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}|b_i(s,C_i(j_i))|^{50}\Big)^{1/50},\\
& \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{s\le t}|w_1(s,C_1(j_1))|^{50}\Big)^{1/50}\Bigg),
\end{aligned}
$$

we also have

$$
\begin{aligned}
&|||\,W\,|||_{\mathrm{samp},t},\ \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z,C_i(j_i);W(s))|^{50}\Big)^{1/50}\\
&\qquad\le K^{\kappa_L}(1+t^{\kappa_L})(1+\max(|||W|||_0^{\kappa_L},|||W|||_{\mathrm{samp},0}^{\kappa_L})).
\end{aligned}
$$

Proof. The proof follows the same argument as Lemma 3.2. This is obvious for the statements concerning $\tilde{W}$ and $\mathbf{W}$. To prove the latter claims that involve $|||W|||_{\mathrm{samp},t}$, the argument follows similarly. In particular, let us denote

$$
\begin{aligned}
&|||\,w_i\,|||_{\mathrm{right},t} = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_{C_{i-1}}\big[\sup_{s\le t}|w_i(s,C_{i-1},C_i(j_i))|^{50}\big]\Big)^{1/50}, && 2\le i\le L,\\
&|||\,w_i\,|||_{\mathrm{left},t} = \Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_{C_i}\big[\sup_{s\le t}|w_i(s,C_{i-1}(j_{i-1}),C_i)|^{50}\big]\Big)^{1/50}, && 2\le i\le L,\\
&|||\,w_i\,|||_{\mathrm{cen},t} = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\le t}|w_i(s,C_{i-1}(j_{i-1}),C_i(j_i))|^{50}\Big)^{1/50}, && 2\le i\le L,\\
&|||\,b_i\,|||_{\mathrm{samp},t} = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}|b_i(s,C_i(j_i))|^{50}\Big)^{1/50}, && 2\le i\le L,\\
&|||\,w_1\,|||_{\mathrm{samp},t} = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{s\le t}|w_1(s,C_1(j_1))|^{50}\Big)^{1/50},\\
&|||\,\Delta_i^H\,|||_{\mathrm{samp},t} = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z,C_i(j_i);W(s))|^{50}\Big)^{1/50}, && 1\le i\le L,
\end{aligned}
$$

$$|||\Delta_i^H|||_t = \mathbb{E}\Big[\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z,C_i;W(s))|^{50}\Big]^{1/50}, \qquad 1\le i\le L.$$

Then similar to Lemma 3.2, we obtain, for $2 \le i \le L$,

$$\begin{aligned}
|||\Delta_L^H|||_{\text{samp},t} &\le K,\\
|||\Delta_{i-1}^H|||_{\text{samp},t} &\le K(1+|||\Delta_i^H|||_t)(1+|||w_i|||_{\text{left},t}+|||b_i|||_t),\\
|||w_i|||_{\text{left},t} &\le |||w_i|||_{\text{left},0}+K(1+|||\Delta_i^H|||_t)t,\\
|||w_i|||_{\text{right},t} &\le |||w_i|||_{\text{right},0}+K(1+|||\Delta_i^H|||_{\text{samp},t})t,\\
|||w_i|||_{\text{cen},t} &\le |||w_i|||_{\text{cen},0}+K(1+|||\Delta_i^H|||_{\text{samp},t})t,\\
|||b_i|||_{\text{samp},t} &\le |||b_i|||_{\text{samp},0}+K(1+|||\Delta_i^H|||_{\text{samp},t})t,\\
|||w_1|||_{\text{samp},t} &\le |||w_1|||_{\text{samp},0}+K(1+|||\Delta_1^H|||_{\text{samp},t})t.
\end{aligned}$$

Note that

$$\begin{aligned}
|||W|||_{\text{samp},t} = \max\Big(&\max_{2\le i\le L}|||w_i|||_{\text{left},t}, \max_{2\le i\le L}|||w_i|||_{\text{right},t}, \max_{2\le i\le L}|||w_i|||_{\text{cen},t},\\
&\max_{2\le i\le L}|||b_i|||_{\text{samp},t}, |||w_1|||_{\text{samp},t}\Big).
\end{aligned}$$

Together with the bound on $|||\Delta_i^H|||_t$ given by Lemma 3.2, one can derive the claims. The proof is complete. ∎

C.1.1. Proof of Proposition 4.14.

Proof of Proposition 4.14. In the following, let K_t denote an immaterial positive constant that takes the form

$$K_t = K^{\kappa_L}(1+t^{\kappa_L}),$$

where $\kappa_L = K^L$, such that $K_t \ge 1$ and $K_t \le K_T$ for all $t \le T$. We note that the terminal time T, the constant K_t, as well as the usual immaterial constant K, do not depend on B. We start with some preliminary facts.

Fact 1: *moment bounds.* We first note that at initialization, we have $\mathcal{D}_0(W,\tilde{W}) = 0$ and $|||W|||_0 \le K$. By Assumption 4.6 and Lemma 3.2, $|||W|||_T \le K_T$. Furthermore, by Lemma C.2, with probability at least $1 - KLn_{\max}\exp(-Kn_{\min}^{1/52})$, we have $|||\tilde{W}|||_0 \le K$ and $|||W|||_{\text{samp},0} \le K$, recalling the definition of $|||W|||_{\text{samp},t}$ from the statement of Lemma C.3. Let this event be denoted by $\mathcal{E}$. Unless noticed otherwise, we shall place most of the contexts of our proof upon $\mathcal{E}$. By Lemma C.3, one deduces that

$$\begin{aligned}
&|||\tilde{W}|||_T, |||W|||_{\text{samp},T} \le K_T,\\
&\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(s))|^{50}\Big)^{1/50} \le K_T,
\end{aligned}$$

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z,C_i(j_i);W(s))|^{50}\Big)^{1/50}\le K_T$$

on the event $\mathcal{E}$. We also remark that the fact $|||W|||_T \le K_T$ holds irrespective of $\mathcal{E}$.

Fact 2: *maximal bounds for W*. We note that the assumption ess-sup $\mathsf{max}_0^w(W)$ and ess-sup $\mathsf{max}_0^b(W)$ has an interesting consequence:

$$\begin{aligned}\operatorname*{ess\text{-}sup}\mathsf{max}_T^w(W)&\le K_T(1+B),\\ \operatorname*{ess\text{-}sup}\mathsf{max}_T^b(W)&\le K_T(1+B),\\ \operatorname*{ess\text{-}sup}\max_{1\le i\le L}\sup_{t\le T}|\Delta_i^H(Z,C_i;W(t))|&\le K_T(1+B).\end{aligned}$$

We note that this claim holds irrespective of the event $\mathcal{E}$ from Fact 1. Following this claim, it is immediate that almost surely,

$$\begin{aligned}\operatorname*{ess\text{-}sup}_{C_i}\max_{2\le i\le L}\max_{j_{i-1}\in[n_{i-1}]}|w_i(t,C_{i-1}(j_{i-1}),C_i)|&\le K_T(1+B),\\ \operatorname*{ess\text{-}sup}_{C_{i-1}}\max_{2\le i\le L}\max_{j_i\in[n_i]}|w_i(t,C_{i-1},C_i(j_i))|&\le K_T(1+B),\\ \max_{2\le i\le L}\max_{j_{i-1}\in[n_{i-1}],j_i\in[n_i]}|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|&\le K_T(1+B),\\ \max_{2\le i\le L}\max_{j_i\in[n_i]}|b_i(t,C_i(j_i))|&\le K_T(1+B),\\ \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}\max_{1\le i\le L}\max_{j_i\in[n_i]}\sup_{t\le T}|\Delta_i^H(Z,C_i(j_i);W(t))|&\le K_T(1+B),\end{aligned}$$

since $C_i(j_i)$ is a copy of C_i. Now we prove the claim. First consider $\mathsf{max}_t^w(W)$. By Assumption 2.6, for $\mathcal{P}$-almost every z,

$$\sup_{t\ge0}\sup_{c_{L-1}\in\Omega_{L-1}}|\Delta_L^w(z,c_{L-1},1;W(t))|\le K\big(1+\sup_{t\ge0}|\Delta_L^H(z,1;W(t))|\big)\le K,$$

which implies, by Assumption 2.4, that

$$\operatorname*{ess\text{-}sup}\sup_{t\le T}|w_L(t,C_{L-1},1)|\le\operatorname*{ess\text{-}sup}|w_L^0(C_{L-1},1)|+KT\le K_T(1+B).$$

Next assuming that ess-sup $\sup_{t\le T}|w_i(t,C_{i-1},C_i)|\le K_T(1+B)$ for a given $i\ge2$, by Assumption 2.6, we have for $\mathcal{P}$-almost every z and all $t\le T$,

$$\begin{aligned}&|\Delta_{i-1}^H(z,C_{i-1};W(t))|\\ &\quad\le K\mathbb{E}_{C_i}[(1+|\Delta_i^H(z,C_i;W(t))|)(1+|w_i(t,C_{i-1},C_i)|+|b_i(t,C_i)|)]\\ &\quad\le K\mathbb{E}_{C_i}[(1+|\Delta_i^H(z,C_i;W(t))|)(K_T(1+B)+|b_i(t,C_i)|)]\\ &\quad\le K[(1+\mathbb{E}[|\Delta_i^H(z,C_i;W(t))|^2]^{1/2})(K_T(1+B)+\mathbb{E}[|b_i(t,C_i)|^2]^{1/2})]\\ &\quad\le K_T(1+B),\end{aligned}$$

where the last step follows from the fact $|||W|||_T \leq K_T$ and Lemma 3.2. Again by Assumption 2.6, we then obtain

$$|\Delta_{i-1}^w(z, C_{i-1}, C_i; W(t))| \leq K_T(1+B),$$

which implies, by Assumption 2.4, that

$$\begin{aligned}\text{ess-sup} \sup_{t\leq T}|w_{i-1}(t, C_{i-1}, C_i)| &\leq \text{ess-sup}|w_{i-1}^0(C_{i-1}, C_i)| + K_T(1+B)T \\ &\leq K_T(1+B).\end{aligned}$$

This completes the induction argument to show that ess-sup $\mathsf{max}_t^w(W) \leq K_T(1+B)$. We have also showed that

$$\text{ess-sup} \max_{1\leq i\leq L} \sup_{t\leq T}|\Delta_i^H(Z, C_i; W(t))| \leq K_T(1+B).$$

We thus obtain, from Assumption 2.6,

$$|\Delta_i^b(z, C_i; W(t))| \leq K_T(1+B),$$

for $2 \leq i \leq L$ and $\mathcal{P}$-almost every z. This implies

$$\text{ess-sup} \sup_{t\leq T}|b_i(t, C_i)| \leq \text{ess-sup}|b_i^0(C_i)| + K_T(1+B)T \leq K_T(1+B),$$

which shows ess-sup $\mathsf{max}_t^b(W) \leq K_T(1+B)$, as claimed.

Fact 3: *maximal bounds for* $\tilde{W}$. We also have on the event $\mathcal{E}$, almost surely,

$$\max_{2\leq i\leq L} \max_{j_{i-1}\in[n_{i-1}], j_i\in[n_i]} \sup_{t\leq T}|\tilde{w}_i(t, j_{i-1}, j_i)| \leq K_T(1+B),$$

$$\max_{2\leq i\leq L} \max_{j_i\in[n_i]} \sup_{t\leq T}|\tilde{b}_i(t, j_i)| \leq K_T(1+B).$$

A proof of this fact is similar to the argument for Fact 2. We note that this argument requires the use of the fact $|||\tilde{W}|||_T \leq K_T$, which holds on the event $\mathcal{E}$, and the application of Lemma 3.2. The latter application holds by noticing that $\tilde{W}$ can be viewed as a collection of MF parameter whose neuronal ensemble $(\Omega_{\text{new}}, P_{\text{new}}) = \prod_{i=1}^L (\Omega_{i,\text{new}}, P_{i,\text{new}})$ takes the following specific form: $\Omega_{i,\text{new}} = \{C_i(1), \ldots, C_i(n_i)\}$ and $P_{i,\text{new}}$ is a uniform probability measure on $\Omega_{i,\text{new}}$.

We now decompose the proof into several steps.

Step 1 – Main proof. Let us first define some quantities that represent the difference between W and $\tilde{W}$:[4]

[4]To simplify our notation, here and in the following argument, we denote by ∂_1 the partial derivative with respect to the first variable, so, for example, $\partial_1 w_i(t, C_{i-1}(j_{i-1}), C_i(j_i)) = \frac{\partial}{\partial t} w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))$.

$$D_1^w(t) = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1} |\partial_1\tilde{w}_1(t,j_1) - \partial_1 w_1(t, C_1(j_1))|^2\Big)^{1/2},$$

$$D_i^w(t) = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i} |\partial_1\tilde{w}_i(t,j_{i-1},j_i) - \partial_1 w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))|^2\Big)^{1/2},$$

$$D_i^b(t) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} |\partial_1\tilde{b}_i(t,j_i) - \partial_1 b_i(t, C_i(j_i))|^2\Big)^{1/2}, \quad 2 \le i \le L.$$

We are also interested in the following quantities that represent the smoothness in the time evolution of $W(t)$ and $\tilde{W}(t)$:

$$A_1^w(t,\zeta) = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1} |\partial_1 w_1(t+\zeta, C_1(j_1)) - \partial_1 w_1(t, C_1(j_1))|^2\Big)^{1/2},$$

$$\tilde{A}_1^w(t,\zeta) = \Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1} |\partial_1\tilde{w}_1(t+\zeta, j_1) - \partial_1\tilde{w}_1(t, j_1)|^2\Big)^{1/2},$$

$$A_i^w(t,\zeta) = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i} |\partial_1 w_i(t+\zeta, C_{i-1}(j_{i-1}), C_i(j_i)) - \partial_1 w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))\big|^2\Big)^{1/2},$$

$$\tilde{A}_i^w(t,\zeta) = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i} |\partial_1\tilde{w}_i(t+\zeta, j_{i-1}, j_i) - \partial_1\tilde{w}_i(t, j_{i-1}, j_i)|^2\Big)^{1/2},$$

$$A_i^b(t,\zeta) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} |\partial_1 b_i(t+\zeta, C_i(j_i)) - \partial_1 b_i(t, C_i(j_i))|^2\Big)^{1/2},$$

$$\tilde{A}_i^b(t,\zeta) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} |\partial_1\tilde{b}_i(t+\zeta, j_i) - \partial_1\tilde{b}_i(t, j_i)|^2\Big)^{1/2}, \quad 2 \le i \le L.$$

These quantities give a bound on $\mathcal{D}_t(W, \tilde{W})$:

$$\mathcal{D}_t(W,\tilde{W}) \le K\int_0^t \max_{1\le i\le L} D_i^w(\lfloor s/\zeta\rfloor\zeta)\,ds + K\int_0^t \max_{2\le i\le L} D_i^b(\lfloor s/\zeta\rfloor\zeta)\,ds + Kt \sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}\max_i(A_i^w(s,\zeta'), A_i^b(s,\zeta'), \tilde{A}_i^w(s,\zeta'), \tilde{A}_i^b(s,\zeta')),$$

where we have used the fact $\mathcal{D}_0(W, \tilde{W}) = 0$. The next task is to bound the terms inside the integral.

To find bounds on $D_i^w(t)$, we introduce the following quantities for $1 \le i \le L$:

$$G_i(t) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} \mathbb{E}_Z[|\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(t)) - \Delta_i^H(Z, C_i(j_i); W(t))|]^2\Big)^{1/2}$$

and

$$F_i(t) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} \mathbb{E}_Z[|\mathbf{H}_i(X, j_i; \tilde{W}(t)) - H_i(X, C_i(j_i); W(t))|]^2\Big)^{1/2}.$$

We specify their connection in the following. By Assumptions 2.4 and 2.6, for $i \geq 2$,

$$D_i^w(t) \leq K(D_i^{w,1}(t) + G_i(t) + \mathcal{D}_t(W, \tilde{W}) + F_i(t)),$$

in which

$$D_i^{w,1}(t) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1 + |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(t))|^2 + |\Delta_i^{H}(Z, C_i(j_i); W(t))|^2)$$
$$\times \frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}} \mathbb{E}_Z[|\mathbf{H}_{i-1}(X, j_{i-1}; \tilde{W}(t)) - H_{i-1}(X, C_{i-1}(j_{i-1}); W(t))|]^2\Big)^{1/2}.$$

By Lemma C.3, on the event $\mathcal{E}$,

$$D_i^{w,1}(t) \leq K_T F_{i-1}(t).$$

As such, on the event $\mathcal{E}$,

$$D_i^w(t) \leq K_T(F_{i-1}(t) + G_i(t) + \mathcal{D}_t(W, \tilde{W}) + F_i(t)).$$

Similarly, we also have

$$D_1^w(t) \leq K(G_1(t) + \mathcal{D}_t(W, \tilde{W})).$$

Together with the previously derived bound on $\mathcal{D}_t(W, \tilde{W})$, we obtain, on the event $\mathcal{E}$,

$$\begin{aligned}\mathcal{D}_t(W, \tilde{W}) \leq\ & K_T\int_0^t \mathcal{D}_s(W, \tilde{W})\,ds + K_T\int_0^t \max_i(G_i(\lfloor s/\zeta\rfloor\zeta) + F_i(\lfloor s/\zeta\rfloor\zeta))\,ds \\ & + K\int_0^t \max_{2\leq i\leq L} D_i^b(\lfloor s/\zeta\rfloor\zeta)\,ds \\ & + Kt \sup_{s\leq T-\zeta}\sup_{0\leq\zeta'\leq\zeta}\max_i(A_i^w(s,\zeta'), A_i^b(s,\zeta'), \tilde{A}_i^w(s,\zeta'), \tilde{A}_i^b(s,\zeta')),\end{aligned}$$

which holds for all $t \leq T$.

Next we make the following claims:

- *Claim 1*: For any $\zeta \in [0, T]$, on the event $\mathcal{E}$, almost surely,

$$\sup_{t\leq T-\zeta}\sup_{0\leq\zeta'\leq\zeta}\max_i(A_i^w(t,\zeta'), A_i^b(t,\zeta'), \tilde{A}_i^w(t,\zeta'), \tilde{A}_i^b(t,\zeta')) \leq K_T(1+B)\zeta.$$

- *Claim 2*: For a sequence $\{\gamma_j > 0, j = 2, \dots, L\}$ and $t \le T$, let $\mathcal{E}^{\mathbf{H}}_{t,i}$ denote the event in which for all $k \in \{1, 2, \dots, i\}$,

$$F_k(t) \le K_T^k \Big(\mathcal{D}_t(W, \tilde{W}) + (1 + B) \sum_{j=1}^{k-1} \gamma_{j+1} \Big).$$

(The summation $\sum_{j=1}^{k-1}$ equals 0 if $k = 1$.) We claim that for each $i = 1, \dots, L$,

$$\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,i}; \mathcal{E}) \ge 1 - \sum_{j=1}^{i-1} \frac{n_{j+1}}{\gamma_{j+1}} \exp\Big(-\frac{n_j \gamma_{j+1}^2}{K_T} \Big).$$

- *Claim 3*: For a sequence $\{\beta_j > 0, j = 1, \dots, L-2\}$ and $t \le T$, let $\mathcal{E}^{\Delta}_{t,i}$ denote the event that for all $k \in \{i, i+1, \dots, L\}$,

$$G_k(t) \le K_T^{2L-k+1} \bigg((1 + B) \mathcal{D}_t(W, \tilde{W}) + (1 + B^2) \Big(\delta_L^{\Delta} + \sum_{j=k}^{L-2} \beta_j \Big) \bigg),$$

where $\delta_L^{\Delta} = \sum_{j=1}^{L-1} \gamma_{j+1}$. (The summation $\sum_{j=k}^{L-2}$ equals 0 if $k \ge L - 1$.) We claim that for each $i = 1, \dots, L$,

$$\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L} \cap \mathcal{E}^{\Delta}_{t,i}; \mathcal{E}) \ge \mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L}; \mathcal{E}) - \sum_{j=i}^{L-2} \frac{n_j}{\beta_j} \exp\Big(-\frac{n_{j+1} \beta_j^2}{K_T} \Big).$$

- *Claim 4*: For $t \le T$, let $\mathcal{E}^b_t$ denote the event that for all $k \in \{2, \dots, L\}$,

$$D_k^b(t) \le K_T \big((1 + B) \mathcal{D}_t(W, \tilde{W}) + (1 + B^2) \delta_L^b \big),$$

where $\delta_L^b = \delta_L^{\Delta} + \sum_{j=1}^{L-2} \beta_j$. We claim that

$$\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L} \cap \mathcal{E}^{\Delta}_{t,1} \cap \mathcal{E}^b_t; \mathcal{E}) \ge \mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L} \cap \mathcal{E}^{\Delta}_{t,1}; \mathcal{E}) - \sum_{j=1}^{i-1} \frac{n_{j+1}}{\gamma_{j+1}} \exp\Big(-\frac{n_j \gamma_{j+1}^2}{K_T} \Big).$$

Let us assume these claims. Using the bounds on $\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L}; \mathcal{E})$, $\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L} \cap \mathcal{E}^{\Delta}_{t,1}; \mathcal{E})$ and $\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L} \cap \mathcal{E}^{\Delta}_{t,1} \cap \mathcal{E}^b_t; \mathcal{E})$, combining the previous bound, applying the union bound over $t \in \{0, \zeta, 2\zeta, \dots, \lfloor T/\zeta \rfloor \zeta\}$ and recalling $\mathcal{D}_0(W, \tilde{W}) = 0$, we then get

$$\begin{aligned} \mathcal{D}_t(W, \tilde{W}) \le K_T \int_0^t \Big[& (1 + B) \mathcal{D}_s(W, \tilde{W}) \\ & + (1 + B^2) \Big(\sum_{j=1}^{L-1} \gamma_{j+1} + \sum_{j=1}^{L-2} \beta_j \Big) + (1 + B) \zeta \Big] ds, \end{aligned}$$

for all $t \le T$, with probability at least

$$1 - \frac{T}{\zeta}\Big(\sum_{j=1}^{L-1}\frac{n_{j+1}}{\gamma_{j+1}}\exp\Big(-\frac{n_j\gamma_{j+1}^2}{K_T}\Big) + \sum_{j=1}^{L-2}\frac{n_j}{\beta_j}\exp\Big(-\frac{n_{j+1}\beta_j^2}{K_T}\Big)\Big) - KLn_{\max}\exp(-Kn_{\min}^{1/52}),$$

for any $\zeta \in [0, T]$. By Gronwall's lemma, the above implies that for all $t \le T$,

$$\mathcal{D}_t(W, \tilde{W}) \le K_T\Big[(1+B^2)\Big(\sum_{j=1}^{L-1}\gamma_{j+1} + \sum_{j=1}^{L-2}\beta_j\Big) + (1+B)\zeta\Big]\exp(K_T(1+B)T)$$
$$\le K_T\Big(\sum_{j=1}^{L-1}\gamma_{j+1} + \sum_{j=1}^{L-2}\beta_j + \zeta\Big)\exp(K_T(1+B)).$$

The proposition statement is then easily obtained by choosing

$$\gamma_{j+1} = \sqrt{\frac{1}{K_T n_j}\log\Big(\frac{2TLn_{\max}^2}{\delta} + e\Big)}, \qquad j = 1, \dots, L-1,$$
$$\beta_j = \sqrt{\frac{1}{K_T n_{j+1}}\log\Big(\frac{2TLn_{\max}^2}{\delta} + e\Big)}, \quad j = 1, \dots, L-2,$$

and

$$\zeta = \frac{1}{\sqrt{n_{\max}}}.$$

We are left with verifying the claims.

Step 2 – Claim 1. We first note that, by Assumptions 2.4 and 2.6, Lemma C.3, and the fact $|||W|||_0, |||W|||_{\text{samp},0} \le K$, on the event $\mathcal{E}$:

$$\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\le t}|\partial_1 w_i(s, C_{i-1}(j_{i-1}), C_i(j_i))|^{50}\Big)^{1/50}$$
$$\le K + K\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le t}\operatorname*{ess\,sup}_{Z\sim\mathcal{P}}|\Delta_i^H(Z, C_i(j_i); W(s))|^{50}\Big)^{1/50} \le K_T,$$

for any $t \le T$. Therefore,

$$\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}\big|w_i(s+\zeta', C_{i-1}(j_{i-1}), C_i(j_i)) - w_i(s, C_{i-1}(j_{i-1}), C_i(j_i))\big|^2\Big)^{1/2} \le K_T\zeta.$$

Similarly, we also have that on the event $\mathcal{E}$,

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_{C_{i-1}}\Big[\sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}\big|w_i(s+\zeta',C_{i-1},C_i(j_i)) - w_i(s,C_{i-1},C_i(j_i))\big|^2\Big]\Big)^{1/2} \le K_T\zeta,$$

$$\Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_{C_i}\Big[\sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}\big|w_i(s+\zeta',C_{i-1}(j_{i-1}),C_i) - w_i(s,C_{i-1}(j_{i-1}),C_i)\big|^2\Big]\Big)^{1/2} \le K_T\zeta,$$

$$\mathbb{E}\Big[\sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}|w_i(s+\zeta',C_{i-1},C_i)-w_i(s,C_{i-1},C_i)|^2\Big]^{1/2} \le K_T\zeta,$$

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}|b_i(s+\zeta',C_i(j_i))-b_i(s,C_i(j_i))|^2\Big)^{1/2} \le K_T\zeta,$$

$$\mathbb{E}\Big[\sup_{s\le T-\zeta}\sup_{0\le\zeta'\le\zeta}|b_i(s+\zeta',C_i)-b_i(s,C_i)|^2\Big]^{1/2} \le K_T\zeta,$$

$$\Big(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\sup_{s\le t}\sup_{0\le\zeta'\le\zeta}|w_1(s+\zeta',C_1(j_1))-w_1(s,C_1(j_1))|^2\Big)^{1/2} \le K_T\zeta,$$

$$\mathbb{E}\Big[\sup_{s\le t}\sup_{0\le\zeta'\le\zeta}|w_1(s+\zeta',C_1)-w_1(s,C_1)|^2\Big]^{1/2} \le K_T\zeta.$$

Together with Lemma B.4, this fact gives us a bound on $A_i^w(t,\zeta)$. In particular, defining $W_\zeta(t) = W(t+\zeta)$, we apply Lemma B.4 to the two MF parameter collections W and W_ζ along with the new random variable $\tilde{C}_i$ that is drawn uniformly from the set $\{C_i(1),\dots,C_i(n_i)\}$. Recalling the metric $\tilde{d}_t(W,W_\zeta)$ in this lemma, the above fact shows that $\tilde{d}_{T-\zeta}(W,W_\zeta) \le K_T\zeta$ on the event $\mathcal{E}$. The lemma holds due to Fact 1 and Fact 2. The conclusion of the lemma then reads

$$\sup_{t\le T-\zeta}\sup_{0\le\zeta'\le\zeta}\max_i(A_i^w(t,\zeta'),A_i^b(t,\zeta')) \le K_T(1+B)\,\tilde{d}_{T-\zeta}(W,W_\zeta) \le K_T(1+B)\zeta,$$

almost surely on the event $\mathcal{E}$.

By a similar argument, we have, almost surely on the event $\mathcal{E}$,

$$\sup_{t\le T-\zeta}\sup_{0\le\zeta'\le\zeta}\max_i(\tilde{A}_i^w(t,\zeta'),\tilde{A}_i^b(t,\zeta')) \le K_T(1+B)\zeta.$$

Indeed, one can repeat the argument by noticing that $\tilde{W}$ can be viewed as a collection of MF parameters whose neuronal ensemble $(\Omega_{\text{new}},P_{\text{new}}) = \prod_{i=1}^L(\Omega_{i,\text{new}},P_{i,\text{new}})$ takes the following specific form: $\Omega_{i,\text{new}} = \{C_i(1),\dots,C_i(n_i)\}$ and $P_{i,\text{new}}$ is a uniform probability measure on $\Omega_{i,\text{new}}$.

Step 3 – Claim 2. We show the claim by induction. Consider F_1:

$$|\mathbf{H}_1(x, j_1; \tilde{W}(t)) - H_1(x, C_i(j_i); W(t))| = |\phi_1(\tilde{w}_1(t, j_1), x) - \phi_1(w_1(t, C_1(j_1)), x)| \\ \leq K\mathcal{D}_t(W, \tilde{W})$$

for $\mathcal{P}$-almost every x by Assumption 2.5, and therefore

$$F_1(t) \leq K\mathcal{D}_t(W, \tilde{W}).$$

That is, $\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,1}) = 1$.

Now let us assume the claim for F_{i-1} with $i \geq 2$ and consider the claim for F_i. We have the following decomposition:

$$\begin{aligned}
&|\mathbf{H}_i(X, j_i; \tilde{W}(t)) - H_i(X, C_i(j_i); W(t))| \\
&\quad = \Big| \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \phi_i\big(\tilde{w}_i(t, j_{i-1}, j_i), \tilde{b}_i(t, j_i), \mathbf{H}_{i-1}(X, j_{i-1}; \tilde{W}(t))\big) \\
&\qquad - \mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i(t, C_{i-1}, C_i(j_i)), b_i(t, C_i(j_i)), H_{i-1}(X, C_{i-1}; W(t))\big)\big] \Big| \\
&\quad \leq Q_{1,i}(t) + Q_{2,i}(t),
\end{aligned}$$

which gives

$$F_i(t) \leq \Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \mathbb{E}_Z[|Q_{1,i}(s)| + |Q_{2,i}(s)|]^2 \Big)^{1/2},$$

where we define

$$\begin{aligned}
Q_{1,i}(t) &= \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \Big| \phi_i\big(\tilde{w}_i(t, j_{i-1}, j_i), \tilde{b}_i(t, j_i), \mathbf{H}_{i-1}(X, j_{i-1}; \tilde{W}(t))\big) \\
&\qquad - \phi_i\big(w_i(t, C_{i-1}(j_{i-1}), C_i(j_i)), b_i(t, C_i(j_i)), H_{i-1}(X, C_{i-1}(j_{i-1}); W(t))\big) \Big|, \\
Q_{2,i}(t) &= \Big| \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \phi_i\big(w_i(t, C_{i-1}(j_{i-1}), C_i(j_i)), b_i(t, C_i(j_i)), \\
&\qquad H_{i-1}(X, C_{i-1}(j_{i-1}); W(t))\big) \\
&\qquad - \mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i(t, C_{i-1}, C_i(j_i)), b_i(t, C_i(j_i)), H_{i-1}(X, C_{i-1}; W(t))\big)\big] \Big|.
\end{aligned}$$

By Assumption 2.5 and Cauchy–Schwarz's inequality, we obtain a bound on $Q_{1,i}$:

$$\begin{aligned}
\mathbb{E}_Z[|Q_{1,i}(t)|]^2 \leq \frac{K}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} &\big(1 + |\tilde{w}_i(t, j_{i-1}, j_i)|^2 \\
&+ |w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))|^2 + |\tilde{b}_i(t, j_i)|^2 + |b_i(t, C_i(j_i))|^2\big)
\end{aligned}$$

$$\times \frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \mathbb{E}_Z[|\mathbf{H}_{i-1}(X, j_{i-1}; \tilde{W}(t)) - H_{i-1}(X, C_{i-1}(j_{i-1}); W(t))|]^2$$
$$+ \frac{K}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} |\tilde{w}_i(t, j_{i-1}, j_i) - w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))|^2$$
$$+ K|\tilde{b}_i(t, j_i) - b_i(t, C_i(j_i))|^2,$$

and therefore, by Fact 1, under the events $\mathcal{E}^{\mathbf{H}}_{t,i-1}$ and $\mathcal{E}$,

$$\Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \mathbb{E}_Z[|Q_{1,i}(t)|]^2\Big)^{1/2} \le K_T F_{i-1}(t) + K\mathcal{D}_t(W, \tilde{W}).$$

Let us bound $Q_{2,i}$. For brevity, let us write

$$Z_i^H(t, c_{i-1}, c_i) = \phi_i\big(w_i(t, c_{i-1}, c_i), b_i(t, c_i), H_{i-1}(x, c_{i-1}; W(t))\big).$$

Recall that $C_{i-1}(j_{i-1})$ and $C_i(j_i)$ are independent. We thus have

$$\mathbb{E}[Z_i^H(t, C_{i-1}(j_{i-1}), C_i(j_i)) \mid C_i(j_i)] = \mathbb{E}_{C_{i-1}}[Z_i^H(t, C_{i-1}, C_i(j_i))].$$

Furthermore, $\{C_{i-1}(j_{i-1})\}_{j_{i-1}\in[n_{i-1}]}$ are η_{i-1}-independent by Assumption 4.4. We also have that for $\mathcal{P}$-almost every x, almost surely,

$$|Z_i^H(t, C_{i-1}(j_{i-1}), C_i(j_i))| \le K_T(1 + B),$$

by Assumption 2.5 and Fact 2. Then, by Lemma C.1, noting that $\gamma_i \ge K\eta_{i-1}$,

$$\mathbb{P}\big(\mathbb{E}_Z[Q_{2,i}] \ge K_T(1 + B)\gamma_i\big) \le \frac{1}{\gamma_i} \exp\Big(-\frac{n_{i-1}\gamma_i^2}{K_T}\Big).$$

By taking a union bound of the above probabilistic bound over $j_i \in [n_i]$, we thus have, on the events $\mathcal{E}^{\mathbf{H}}_{t,i-1}$ and $\mathcal{E}$,

$$F_i(t) \le K_T F_{i-1}(t) + K\mathcal{D}_t(W, \tilde{W}) + K_T(1 + B)\gamma_i$$
$$\le K_T^i\Big(\mathcal{D}_t(W, \tilde{W}) + (1 + B)\sum_{j=1}^{i-1} \gamma_{j+1}\Big),$$

with probability at least $1 - (n_i/\gamma_i)\exp(-n_{i-1}\gamma_i^2/K_T)$. We thus get

$$\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,i}; \mathcal{E}) \ge \mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,i-1}; \mathcal{E}) - \frac{n_j}{\gamma_j} \exp\Big(-\frac{n_{i-1}\gamma_i^2}{K_T}\Big) \ge 1 - \sum_{j=1}^{i-1} \frac{n_{j+1}}{\gamma_{j+1}} \exp\Big(-\frac{n_j\gamma_{j+1}^2}{K_T}\Big),$$

which proves the claim.

Step 4 – Claim 3. We show the claim by backward induction. The proof is similar to Claim 2. Consider $i = L$. Notice that on the event $\mathcal{E}^{\mathbf{H}}_{t,L}$,

$$\mathbb{E}_Z[|\hat{\mathbf{y}}(X;\tilde{W}(t)) - \hat{y}(X;W(t))|] \leq KF_L(t) \leq K_T^L(\mathcal{D}_t(W,\tilde{W}) + (1+B)\delta_L^{\Delta}),$$

by Assumption 2.5. We thus get from Assumption 2.6 that on the events $\mathcal{E}^{\mathbf{H}}_{t,L}$ and $\mathcal{E}$,

$$\begin{aligned} G_L(t) &\leq K(F_L(t) + \mathbb{E}_Z[|\hat{\mathbf{y}}(X;\tilde{W}(t)) - \hat{y}(X;W(t))|]) \\ &\leq K_T^{L+1}(\mathcal{D}_t(W,\tilde{W}) + (1+B)\delta_L^{\Delta}). \end{aligned}$$

That is, $\mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L} \cap \mathcal{E}^{\Delta}_{t,L}; \mathcal{E}) = \mathbb{P}(\mathcal{E}^{\mathbf{H}}_{t,L}; \mathcal{E})$.

Considering $i = L-1$, by Assumption 2.6, we have

$$G_{L-1}(t) \leq K\big(G^{(1)}_{L-1}(t) + G^{(2)}_{L-1}(t) + G^{(3)}_{L-1}(t) + G^{(4)}_{L-1}(t)\big),$$

in which

$$\begin{aligned} G^{(1)}_{L-1}(t) &= \Big(\frac{1}{n_{L-1}} \sum_{j_{L-1}=1}^{n_{L-1}} 1 + |\tilde{w}_L(t, j_{L-1}, 1)|^2 + |w_L(t, C_{L-1}(j_{L-1}), 1)|^2 \\ &\quad + |\tilde{b}_L(t,1)|^2 + |b_L(t,1)|^2\Big)^{1/2} G_L(t), \\ G^{(2)}_{L-1}(t) &= \big(1 + \mathbb{E}_Z[|\Delta^{\mathbf{H}}_L(Z,1;\tilde{W}(t))|] + \mathbb{E}_Z[|\Delta^{H}_L(Z,1;W(t))|]\big) \\ &\quad \times \Big(\frac{1}{n_{L-1}} \sum_{j_{L-1}=1}^{n_{L-1}} |\tilde{w}_L(t, j_{L-1}, 1) - w_L(t, C_{L-1}(j_{L-1}), 1)|^2 \\ &\quad + |\tilde{b}_L(t,1) - b_L(t,1)|^2\Big)^{1/2}, \\ G^{(3)}_{L-1}(t) &= \Big(\frac{1}{n_{L-1}} \sum_{j_{L-1}=1}^{n_{L-1}} \mathbb{E}_Z\big[(1 + |\Delta^{\mathbf{H}}_L(Z,1;\tilde{W}(t))| + |\Delta^{H}_L(Z,1;W(t))|) \\ &\quad \times \big(1 + |\tilde{w}_L(t, j_{L-1}, 1)| + |w_L(t, C_{L-1}(j_{L-1}), 1)| + |\tilde{b}_L(t,1)| + |b_L(t,1)|\big) \\ &\quad \times |\mathbf{H}_L(X,1;\tilde{W}(t)) - H_L(X,1;W(t))|\big]^2\Big)^{1/2}, \\ G^{(4)}_{L-1}(t) &= \Big(\frac{1}{n_{L-1}} \sum_{j_{L-1}=1}^{n_{L-1}} \mathbb{E}_Z\big[(1 + |\Delta^{\mathbf{H}}_L(Z,1;\tilde{W}(t))| + |\Delta^{H}_L(Z,1;W(t))|) \\ &\quad \times \big(1 + |\tilde{w}_L(t, j_{L-1}, 1)| + |w_L(t, C_{L-1}(j_{L-1}), 1)| + |\tilde{b}_L(t,1)| + |b_L(t,1)|\big) \\ &\quad \times |\mathbf{H}_{L-1}(X, j_{L-1};\tilde{W}(t)) - H_L(X, C_{L-1}(j_{L-1});W(t))|\big]^2\Big)^{1/2}. \end{aligned}$$

Due to Fact 1, on the event $\mathcal{E}$,

$$G^{(1)}_{L-1}(t) \leq K_T G_L(t).$$

By Assumption 2.6, we have, for $\mathcal{P}$-almost every z,

$$|\Delta_L^{\mathbf{H}}(z,1;\tilde{W}(t))|, |\Delta_L^{H}(z,1;W(t))| \le K.$$

Using this fact,

$$G_{L-1}^{(2)}(t) \le K\mathcal{D}_t(W,\tilde{W}).$$

The same fact also applies to $G_{L-1}^{(3)}$, $G_{L-1}^{(4)}$ and $G_{L-1}^{(5)}$. In particular, we obtain for $G_{L-1}^{(3)}$, on the event $\mathcal{E}$,

$$\begin{aligned} G_{L-1}^{(3)}(t) &\le K\Big(\frac{1}{n_{L-1}}\sum_{j_{L-1}=1}^{n_{L-1}}\big(1+|\tilde{w}_L(t,j_{L-1},1)| \\ &\qquad + |w_L(t,C_{L-1}(j_{L-1}),1)| + |\tilde{b}_L(t,1)| + |b_L(t,1)|\big)^2\Big)^{1/2} F_L(t) \\ &\le K_T F_L(t), \end{aligned}$$

where the last display follows from Fact 1. Similarly, by using Fact 2 and Fact 3, we have on the event $\mathcal{E}$,

$$G_{L-1}^{(4)}(t) \le K_T(1+B)F_{L-1}(t).$$

Hence, on the events $\mathcal{E}_{t,L}^{\mathbf{H}}$, $\mathcal{E}_{t,L}^{\Delta}$ and $\mathcal{E}$,

$$\begin{aligned} G_{L-1}(t) &\le K_T\big(G_L(t) + \mathcal{D}_t(W,\tilde{W}) + F_L(t) + (1+B)F_{L-1}(t)\big) \\ &\le K_T^{L+2}\big((1+B)\mathcal{D}_t(W,\tilde{W}) + (1+B^2)\delta_L^{\Delta}\big). \end{aligned}$$

In other words, $\mathbb{P}(\mathcal{E}_{t,L}^{\mathbf{H}} \cap \mathcal{E}_{t,L-1}^{\Delta}; \mathcal{E}) = \mathbb{P}(\mathcal{E}_{t,L}^{\mathbf{H}}; \mathcal{E})$.

Next let us assume the claim for i, and consider the claim for $i-1$, for $2 \le i \le L-1$. For notational brevity, in the following, we let

$$\begin{aligned} \Delta_i^{\mathbf{H}}(j_i) &= \Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(t)), \quad \mathbf{H}_i(j_i) = \mathbf{H}_i(X,j_i;\tilde{W}(t)), \\ \Delta_i^{H}(c_i) &= \Delta_i^{H}(Z,c_i;W(t)), \quad H_i(c_i) = H_i(X,c_i;W(t)). \end{aligned}$$

We have

$$\begin{aligned} &|\Delta_{i-1}^{\mathbf{H}}(j_{i-1}) - \Delta_{i-1}^{H}(C_{i-1}(j_{i-1}))| \\ &= \Big|\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{\mathbf{H}}(j_i),\tilde{w}_i(t,j_{i-1},j_i),\tilde{b}_i(t,j_i),\mathbf{H}_i(j_i),\mathbf{H}_{i-1}(j_{i-1})\big) \\ &\quad - \mathbb{E}_{C_i}\big[\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{H}(C_i),w_i(t,C_{i-1}(j_{i-1}),C_i),b_i(t,C_i),H_i(C_i),H_{i-1}(C_{i-1}(j_{i-1}))\big)\big]\Big| \\ &\le Q_{3,i}(t) + Q_{4,i}(t), \end{aligned}$$

which gives

$$G_{i-1}(t) \le \Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_Z[|Q_{3,i}(t)| + |Q_{4,i}(t)|]^2\Big)^{1/2},$$

in which we define

$$
\begin{aligned}
Q_{3,i}(t) = \frac{1}{n_i}\sum_{j_i=1}^{n_i}\big|&\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{\mathbf{H}}(j_i),\tilde{w}_i(t,j_{i-1},j_i),\tilde{b}_i(t,j_i),\mathbf{H}_i(j_i),\mathbf{H}_{i-1}(j_{i-1})\big)\\
&-\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{H}(C_i(j_i)),w_i(t,C_{i-1}(j_{i-1}),C_i(j_i)),b_i(t,C_i(j_i)),\\
&H_i(C_i(j_i)),H_{i-1}(C_{i-1}(j_{i-1}))\big)\big|,\\
Q_{4,i}(t) = \Big|\frac{1}{n_i}\sum_{j_i=1}^{n_i}&\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{H}(C_i(j_i)),w_i(t,C_{i-1}(j_{i-1}),C_i(j_i)),b_i(t,C_i(j_i)),\\
&H_i(C_i(j_i)),H_{i-1}(C_{i-1}(j_{i-1}))\big)\\
&-\mathbb{E}_{C_i}\big[\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^{H}(C_i),w_i(t,C_{i-1}(j_{i-1}),C_i),b_i(t,C_i),\\
&H_i(C_i),H_{i-1}(C_{i-1}(j_{i-1}))\big)\big]\Big|.
\end{aligned}
$$

Let us first bound $Q_{3,i}$. This is similar to the bounding of G_{L-1}. In particular, by Assumption 2.6, we have

$$
\Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_Z[|Q_{3,i}(t)|]^2\Big)^{1/2} \le K\big(Q_{3,i}^{(1)}(t)+Q_{3,i}^{(2)}(t)+Q_{3,i}^{(3)}(t)+Q_{3,i}^{(4)}(t)\big),
$$

in which

$$
\begin{aligned}
Q_{3,i}^{(1)}(t) &= \Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\big(1+|\tilde{w}_i(t,j_{i-1},j_i)|+|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|\\
&\quad+|\tilde{b}_i(t,j_i)|+|b_i(t,C_i(j_i))|\big)\mathbb{E}_Z[|\Delta_i^{\mathbf{H}}(j_i)-\Delta_i^{H}(C_i(j_i))|]\Big)^2\Big)^{1/2},\\
Q_{3,i}^{(2)}(t) &= \Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[1+|\Delta_i^{\mathbf{H}}(j_i)|+|\Delta_i^{H}(C_i(j_i))|]\\
&\quad\times\big(|\tilde{w}_i(t,j_{i-1},j_i)-w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|+|\tilde{b}_i(t,j_i)-b_i(t,C_i(j_i))|\big)\Big)^2\Big)^{1/2},\\
Q_{3,i}^{(3)}(t) &= \Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z\big[(1+|\Delta_i^{\mathbf{H}}(j_i)|+|\Delta_i^{H}(C_i(j_i))|)\\
&\quad\times\big(1+|\tilde{w}_i(t,j_{i-1},j_i)|+|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|+|\tilde{b}_i(t,j_i)|+|b_i(t,C_i(j_i))|\big)\\
&\qquad\times|\mathbf{H}_i(j_i)-H_i(C_i(j_i))|\big]\Big)^2\Big)^{1/2},\\
Q_{3,i}^{(4)}(t) &= \Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z\big[(1+|\Delta_i^{\mathbf{H}}(j_i)|+|\Delta_i^{H}(C_i(j_i))|)\\
&\quad\times\big(1+|\tilde{w}_i(t,j_{i-1},j_i)|+|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|+|\tilde{b}_i(t,j_i)|+|b_i(t,C_i(j_i))|\big)\\
&\qquad\times|\mathbf{H}_{i-1}(j_{i-1})-H_{i-1}(C_{i-1}(j_{i-1}))|\big]\Big)^2\Big)^{1/2}.
\end{aligned}
$$

To bound $Q_{3,i}^{(1)}$, we use Cauchy–Schwarz's inequality and Fact 1 to obtain that, on the event $\mathcal{E}$,

$$\begin{aligned} Q_{3,i}^{(1)}(t) &\le \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\big(1+|\tilde{w}_i(t,j_{i-1},j_i)|+|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))| \\ &\quad +|\tilde{b}_i(t,j_i)|+|b_i(t,C_i(j_i))|\big)^2\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|\Delta_i^{\mathbf{H}}(j_i)-\Delta_i^H(C_i(j_i))|]^2\Big)^{1/2} \\ &\le K_T G_i(t). \end{aligned}$$

By Cauchy–Schwarz's inequality and Fact 1, we have the following bound on $Q_{3,i}^{(2)}$ on the event $\mathcal{E}$:

$$\begin{aligned} Q_{3,i}^{(2)}(t) &\le K\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[1+|\Delta_i^{\mathbf{H}}(j_i)|^2+|\Delta_i^H(C_i(j_i))|^2]\Big)^{1/2} \\ &\quad\times\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\tilde{w}_i(t,j_{i-1},j_i)-w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|^2 \\ &\quad+|\tilde{b}_i(t,j_i)-b_i(t,C_i(j_i))|^2\Big)^{1/2} \\ &\le K_T\mathcal{D}_t(W,\tilde{W}). \end{aligned}$$

Similarly, on the event $\mathcal{E}$,

$$\begin{aligned} Q_{3,i}^{(3)}(t) &\overset{(a)}{\le} K_T(1+B)\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z\big[(1+|\Delta_i^{\mathbf{H}}(j_i)|+|\Delta_i^H(C_i(j_i))|) \\ &\quad\times|\mathbf{H}_i(j_i)-H_i(C_i(j_i))|\big] \\ &\overset{(b)}{\le} K_T(1+B)\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[1+|\Delta_i^{\mathbf{H}}(j_i)|^2+|\Delta_i^H(C_i(j_i))|^2]\Big)^{1/2}F_i(t) \\ &\overset{(c)}{\le} K_T(1+B)F_i(t), \end{aligned}$$

where we use Fact 2 and Fact 3 in step (a), Cauchy–Schwarz's inequality in step (b) and Fact 1 in step (c). With the same argument, on the event $\mathcal{E}$,

$$\begin{aligned} Q_{3,i}^{(4)}(t) &\overset{(a)}{\le} K_T(1+B)\Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z\big[(1+|\Delta_i^{\mathbf{H}}(j_i)|+|\Delta_i^H(C_i(j_i))|) \\ &\quad\times|\mathbf{H}_{i-1}(j_{i-1})-H_{i-1}(C_{i-1}(j_{i-1}))|\big]\Big)^2\Big)^{1/2} \end{aligned}$$

$$\begin{aligned}
&\overset{\text{(b)}}{\leq} K_T(1+B)\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[1+|\Delta_i^{\mathbf{H}}(j_i)|^2+|\Delta_i^H(C_i(j_i))|^2] \\
&\qquad\times\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_Z[|\mathbf{H}_{i-1}(j_{i-1})-H_{i-1}(C_{i-1}(j_{i-1}))|^2]\Big)^{1/2} \\
&\overset{\text{(c)}}{\leq} K_T(1+B)F_{i-1}(t),
\end{aligned}$$

where again we use Fact 2 and Fact 3 in step (a), Cauchy–Schwarz's inequality in step (b) and Fact 1 in step (c). Therefore, on the events $\mathcal{E}$, $\mathcal{E}_{t,i}^{\Delta}$ and $\mathcal{E}_{t,L}^{\mathbf{H}}$,

$$\begin{aligned}
&\Big(\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\mathbb{E}_Z[|Q_{3,i}(t)||^2\Big)^{1/2} \\
&\quad\leq K_T\big(G_i(t)+\mathcal{D}_t(W,\tilde{W})+(1+B)(F_i(t)+F_{i-1}(t))\big) \\
&\quad\leq K_T^{2L-i+2}\Big((1+B)\mathcal{D}_t(W,\tilde{W})+(1+B^2)\Big(\delta_L^{\Delta}+\sum_{j=i}^{L-2}\beta_j\Big)\Big).
\end{aligned}$$

Next let us bound $Q_{4,i}$. For brevity, let us write

$$Z_i^{\Delta}(t,c_{i-1},c_i)=\sigma_{i-1}^{\mathbf{H}}\big(\Delta_i^H(c_i),w_i(t,c_{i-1},c_i),b_i(t,c_i),H_i(c_i),H_{i-1}(c_{i-1})\big).$$

Recall that $C_{i-1}(j_{i-1})$ and $C_i(j_i)$ are independent. We thus have

$$\mathbb{E}_{C_i(j_i)}[Z_i^{\Delta}(t,C_{i-1}(j_{i-1}),C_i(j_i))\mid C_{i-1}(j_{i-1})]=\mathbb{E}_{C_i}[Z_i^{\Delta}(t,C_{i-1}(j_{i-1}),C_i)].$$

Furthermore, $\{C_i(j_i)\}_{j_i\in[n_i]}$ are η_i-independent by Assumption 4.4. We also have that, almost surely,

$$\begin{aligned}
&|Z_i^{\Delta}(t,C_{i-1}(j_{i-1}),C_i(j_i))| \\
&\quad\leq K(1+|\Delta_i^H(C_i(j_i))|)(1+|w_i(t,C_{i-1}(j_{i-1}),C_i(j_i))|+|b_i(t,C_i(j_i))|) \\
&\quad\leq K_T(1+B^2),
\end{aligned}$$

by Assumption 2.6 and Fact 2. Then, by Lemma C.1, noting that $\beta_{i-1}\geq K\eta_i$,

$$\mathbb{P}\big(\mathbb{E}_Z[Q_{4,i}(t)]\geq K_T(1+B^2)\beta_{i-1}\big)\leq\frac{1}{\beta_{i-1}}\exp\Big(-\frac{n_i\beta_{i-1}^2}{K_T}\Big).$$

We thus have, by taking a union bound over $j_{i-1}\in[n_{i-1}]$, on the events $\mathcal{E}$, $\mathcal{E}_{t,i}^{\Delta}$ and $\mathcal{E}_{t,L}^{\mathbf{H}}$,

$$G_{i-1}(t)\leq K_T^{2L-i+2}\Big((1+B)\mathcal{D}_t(W,\tilde{W})+(1+B^2)\Big(\delta_L^{\Delta}+\sum_{j=i-1}^{L-2}\beta_j\Big)\Big)$$

with probability at least $1-(n_{i-1}/\beta_{i-1})\exp(-n_i\beta_{i-1}^2/K_T)$. We thus get

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{t,L}^{\mathbf{H}}\cap\mathcal{E}_{t,i-1}^{\Delta};\mathcal{E}) &\geq \mathbb{P}(\mathcal{E}_{t,L}^{\mathbf{H}}\cap\mathcal{E}_{t,i}^{\Delta};\mathcal{E}) - \frac{n_{i-1}}{\beta_{i-1}}\exp\Big(-\frac{n_{i-1}\beta_{i-1}^2}{K_T}\Big)\\ &\geq \mathbb{P}(\mathcal{E}_{t,L}^{\mathbf{H}};\mathcal{E}) - \sum_{j=i-1}^{L-2}\frac{n_j}{\beta_j}\exp\Big(-\frac{n_{j+1}\beta_j^2}{K_T}\Big),\end{aligned}$$

which proves the claim.

Step 5 – Claim 4. We reuse the notations introduced in the previous step. For $2\leq i\leq L$, we have

$$\begin{aligned}&|\Delta_i^{\mathbf{b}}(Z,j_i;\tilde{W}(t)) - \Delta_i^{b}(Z,C_i(j_i);W(t))|\\ &\quad= \Big|\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\sigma_i^{\mathbf{b}}\big(\Delta_i^{\mathbf{H}}(j_i),\tilde{w}_i(t,j_{i-1},j_i),\tilde{b}_i(t,j_i),\mathbf{H}_i(j_i),\mathbf{H}_{i-1}(j_{i-1})\big)\\ &\qquad-\mathbb{E}_{C_{i-1}}\big[\sigma_i^{\mathbf{b}}\big(\Delta_i^{H}(C_i(j_i)),w_i(t,C_{i-1},C_i(j_i)),b_i(t,C_i(j_i)),\\ &\qquad\quad H_i(C_i(j_i)),H_{i-1}(C_{i-1})\big)\big]\Big|\\ &\quad\leq Q_{5,i}(t)+Q_{6,i}(t),\end{aligned}$$

which gives, by Assumption 2.4,

$$D_i^b(t)\leq K\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|Q_{5,i}(t)|+|Q_{6,i}(t)|]^2\Big)^{1/2},$$

in which we define

$$\begin{aligned}Q_{5,i}(t) = \frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}&\big|\sigma_i^{\mathbf{b}}(\Delta_i^{\mathbf{H}}(j_i),\tilde{w}_i(t,j_{i-1},j_i),\tilde{b}_i(t,j_i),\mathbf{H}_i(j_i),\mathbf{H}_{i-1}(j_{i-1}))\\ &-\sigma_i^{\mathbf{b}}\big(\Delta_i^{H}(C_i(j_i)),w_i(t,C_{i-1}(j_{i-1}),C_i(j_i)),b_i(t,C_i(j_i)),\\ &\quad H_i(C_i(j_i)),H_{i-1}(C_{i-1}(j_{i-1}))\big)\big|\end{aligned}$$

and

$$\begin{aligned}Q_{6,i}(t) = \Big|\frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}&\sigma_i^{\mathbf{b}}\big(\Delta_i^{H}(C_i(j_i)),w_i(t,C_{i-1}(j_{i-1}),C_i(j_i)),b_i(t,C_i(j_i)),\\ &\quad H_i(C_i(j_i)),H_{i-1}(C_{i-1}(j_{i-1}))\big)\\ &-\mathbb{E}_{C_{i-1}}\big[\sigma_i^{\mathbf{b}}\big(\Delta_i^{H}(C_i(j_i)),w_i(t,C_{i-1},C_i(j_i)),b_i(t,C_i(j_i)),\\ &\quad H_i(C_i(j_i)),H_{i-1}(C_{i-1})\big)\big]\Big|.\end{aligned}$$

Similar to the bounding of $D_i^w(t)$, we have, on the event $\mathcal{E}$,

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|Q_{5,i}(t)|]^2\Big)^{1/2} \le K_T\big(F_{i-1}(t)+G_i(t)+\mathcal{D}_t(W,\tilde{W})+F_i(t)\big).$$

To bound $Q_{6,i}$, for brevity, let us write

$$Z_i^b(t,c_{i-1},c_i)=\sigma_i^{\mathbf{b}}\big(\Delta_i^H(c_i),w_i(t,c_{i-1},c_i),b_i(t,c_i),H_i(c_i),H_{i-1}(c_{i-1})\big).$$

Recall that $C_{i-1}(j_{i-1})$ and $C_i(j_i)$ are independent. We thus have

$$\mathbb{E}_{C_{i-1}(j_{i-1})}[Z_i^b(t,C_{i-1}(j_{i-1}),C_i(j_i))\mid C_i(j_i)]=\mathbb{E}_{C_{i-1}}[Z_i^b(t,C_{i-1},C_i(j_i))].$$

Furthermore, $\{C_{i-1}(j_{i-1})\}_{j_{i-1}\in[n_{i-1}]}$ are η_{i-1}-independent by Assumption 4.4. We also have that, almost surely,

$$|Z_i^b(t,C_{i-1}(j_{i-1}),C_i(j_i))|\le K(1+|\Delta_i^H(C_i(j_i))|)\le K_T(1+B),$$

by Assumption 2.6 and Fact 2. Then, by Lemma C.1, and since $\gamma_i \ge K\eta_{i-1}$,

$$\mathbb{P}\big(\mathbb{E}_Z[Q_{6,i}]\ge K_T(1+B)\gamma_i\big)\le\frac{1}{\gamma_i}\exp\Big(-\frac{n_{i-1}\gamma_i^2}{K_T}\Big).$$

Notice that $\delta_L^b\ge\gamma_i$. We thus have, by taking a union bound over $j_i\in[n_i]$, on the events $\mathcal{E}$, $\mathcal{E}_{t,1}^{\Delta}$ and $\mathcal{E}_{t,L}^{\mathbf{H}}$,

$$D_i^b(t)\le K_T\big((1+B)\mathcal{D}_t(W,\tilde{W})+(1+B^2)\delta_L^b\big),$$

with probability at least $1-(n_i/\gamma_i)\exp(-n_{i-1}\gamma_i^2/K_T)$. The claim then follows again from the union bound. ∎

C.1.2. Proof of Proposition 4.15.

Proof of Proposition 4.15. We consider $t\le T$, for a given terminal time $T\in\epsilon\mathbb{N}_{\ge0}$. We again reuse the notation K_t from the proof of Proposition 4.14. Note that $K_t\le K_T$ for all $t\le T$. We also note that at initialization, $\mathcal{D}_0(\mathbf{W},\tilde{W})=0$. We start with a few preliminary facts:

Fact 1: *moment bounds.* We recall a useful fact from the proof of Proposition 4.14: with probability at least $1-KLn_{\max}\exp(-Kn_{\min}^{1/52})$, the event $\mathcal{E}$ occurs, and $\mathcal{E}$ contains the following:

$$|||\mathbf{W}|||_0=|||\tilde{W}|||_0\le K,$$

$$\max_{1\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{t\le T}\operatorname*{ess\,sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(t))|^{50}\Big)^{1/50},\ |||\tilde{W}|||_T\le K_T.$$

We further remark that since $|||\mathbf{W}|||_0 \le K$, from Lemma C.3, we have on the event $\mathcal{E}$:

$$\max_{1\le i\le L}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{t\le T}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(\lfloor t/\epsilon\rfloor))|^{50}\Big)^{1/50},\ |||\mathbf{W}|||_{\lfloor T/\epsilon\rfloor}\le K_T.$$

We also observe that the randomness of the event $\mathcal{E}$ is entirely by the samples of the coupling procedure $\{C_1(j_1),\dots,C_L(j_L): j_i\in[n_i],\ i=1,\dots,L\}$.

Fact 2: *maximal bounds.* We also recall another useful fact from the proof of Proposition 4.14: on the event $\mathcal{E}$, almost surely,

$$\max_{2\le i\le L}\max_{j_{i-1}\in[n_{i-1}],j_i\in[n_i]}\sup_{t\le T}|\tilde{w}_i(t,j_{i-1},j_i)|\le K_T(1+B),$$
$$\max_{2\le i\le L}\max_{j_i\in[n_i]}\sup_{t\le T}|\tilde{b}_i(t,j_i)|\le K_T(1+B).$$

In fact, the same extends to $\mathbf{W}$: on the event $\mathcal{E}$, almost surely,

$$\max_{2\le i\le L}\max_{j_{i-1}\in[n_{i-1}],j_i\in[n_i]}\sup_{t\le T}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|\le K_T(1+B),$$
$$\max_{2\le i\le L}\max_{j_i\in[n_i]}\sup_{t\le T}|\mathbf{b}_i(\lfloor t/\epsilon\rfloor,j_i)|\le K_T(1+B),$$
$$\max_{1\le i\le L}\max_{j_i\in[n_i]}\sup_{t\le T}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(\lfloor t/\epsilon\rfloor))|\le K_T(1+B).$$

Indeed, let us consider the claim for $\mathbf{w}_i$. By Assumption 2.6, for $\mathcal{P}$-almost every z,

$$\sup_{t\ge 0}\max_{j_{L-1}\in[n_{L-1}]}|\Delta_L^{\mathbf{w}}(z,j_{L-1},1;\mathbf{W}(\lfloor t/\epsilon\rfloor))|\le K\big(1+\sup_{t\ge 0}|\Delta_L^{\mathbf{H}}(z,1;\mathbf{W}(\lfloor t/\epsilon\rfloor))|\big)\le K,$$

which implies, by Assumption 2.4, that almost surely, for any $j_{L-1}\in[n_{L-1}]$,

$$\sup_{t\le T}|\mathbf{w}_L(\lfloor t/\epsilon\rfloor,j_{L-1},1)|\le \operatorname*{ess\text{-}sup}|w_L^0(C_{L-1},1)|+KT\le K_T(1+B).$$

Next assuming that $\sup_{t\le T}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|\le K_T(1+B)$ almost surely for a given $i\ge 2$, by Assumption 2.6, we have on the event $\mathcal{E}$, for any $j_{i-1}\in[n_{i-1}]$, $t\le T$ and $\mathcal{P}$-almost every z:

$$\begin{aligned}
&|\Delta_{i-1}^{\mathbf{H}}(z,j_{i-1};\mathbf{W}(\lfloor t/\epsilon\rfloor))|\\
&\quad\le\frac{K}{n_i}\sum_{j_i=1}^{n_i}(1+\sup_{t\le T}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(\lfloor t/\epsilon\rfloor))|)\\
&\qquad\times(1+|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|+|\mathbf{b}_i(\lfloor t/\epsilon\rfloor,j_i)|)\\
&\quad\le\frac{K}{n_i}\sum_{j_i=1}^{n_i}\big(1+\sup_{t\le T}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(\lfloor t/\epsilon\rfloor))|\big)(K_T(1+B)+|\mathbf{b}_i(\lfloor t/\epsilon\rfloor,j_i)|)
\end{aligned}$$

$$\leq K\Big(1+\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\sup_{t\leq T}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(\lfloor t/\epsilon\rfloor))|^2\Big)^{1/2}\Big)$$
$$\times\Big(K_T(1+B)+\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}|\mathbf{b}_i(\lfloor t/\epsilon\rfloor,j_i)|^2\Big)^{1/2}\Big)$$
$$\leq K_T(1+B),$$

where the last step follows from Fact 1. Again by Assumption 2.6, we then obtain

$$|\Delta_{i-1}^{\mathbf{w}}(z,j_{i-1},j_i;\mathbf{W}(\lfloor t/\epsilon\rfloor))|\leq K_T(1+B),$$

which implies, by Assumption 2.4, that almost surely on the event $\mathcal{E}$, for any $j_{i-1}\in[n_{i-1}]$ and $j_i\in[n_i]$,

$$\sup_{t\leq T}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|\leq \operatorname*{ess\text{-}sup}|w_{i-1}^0(C_{i-1},C_i)|+K_T(1+B)T$$
$$\leq K_T(1+B).$$

This proves the claim for $\mathbf{w}_i$, and the rest of the claims are similarly proven.

Now let us consider $2\leq i\leq L$ and particularly the task of bounding

$$\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)-\tilde{w}_i(t,j_{i-1},j_i)|^2\Big)^{1/2},$$

which is a quantity in $\mathcal{D}_T(\tilde{W},\mathbf{W})$. As shown in the proof of Proposition 4.14,

$$\sup_{t\leq T-\zeta}\sup_{0\leq\zeta'\leq\zeta}\max_i(\tilde{A}_i^w(t,\zeta'),\tilde{A}_i^b(t,\zeta'))\leq K_T(1+B)\zeta$$

almost surely, where we recall

$$\tilde{A}_i^w(t,\zeta)=\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\partial_1\tilde{w}_i(t+\zeta,j_{i-1},j_i)-\partial_1\tilde{w}_i(t,j_{i-1},j_i)|^2\Big)^{1/2}.$$

As such, by Assumption 2.4, we have the decomposition

$$\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)-\tilde{w}_i(t,j_{i-1},j_i)|^2\Big)^{1/2}$$
$$=\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\Big|\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\xi_i^{\mathbf{w}}(k\epsilon)\Delta_i^{\mathbf{w}}(z(k),j_{i-1},j_i;\mathbf{W}(k))$$
$$-\int_{s=0}^{t}\partial_1\tilde{w}_i(s,j_{i-1},j_i)\,ds\Big|^2\Big)^{1/2}$$

$$\le \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\Big|\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\xi_i^{\mathbf{w}}(k\epsilon)\Delta_i^{\mathbf{w}}(z(k),j_{i-1},j_i;\mathbf{W}(k)) - \epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\partial_1\tilde{w}_i(k\epsilon,j_{i-1},j_i)\Big|^2\Big)^{1/2} + tK_T(1+B)\epsilon$$

$$\le K\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|Q_1(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|^2 + |Q_2(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|^2\Big)^{1/2} + tK_T(1+B)\epsilon,$$

where we define

$$Q_{1,i}(\lfloor t/\epsilon\rfloor,j_{i-1},j_i) = \epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\mathbb{E}_Z[|\Delta_i^{\mathbf{w}}(Z,j_{i-1},j_i;\mathbf{W}(k)) - \Delta_i^{\mathbf{w}}(Z,j_{i-1},j_i;\tilde{W}(k\epsilon))|],$$

$$Q_{2,i}(\lfloor t/\epsilon\rfloor,j_{i-1},j_i) = \Big|\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\xi_i^{\mathbf{w}}(k\epsilon)\big(\Delta_i^{\mathbf{w}}(z(k),j_{i-1},j_i;\mathbf{W}(k)) - \mathbb{E}_Z[\Delta_i^{\mathbf{w}}(Z,j_{i-1},j_i;\mathbf{W}(k))]\big)\Big|.$$

(Here $\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}=0$ if $\lfloor t/\epsilon\rfloor=0$.) The task is then to bound $Q_{1,i}$ and $Q_{2,i}$.

Bounding $Q_{1,i}$. We take note of a simple identity:

$$\frac{1}{|J|}\sum_{j\in J}\Big(\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}f(j,k)\Big)^2 = \epsilon^2\sum_{k_1=0}^{\lfloor t/\epsilon\rfloor-1}\sum_{k_2=0}^{\lfloor t/\epsilon\rfloor-1}\frac{1}{|J|}\sum_{j\in J}f(j,k_1)f(j,k_2)$$

$$\le \epsilon^2\sum_{k_1=0}^{\lfloor t/\epsilon\rfloor-1}\sum_{k_2=0}^{\lfloor t/\epsilon\rfloor-1}\Big(\frac{1}{|J|}\sum_{j\in J}|f(j,k_1)|^2\Big)^{1/2}\Big(\frac{1}{|J|}\sum_{j\in J}|f(j,k_2)|^2\Big)^{1/2} = \Big(\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\Big(\frac{1}{|J|}\sum_{j\in J}|f(j,k)|^2\Big)^{1/2}\Big)^2.$$

As such, by Assumption 2.6,

$$\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|Q_{1,i}(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|^2\Big)^{1/2} \le \epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}D_i(k) \le K\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\big(D_i^{(1)}(k) + G_i(k) + \mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W}) + F_i(k)\big),$$

in which we define

$$D_i(k) = \Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{w}}(Z, j_{i-1}, j_i; \mathbf{W}(k)) - \Delta_i^{\mathbf{w}}(Z, j_{i-1}, j_i; \tilde{W}(k\epsilon))|^2\Big)^{1/2},$$

$$D_i^{(1)}(k) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1 + |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon))|^2 + |\Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|^2) \times \frac{1}{n_{i-1}}\sum_{j_{i-1}=1}^{n_{i-1}}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\mathbf{H}_{i-1}(X, j_{i-1}; \tilde{W}(k\epsilon)) - \mathbf{H}_{i-1}(X, j_{i-1}; \mathbf{W}(k))|^2\Big)^{1/2},$$

$$G_i(k) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon)) - \Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|^2\Big)^{1/2},$$

$$F_i(k) = \Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\mathbf{H}_i(X, j_i; \tilde{W}(k\epsilon)) - \mathbf{H}_i(X, j_i; \mathbf{W}(k))|^2\Big)^{1/2}.$$

By Lemma C.3 and Fact 1, on the event $\mathcal{E}$,

$$D_i^{(1)}(k) \le K_T F_{i-1}(k),$$

which implies

$$\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|Q_{1,i}(\lfloor t/\epsilon\rfloor, j_{i-1}, j_i)|^2\Big)^{1/2} \le K_T\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}(F_{i-1}(k) + G_i(k) + \mathcal{D}_{k\epsilon}(\mathbf{W}, \tilde{W}) + F_i(k)).$$

We proceed with bounding F_i and G_i.

To bound F_i, by Assumption 2.5 and Cauchy–Schwarz's inequality,

$$\begin{aligned}|F_i(k)|^2 &\le \frac{K}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}\big(1 + |\tilde{w}_i(k\epsilon, j_{i-1}, j_i)|^2 + |\mathbf{w}_i(k, j_{i-1}, j_i)|^2 \\ &\quad + |\tilde{b}_i(k\epsilon, j_i)|^2 + |\mathbf{b}_i(k, j_i)|^2\big)|F_{i-1}(k)|^2 + K\mathcal{D}_{k\epsilon}^2(\mathbf{W}, \tilde{W}) \\ &\le K_T|F_{i-1}(k)|^2 + K\mathcal{D}_{k\epsilon}^2(\mathbf{W}, \tilde{W}),\end{aligned}$$

where the last display holds on the event $\mathcal{E}$ by Fact 1. Notice that, by Assumption 2.5, $|F_1(k)| \le K\mathcal{D}_{k\epsilon}(\mathbf{W}, \tilde{W})$. Therefore, on the event $\mathcal{E}$,

$$\max_{1\le i\le L}|F_i(k)| \le K_T\mathcal{D}_{k\epsilon}(\mathbf{W}, \tilde{W}),$$

which is the desired bound for F_i.

Next let us bound G_i. By Assumption 2.6, we have:

$$G_{i-1}(k) \le K\big(G_i^{(1)}(k) + G_i^{(2)}(k) + G_i^{(3)}(k) + G_i^{(4)}(k)\big),$$

in which

$$\begin{aligned}
G_{i-1}^{(1)}(k) = \Big(\frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \big(1 + |\tilde{w}_i(k\epsilon, j_{i-1}, j_i)| + |\mathbf{w}_i(k, j_{i-1}, j_i)| \\
+ |\tilde{b}_i(k\epsilon, j_i)| + |\mathbf{b}_i(k, j_i)|\big) \\
\times \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon)) - \Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|\Big)^2\Big)^{1/2}, \\
G_{i-1}^{(2)}(k) = \Big(\frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} \big(1 + |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon))| \\
+ |\Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|\big) \\
\times (|\tilde{w}_i(k\epsilon, j_{i-1}, j_i) - \mathbf{w}_i(k, j_{i-1}, j_i)| + |\tilde{b}_i(k\epsilon, j_i) - \mathbf{b}_i(k, j_i)|)\Big)^2\Big)^{1/2}, \\
G_{i-1}^{(3)}(k) = \Big(\frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} \big(1 + |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon))| \\
+ |\Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|\big) \\
\times (1 + |\tilde{w}_i(k\epsilon, j_{i-1}, j_i)| + |\mathbf{w}_i(k, j_{i-1}, j_i)| + |\tilde{b}_i(k\epsilon, j_i)| + |\mathbf{b}_i(k, j_i)|) \\
\times \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\mathbf{H}_i(X, j_i; \tilde{W}(k\epsilon)) - \mathbf{H}_i(X, j_i; \mathbf{W}(k))|\Big)^2\Big)^{1/2}, \\
G_{i-1}^{(4)}(k) = \Big(\frac{1}{n_{i-1}} \sum_{j_{i-1}=1}^{n_{i-1}} \Big(\frac{1}{n_i} \sum_{j_i=1}^{n_i} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} \big(1 + |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon))| \\
+ |\Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|\big) \\
\times (1 + |\tilde{w}_i(k\epsilon, j_{i-1}, j_i)| + |\mathbf{w}_i(k, j_{i-1}, j_i)| + |\tilde{b}_i(k\epsilon, j_i)| + |\mathbf{b}_i(k, j_i)|) \\
\times \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\mathbf{H}_{i-1}(X, j_{i-1}; \tilde{W}(k\epsilon)) - \mathbf{H}_{i-1}(X, j_{i-1}; \mathbf{W}(k))|\Big)^2\Big)^{1/2}.
\end{aligned}$$

To bound $G_{i-1}^{(1)}$, by Cauchy–Schwarz's inequality and Fact 1, on the event $\mathcal{E}$,

$$\begin{aligned}
G_{i-1}^{(1)}(k) &\le \Big(\frac{1}{n_{i-1}n_i} \sum_{j_{i-1}=1}^{n_{i-1}} \sum_{j_i=1}^{n_i} \big(1 + |\tilde{w}_i(k\epsilon, j_{i-1}, j_i)| + |\mathbf{w}_i(k, j_{i-1}, j_i)| + |\tilde{b}_i(k\epsilon, j_i)| \\
&\quad + |\mathbf{b}_i(k, j_i)|\big)^2 \frac{1}{n_i} \sum_{j_i=1}^{n_i} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}} |\Delta_i^{\mathbf{H}}(Z, j_i; \tilde{W}(k\epsilon)) - \Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|^2\Big)^{1/2} \\
&\le K_T G_i(k).
\end{aligned}$$

We also have a bound on $G_{i-1}^{(2)}$ on the event $\mathcal{E}$:

$$\begin{aligned}
G_{i-1}^{(2)}(k) &\le K\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i} \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1+|\Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(k\epsilon))|^2+|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(k))|^2)\Big)^{1/2} \\
&\quad\times\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\tilde{w}_i(k\epsilon,j_{i-1},j_i)-\mathbf{w}_i(k,j_{i-1},j_i)|^2 \\
&\qquad+|\tilde{b}_i(k\epsilon,j_i)-\mathbf{b}_i(k,j_i)|^2\Big)^{1/2} \\
&\le K_T\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W}).
\end{aligned}$$

Similarly, by Fact 1 and Fact 2, on the event $\mathcal{E}$,

$$\begin{aligned}
&G_{i-1}^{(3)}(k) \\
&\quad\le K_T(1+B)\frac{1}{n_i}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1+|\Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(k\epsilon))|+|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(k))|) \\
&\qquad\times\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\mathbf{H}_i(X,j_i;\tilde{W}(k\epsilon))-\mathbf{H}_i(X,j_i;\mathbf{W}(k))| \\
&\quad\le K_T(1+B) \\
&\qquad\times\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1+|\Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(k\epsilon))|^2+|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(k))|^2)\Big)^{1/2}F_i(k) \\
&\quad\le K_T(1+B)F_i(k), \\
&G_{i-1}^{(4)}(k) \\
&\quad\le K_T(1+B)\frac{1}{n_i}\sum_{j_i=1}^{n_i}\operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}(1+|\Delta_i^{\mathbf{H}}(Z,j_i;\tilde{W}(k\epsilon))|+|\Delta_i^{\mathbf{H}}(Z,j_i;\mathbf{W}(k))|)F_{i-1}(k) \\
&\quad\le K_T(1+B)F_{i-1}(k).
\end{aligned}$$

Therefore, on the event $\mathcal{E}$,

$$\begin{aligned}
G_{i-1}(k) &\le K_T(G_i(k)+\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W})+(1+B)(F_i(k)+F_{i-1}(k))) \\
&\le K_T(G_i(k)+(1+B)\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W})).
\end{aligned}$$

Notice that, by Assumption 2.6,

$$G_L(k)\le KF_L(k)\le K_T\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W}).$$

Therefore, on the event $\mathcal{E}$,

$$\max_{1\le i\le L}|G_i(k)|\le K_T(1+B)\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W}),$$

which is the desired bound for G_i.

Together these bounds yield

$$\max_{2\le i\le L}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|Q_{1,i}(\lfloor t/\epsilon\rfloor, j_{i-1}, j_i)|^2\Big)^{1/2} \le \epsilon K_T(1+B)\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W}).$$

Bounding $Q_{2,i}$. For brevity, let us write

$$Z_k = \xi_i^{\mathbf{w}}(k\epsilon)\big(\Delta_i^{\mathbf{w}}(z(k), j_{i-1}, j_i; \mathbf{W}(k)) - \mathbb{E}_Z[\Delta_i^{\mathbf{w}}(Z, j_{i-1}, j_i; \mathbf{W}(k))]\big),$$
$$\underline{Z}_k = \sum_{\ell=0}^{k-1} Z_\ell, \quad \underline{Z}_0 = 0.$$

Let $\mathcal{F}_k$ be the sigma-algebra generated by $\{z(s) : s \in \{0,\dots,k-1\}\}$. Recall that it is independent of the samples $\{C_1(j_1),\dots,C_L(j_L) : j_i \in [n_i],\ i = 1,\dots,L\}$ and hence the event $\mathcal{E}$. Note that $\{\underline{Z}_k\}_{k\in\mathbb{N}}$ is a martingale adapted to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$. Furthermore, for $k \le T/\epsilon$, the martingale difference is bounded:

$$\begin{aligned}|Z_k| &\le K \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{w}}(Z, j_{i-1}, j_i; \mathbf{W}(k))| \\ &\le K\big(1 + \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{\mathbf{H}}(Z, j_i; \mathbf{W}(k))|\big) \le K_T(1+B),\end{aligned}$$

which holds on the event $\mathcal{E}$, by Assumptions 2.4 and 2.6 and Fact 2. Therefore, by Theorem A.1, we have

$$\mathbb{P}\Big(\max_{u\in\{0,1,\dots,T/\epsilon\}} Q_{2,i}(u, j_{i-1}, j_i) \ge (1+B)\zeta; \mathcal{E}\Big) \le 2\exp\Big(-\frac{\zeta^2}{K_T T\epsilon}\Big).$$

Putting it all together. Applying the union bound to the bound on $Q_{2,i}$, we then get that on the event $\mathcal{E}$, with probability at least $1 - 2n_i n_{i-1}\exp(-\zeta^2/(K_T T\epsilon))$, for all $t \le T$,

$$\begin{aligned}&\sup_{s\le t, j_{i-1}\in[n_{i-1}], j_i\in[n_i]}\Big(\frac{1}{n_{i-1}n_i}\sum_{j_{i-1}=1}^{n_{i-1}}\sum_{j_i=1}^{n_i}|\mathbf{w}_i(\lfloor s/\epsilon\rfloor, j_{i-1}, j_i) - \tilde{w}_i(s, j_{i-1}, j_i)|^2\Big)^{1/2} \\ &\le K_T(1+B)\Big(\epsilon\sum_{k=0}^{\lfloor t/\epsilon\rfloor-1}\mathcal{D}_{k\epsilon}(\mathbf{W},\tilde{W}) + \zeta + \epsilon\Big).\end{aligned}$$

One can obtain similar bounds for $\mathbf{b}_i$ and $\mathbf{w}_1$. Together these bounds yield that with probability at least

$$1 - 2\Big(n_1 + \sum_{i=2}^{L} n_i n_{i-1}\Big)\exp\Big(-\frac{\zeta^2}{K_T T\epsilon}\Big) - KLn_{\max}\exp(-Kn_{\min}^{1/52}),$$

we have, for all $t \leq T$,

$$\mathcal{D}_{\lfloor t/\epsilon \rfloor \epsilon}(\tilde{W}, \mathbf{W}) \leq K_T(1+B)\Big(\epsilon \sum_{k=0}^{\lfloor t/\epsilon \rfloor - 1} \mathcal{D}_{k\epsilon}(\mathbf{W}, \tilde{W}) + \zeta + \epsilon\Big),$$

which implies, by Gronwall's lemma,

$$\mathcal{D}_T(\tilde{W}, \mathbf{W}) \leq (\zeta + \epsilon) \exp(K_T(1+B)).$$

Choosing $\zeta = \sqrt{K_T \epsilon \log(2(n_1 + \sum_{i=2}^{L} n_i n_{i-1})/\delta)}$ completes the proof. ∎

C.1.3. Proof of Proposition 4.16.

Proof of Proposition 4.16. We again reuse the notation K_t from the proof of Proposition 4.14. Note that $K_t \leq K_T$ for all $t \leq T$. It is easy to see from Theorem 3.1 that the trajectory $\underline{W}(t)$ exists and is unique. Let us recall the mapping F and the space $\mathcal{W}_T$ from the proof Theorem 3.1; we note that F is associated with the initialization $W(0)$. Since $F(\underline{W})(t) - W(0) = \underline{W}(t) - \underline{W}(0)$ and W is a fixed point of F, we have

$$\begin{aligned} \|W - \underline{W}\|_t &\leq \|W - \underline{W}\|_0 + \|W - F(\underline{W})\|_t \\ &= \|W - \underline{W}\|_0 + \|F(W) - F(\underline{W})\|_t. \end{aligned}$$

Due to truncation, it is immediate that for $2 \leq i \leq L$,

$$\begin{aligned} |\underline{w}_1(0, c_1)| &= |w_1(0, c_1)|, \\ |\underline{w}_i(0, c_{i-1}, c_i)| &\leq |w_i(0, c_{i-1}, c_i)|, \\ |\underline{b}_i(0, c_i)| &\leq |b_i(0, c_i)|. \end{aligned}$$

As such, by repeating the argument of Lemma 3.2, one can show that $\underline{W} \in \mathcal{W}_T$ and that

$$\mathbb{P}\left(\mathsf{max}_T^w(\underline{W}) \geq K_0(T)u\right) \leq 2Le^{1-K_1u^2} \quad \text{for all } u \geq 0.$$

Thus, Lemma 3.4 gives

$$\|F(W) - F(\underline{W})\|_t \leq K_T\Big((1+B)\int_0^t \|W - \underline{W}\|_s \, ds + e^{-KB^2}\Big),$$

which implies, by the previous bound,

$$\|W - \underline{W}\|_t \leq K_T\Big((1+B)\int_0^t \|W - \underline{W}\|_s \, ds + e^{-KB^2}\Big) + \|W - \underline{W}\|_0.$$

Hence, Gronwall's lemma yields

$$\|W - \underline{W}\|_T \leq (\|W - \underline{W}\|_0 + e^{-KB^2}) e^{K_T(1+B)}.$$

Notice that, for $2 \le i \le L$,

$$\begin{aligned}
&\mathbb{E}[|w_i^0(C_{i-1}, C_i) - \underline{w}_i(0, C_{i-1}, C_i)|^2] \\
&\quad = \mathbb{E}[|w_i^0(C_{i-1}, C_i) - B|^2 \mathbb{I}(|w_i^0(C_{i-1}, C_i)| > B)] \\
&\quad \le \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^2 \mathbb{I}(|w_i^0(C_{i-1}, C_i)| > B)] \\
&\quad \le \mathbb{E}[|w_i^0(C_{i-1}, C_i)|^4]^{1/2} \mathbb{P}(|w_i^0(C_{i-1}, C_i)| > B)^{1/2} \le K e^{-KB^2},
\end{aligned}$$

where the last displays comes from Assumption 4.6 and, in particular, we have [37]

$$\mathbb{P}\big(|w_i^0(C_{i-1}, C_i)| \ge r\big) \le K e^{-Kr^2} \quad \text{for all } r \ge 0.$$

Similarly,

$$\mathbb{E}[|b_i^0(C_i) - \underline{b}_i(0, C_i)|^2] \le K e^{-KB^2}.$$

Also recall that $\underline{w}_1(0, c_1) = w_1^0(c_1)$. As such, a similar bound holds for $\|W - \underline{W}\|_0$ and this gives the desired bound on $\|W - \underline{W}\|_T$.

The derivation for $\|\tilde{W} - \underline{\tilde{W}}\|_T$ is similar. Indeed, Lemma C.2 indicates that for any fixed $r \ge 0$, with probability at least $1 - KLn_{\max} \exp(-Ke^{-Kr^2} n_{\min}^{1/52})$, we have

$$|||\tilde{W}|||_0 \le |||W|||_0 + e^{-Kr^2} \le K,$$

as well as that for all $i \in \{2, \dots, L\}$,

$$\begin{aligned}
&\frac{1}{n_{i-1} n_i} \sum_{j_{i-1}=1}^{n_{i-1}} \sum_{j_i=1}^{n_i} \mathbb{I}\big(|w_i^0(C_{i-1}(j_{i-1}), C_i(j_i))| \ge r\big) \\
&\qquad\qquad \le \mathbb{P}\big(|w_i^0(C_{i-1}, C_i)| \ge r\big) + e^{-Kr^2} \le K e^{-Kr^2}, \\
&\frac{1}{n_i} \sum_{j_i=1}^{n_i} \mathbb{I}\big(|b_i^0(C_i(j_i))| \ge r\big) \le \mathbb{P}\big(|b_i^0(C_i)| \ge r\big) + e^{-Kr^2} \le K e^{-Kr^2}.
\end{aligned}$$

By taking $r = B$ and performing an argument similar to the bounding of $\|W - \underline{W}\|_T$, we obtain

$$\|\tilde{W} - \underline{\tilde{W}}\|_T \le (\|\tilde{W} - \underline{\tilde{W}}\|_0 + e^{-KB^2}) e^{K_T(1+B)} \le K e^{-KB^2 + K_T(1+B)},$$

with probability at least $1 - KLn_{\max} \exp(-Ke^{-KB^2} n_{\min}^{1/52})$. The derivation for $\|\mathbf{W} - \underline{\mathbf{W}}\|_T$ is also similar. ∎

C.2. Proofs of Corollaries 4.9, 4.10 and 4.11

Lemma C.4. *Consider the MF trajectory $W(t)$, $t \le T$, under Assumptions 2.4–2.6 and 4.6. For any $\zeta \ge 0$, $\|W - W_\zeta\|_T \le K_{T+\zeta} \zeta$, where $W_\zeta(t) = W(t + \zeta)$ and $K_{T+\zeta}$ is a finite constant that depends on the initialization $W(0)$ and grows continuously with ζ.*

Proof. We reuse the notation K_t from the proof of Proposition 4.14. By Assumption 2.6 and Lemma C.3, for $2 \le i \le L$,

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\le T+\zeta}\mathbb{E}_Z[|\Delta_i^w(t,Z,C_{i-1},C_i)|^2]\Big] &\le K\Big(1+\mathbb{E}\Big[\sup_{t\le T+\zeta}\mathbb{E}_Z[|\Delta_i^H(t,Z,C_i)|^2]\Big]\Big)\\ &\le K_{T+\zeta},\end{aligned}$$

and therefore, by Assumption 2.4,

$$\mathbb{E}\Big[\sup_{t\le T}|w_i(t+\zeta,C_{i-1},C_i)-w_i(t,C_{i-1},C_i)|^2\Big]^{1/2} \le K_{T+\zeta}\,\zeta.$$

One can also deduce a similar bound for b_i and w_1. ∎

Proof of Corollary 4.9. We reuse the notation K_t from the proof of Proposition 4.14. We have the following decomposition for $i \in [L]$:

$$\begin{aligned}&\Big|\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[\psi(\mathbf{H}_i(\lfloor t/\epsilon\rfloor,X,j_i))]-\mathbb{E}_Z\mathbb{E}_{C_i}[\psi(H_i(t,X,C_i))]\Big|\\ &\quad\le \frac{1}{n_i}\sum_{j_i=1}^{n_i}\Big|\mathbb{E}_Z[\psi(\mathbf{H}_i(\lfloor t/\epsilon\rfloor,X,j_i))]-\mathbb{E}_Z[\psi(H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i(j_i)))]\Big|\\ &\qquad+\Big|\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[\psi(H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i(j_i)))]-\mathbb{E}_Z\mathbb{E}_{C_i}[\psi(H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i))]\Big|\\ &\qquad+\mathbb{E}_Z\mathbb{E}_{C_i}[|\psi(H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i))-\psi(H_i(t,X,C_i))|]\\ &\quad\le K\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|\mathbf{H}_i(\lfloor t/\epsilon\rfloor,X,j_i)-H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i(j_i))|^2]\Big)^{1/2}\\ &\qquad+\Big|\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[\psi(H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i(j_i)))]-\mathbb{E}_Z\mathbb{E}_{C_i}[\psi(H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i))]\Big|\\ &\qquad+K\mathbb{E}_Z\mathbb{E}_{C_i}[|H_i(\lfloor t/\epsilon\rfloor\epsilon,X,C_i)-H_i(t,X,C_i)|]\\ &\quad= Q_{1,i}(t)+Q_{2,i}(t)+Q_{3,i}(t),\end{aligned}$$

where we use the fact ψ is K-Lipschitz and Cauchy–Schwarz's inequality. We provide bounds on each term. Note that by the fact ψ is K-Lipschitz and Assumption 2.5:

$$\begin{aligned}&|\mathbb{E}_Z[\psi(Y,\hat{\mathbf{y}}(\lfloor t/\epsilon\rfloor,X))]-\mathbb{E}_Z[\psi(Y,\hat{y}(t,X))]|\\ &\qquad\le K\mathbb{E}_Z[|\mathbf{H}_L(\lfloor t/\epsilon\rfloor,X,1)-H_L(t,X,1)|],\end{aligned}$$

and as such, bounding $Q_{1,L}$ gives the last claim in the corollary.

Bounding $Q_{1,i}$. By Assumption 2.5 and Cauchy–Schwarz's inequality, for $i \geq 2$,

$$
\begin{aligned}
&|Q_{1,i}(t)|^2 \\
&\quad \leq \frac{K}{n_{i-1}n_i} \sum_{j_{i-1}=1}^{n_{i-1}} \sum_{j_i=1}^{n_i} \big(1 + |\mathbf{w}_i(\lfloor t/\epsilon \rfloor, j_{i-1}, j_i)|^2 + |w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))|^2 \\
&\qquad + |\mathbf{b}_i(\lfloor t/\epsilon \rfloor, j_i)|^2 + |b_i(t, C_i(j_i))|^2\big)|Q_{1,i-1}(t)|^2 + K\mathcal{D}_t^2(W, \mathbf{W}) \\
&\quad \leq \frac{K}{n_{i-1}n_i} \sum_{j_{i-1}=1}^{n_{i-1}} \sum_{j_i=1}^{n_i} \big(1 + |w_i(t, C_{i-1}(j_{i-1}), C_i(j_i))|^2 + |b_i(t, C_i(j_i))|^2 \\
&\qquad + \mathcal{D}_t^2(W, \mathbf{W})\big)|Q_{1,i-1}(t)|^2 + K\mathcal{D}_t^2(W, \mathbf{W}) \\
&\quad \overset{\text{(a)}}{\leq} K_T(1 + \mathcal{D}_t^2(W, \mathbf{W}))|Q_{1,i-1}(t)|^2 + K\mathcal{D}_t^2(W, \mathbf{W}) \\
&\quad \leq K_T(1 + \mathcal{D}_t^2(W, \mathbf{W}))|Q_{1,i-1}(t)|^2,
\end{aligned}
$$

where, by Lemmas C.2 and C.3, (a) holds for all $t \leq T$ and all $i \geq 2$ with probability at least $1 - KLn_{\max}\exp(-Kn_{\min}^{1/52})$. Also, from Assumption 2.5, $Q_{1,1}(t) \leq K\mathcal{D}_t(W, \mathbf{W})$. We thus have

$$
\max_{i \in [L]} \sup_{t \leq T} Q_{1,i}(t) \leq K_T(1 + \mathcal{D}_T^L(W, \mathbf{W}))\mathcal{D}_T(W, \mathbf{W}),
$$

with probability at least $1 - KLn_{\max}\exp(-Kn_{\min}^{1/52})$.

Bounding $Q_{2,i}$. Recall that $\{C_i(j_i)\}_{j_i \in [n_i]}$ are η_i-independent and $\eta_i \leq n_i^{-1/2}$. Since ψ is K-bounded, we have, by Theorem A.2 and the union bound, that

$$
\sup_{t \leq T} Q_{2,i}(t) \leq \sqrt{\frac{K}{n_i} \log\Big(\frac{KT}{\epsilon\delta}\Big)},
$$

with probability at least $1 - \delta$.

Bounding $Q_{3,i}$. By Assumption 2.5 and Lemma C.3, for all $t \leq T$ and $i \geq 2$,

$$
\begin{aligned}
|Q_{3,i}(t)|^2 &\leq K\mathbb{E}\big[1 + |w_i(t, C_{i-1}, C_i)|^2 + |b_i(t, C_i)|^2 \\
&\qquad + |w_i(\lfloor t/\epsilon \rfloor \epsilon, C_{i-1}, C_i)|^2 + |b_i(\lfloor t/\epsilon \rfloor \epsilon, C_i)|^2\big]|Q_{3,i-1}(t)|^2 \\
&\qquad + K\mathbb{E}\big[|w_i(t, C_{i-1}, C_i) - w_i(\lfloor t/\epsilon \rfloor \epsilon, C_{i-1}, C_i)|^2 \\
&\qquad + |b_i(t, C_i) - b_i(\lfloor t/\epsilon \rfloor \epsilon, C_i)|^2\big] \\
&\leq K_T|Q_{3,i-1}(t)|^2 + K\mathbb{E}\big[|w_i(t, C_{i-1}, C_i) - w_i(\lfloor t/\epsilon \rfloor \epsilon, C_{i-1}, C_i)|^2 \\
&\qquad + |b_i(t, C_i) - b_i(\lfloor t/\epsilon \rfloor \epsilon, C_i)|^2\big].
\end{aligned}
$$

Similarly,

$$
|Q_{3,1}(t)|^2 \leq K\mathbb{E}[|w_1(t, C_1) - w_1(\lfloor t/\epsilon \rfloor \epsilon, C_1)|^2].
$$

We thus obtain, from Lemma C.4,

$$\max_{i\in[L]} \sup_{t\le T} Q_{3,i}(t) \le K_T \epsilon.$$

Putting it all together. All previous bounds show that

$$\sup_{t\le T} \Big| \frac{1}{n_i} \sum_{j_i=1}^{n_i} \mathbb{E}_Z[\psi(\mathbf{H}_i(\lfloor t/\epsilon \rfloor, X, j_i))] - \mathbb{E}_Z \mathbb{E}_{C_i}[\psi(H_i(t, X, C_i))] \Big|$$
$$\le K_T(1 + \mathcal{D}_T^L(W, \mathbf{W}))\mathcal{D}_T(W, \mathbf{W}) + \sqrt{\frac{K}{n_i} \log\Big(\frac{KT}{\epsilon\delta}\Big)} + K_T\epsilon,$$

with probability at least $1 - \delta - KLn_{\max} \exp(-Kn_{\min}^{1/52})$. Together with Theorem 4.7, we obtain the claim. ∎

Proof of Corollary 4.10. In the following, for a set $J = \{N_1, \dots, N_L\}$ with $N_L = 1$, we write $J \to \infty$ to mean that $N_1, \dots, N_{L-1} \to \infty$ such that for $N_{\max} = \max J$ and $N_{\min} = \min\{N_1, \dots, N_{L-1}\}$, we have $N_{\min} \to \infty$ and $N_{\min}^{-c} \log N_{\max} \to 0$ for any $c > 0$.

For a given $T \ge 0$ and a set of integers $I = \{n_1, \dots, n_L\}$, for any two sets $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ of the form

$$\mathcal{W}^{(1)} = \big\{w_1^{(1)}(t, r_1), w_i^{(1)}(t, r_{i-1}, r_i), b_i^{(1)}(t, r_i) : r_i \in [n_i], i \in [L], t \in [0, T]\big\},$$

and similar for $\mathcal{W}^{(2)}$, let us equip a distance metric:

$$d_{I,T}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)}) = \max\Big(\max_{1\le i\le L} d_{I,T}^{w,i}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)}), \max_{2\le i\le L} d_{I,T}^{b,i}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)})\Big),$$

$$d_{I,T}^{w,1}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)}) = \Big(\frac{1}{n_1} \sum_{r_1=1}^{n_1} \sup_{t\le T} |w_1^{(1)}(t, r_1) - w_1^{(2)}(t, r_1)|^2\Big)^{1/2},$$

$$d_{I,T}^{w,i}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)}) = \Big(\frac{1}{n_{i-1}n_i} \sum_{r_{i-1}=1}^{n_{i-1}} \sum_{r_i=1}^{n_i} \sup_{t\le T} |w_i^{(1)}(t, r_{i-1}, r_i) - w_i^{(2)}(t, r_{i-1}, r_i)|^2\Big)^{1/2},$$

$$d_{I,T}^{b,i}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)}) = \Big(\frac{1}{n_i} \sum_{r_i=1}^{n_i} \sup_{t\le T} |b_i^{(1)}(t, r_i) - b_i^{(2)}(t, r_i)|^2\Big)^{1/2}, \quad 2 \le i \le L.$$

Let us also consider the space $\mathcal{F}_{I,T}$ of 1-bounded Lipschitz functions f with respect to this distance metric:

$$|f(\mathcal{W}^{(1)}) - f(\mathcal{W}^{(2)})| \le 2 \wedge d_{I,T}(\mathcal{W}^{(1)}, \mathcal{W}^{(2)}).$$

Step 1: *Coupling via finite-width networks.* Recall that $(\Omega, P, \{w_i^0\}_{i\in[L]}, \{b_i^0\}_{2\le i\le L})$ satisfies Assumption 4.4, i.e., $\bar{\eta}$-independence. Thus, for each index $J = \{N_1, \dots, N_L\}$

of Init, one can find a sampling rule for which the samples $\{C_i(j_i)\}_{j_i\in[N_i]}$ are η_i-independent for $i \le L-1$ and $\eta_i \to 0$ as $N_i \to \infty$. Then one obtains a neural network initialization $\mathbf{W}(0)$ with law ρ by setting

$$\mathbf{w}_1(0,j_1) = w_1(0,C_1(j_1)), \quad \mathbf{w}_i(0,j_{i-1},j_i) = w_i(0,C_{i-1}(j_{i-1}),C_i(j_i)),$$
$$\mathbf{b}_i(0,j_i) = b_i(0,C_i(j_i)), \quad 2 \le i \le L.$$

Similarly using $(\hat{\Omega},\hat{P},\{\hat{w}_i^0\}_{i\in[L]},\{\hat{b}_i^0\}_{2\le i\le L})$, we obtain $\hat{\mathbf{W}}(0)$ with the same law ρ by setting

$$\hat{\mathbf{w}}_1(0,j_1) = \hat{w}_1(0,\hat{C}_1(j_1)), \quad \hat{\mathbf{w}}_i(0,j_{i-1},j_i) = \hat{w}_i(0,\hat{C}_{i-1}(j_{i-1}),\hat{C}_i(j_i)),$$
$$\hat{\mathbf{b}}_i(0,j_i) = \hat{b}_i(0,\hat{C}_i(j_i)), \quad 2 \le i \le L,$$

where $\{\hat{C}_i(j_i)\}_{j_i\in[N_i]}$ are η_i-independent for $i \le L-1$. We consider the evolution $\mathbf{W}(t)$ starting from $\mathbf{W}(0)$ (which is independent of W once $\mathbf{W}(0)$ is fixed). Note that $\mathbf{W}(t)$ is a deterministic function of its initialization $\mathbf{W}(0)$ and the data $\{z(s)\}_{s\le t}$. Similarly, we consider the counterpart for $\hat{W}$: the evolution $\hat{\mathbf{W}}(t)$ as a function of the initialization $\hat{\mathbf{W}}(0)$ and the data $\{\hat{z}(s)\}_{s\le t}$. Due to sharing the same distribution for both the initialization and the data, these evolutions have the same law. In other words, for any $\theta > 0$,

$$\inf_{\text{coupling of }(\mathbf{W},\hat{\mathbf{W}})} \mathbb{P}(\|\mathbf{W}-\hat{\mathbf{W}}\|_T \ge \theta) = 0,$$

in which

$$\begin{aligned}
&\|\mathbf{W}-\hat{\mathbf{W}}\|_T \\
&\quad = \max\Bigg(\Big(\frac{1}{N_1}\sum_{j_1=1}^{N_1}\sup_{t\le T}|\mathbf{w}_1(\lfloor t/\epsilon\rfloor,j_1)-\hat{\mathbf{w}}_1(\lfloor t/\epsilon\rfloor,j_1)|^2\Big)^{1/2}, \\
&\qquad \max_{2\le i\le L}\Big(\frac{1}{N_{i-1}N_i}\sum_{j_{i-1}=1}^{N_{i-1}}\sum_{j_i=1}^{N_i}\sup_{t\le T}|\mathbf{w}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)-\hat{\mathbf{w}}_i(\lfloor t/\epsilon\rfloor,j_{i-1},j_i)|^2\Big)^{1/2}, \\
&\qquad \max_{2\le i\le L}\Big(\frac{1}{N_i}\sum_{j_i=1}^{N_i}\sup_{t\le T}|\mathbf{b}_i(\lfloor t/\epsilon\rfloor,j_i)-\hat{\mathbf{b}}_i(\lfloor t/\epsilon\rfloor,j_i)|^2\Big)^{1/2}\Bigg).
\end{aligned}$$

Theorem 4.7 implies that, following the coupling procedure, for any $\delta > 0$, with probability at least $1-\delta-o_{J;L}$,

$$\mathcal{D}_T(W,\mathbf{W}) \le o_{\epsilon,J;\delta,T,L},$$

where here and in the following, we denote by $o_{J;L}$ and $o_{\epsilon,J;\delta,T,L}$ appropriate quantities that may change from line to line with $o_{J;L}\to 0$ and $o_{\epsilon,J;\delta,T,L}\to 0$ as the learning rate $\epsilon\to 0$ and $J\to\infty$. Here without loss of generality, we assume $o_{\epsilon,J;\delta,T,L} > 0$.

We also have a similar result for $\mathcal{D}_T(\hat{W}, \hat{\mathbf{W}})$. As such,

$$\begin{aligned}
&\inf_{\text{coupling of } (\mathcal{W}, \hat{\mathcal{W}})} \mathbb{P}(d_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \geq o_{\epsilon,J;\delta,T,L}) \\
&\quad \leq \mathbb{P}(\mathcal{D}_T(W, \mathbf{W}) \geq o_{\epsilon,J;\delta,T,L}) + \mathbb{P}(\mathcal{D}_T(\hat{W}, \hat{\mathbf{W}}) \geq o_{\epsilon,J;\delta,T,L}) \\
&\qquad + \inf_{\text{coupling of } (\mathbf{W}, \hat{\mathbf{W}})} \mathbb{P}(\|\mathbf{W} - \hat{\mathbf{W}}\|_T \geq o_{\epsilon,J;\delta,T,L}) \\
&\quad \leq 2\delta + 2o_{J;L},
\end{aligned}$$

where we define

$$\begin{aligned}
\mathcal{W}_{J,T} &= \big\{w_1(t, C_1(j_1)), w_i(t, C_{i-1}(j_{i-1}), C_i(j_i)), b_i(t, C_i(j_i)) : \\
&\qquad j_i \in [N_i], i \in [L], t \in [0, T]\big\}, \\
\hat{\mathcal{W}}_{J,T} &= \big\{\hat{w}_1(t, \hat{C}_1(j_1)), \hat{w}_i(t, \hat{C}_{i-1}(j_{i-1}), \hat{C}_i(j_i)), \hat{b}_i(t, \hat{C}_i(j_i)) : \\
&\qquad j_i \in [N_i], i \in [L], t \in [0, T]\big\}.
\end{aligned}$$

This gives a sense of approximate closeness between W and $\hat{W}$ on a set $J = \{N_1, \ldots, N_L\}$ with sufficiently large size N_i, importantly under the assumption of $\bar{\eta}$-independence. To extend this to arbitrary finite sizes, we perform the following argument.

Step 2: *Extension to finite sizes.* For a given fixed set $I = \{n_1, \ldots, n_L\}$ with $n_L = 1$, let us consider the following sub-sampling procedure: for each $i \in [L]$ and each $r_i \in [n_i]$, we independently sample $V_i(r_i)$ uniformly from $[N_i]$, and then set $S_i(r_i) = C_i(V_i(r_i))$ and $\hat{S}_i(r_i) = \hat{C}_i(V_i(r_i))$. Let us define

$$\begin{aligned}
\mathcal{W}_{I,T} = \big\{&w_1(t, S_1(r_1)), w_i(t, S_{i-1}(r_{i-1}), S_i(r_i)), b_i(t, S_i(r_i)) : \\
&r_i \in [n_i], i \in [L], t \in [0, T]\big\},
\end{aligned}$$

and $\hat{\mathcal{W}}_{I,T}$ similarly. We prove that $\text{Law}(\mathcal{W}_{I,T})$ and $\text{Law}(\hat{\mathcal{W}}_{I,T})$ are close in an appropriate sense. This shall be done via a connection with $d_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T})$ on the set J.

Let $\mathbb{E}_V$ denote the expectation with respect to the sub-sampling procedure only (i.e., with respect to the randomness of $\{V_i(r_i) : r_i \in [n_i],\ i \in [L]\}$). Notice that

$$\begin{aligned}
\mathbb{E}_V[|d^{w,i}_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T})|^2] &= |d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T})|^2, \\
\mathbb{E}_V[|d^{b,i}_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T})|^2] &= |d^{b,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T})|^2.
\end{aligned}$$

Using this fact and Markov's inequality, for any $\theta > 0$,

$$\begin{aligned}
&\mathbb{P}(d^{w,i}_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T}) \geq \theta) \\
&\quad \leq \mathbb{P}(d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \geq \theta^2) \\
&\qquad + \mathbb{E}\big[\mathbb{E}_V[\mathbb{I}(d^{w,i}_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T}) \geq \theta)]\, \mathbb{I}(d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \leq \theta^2)\big]
\end{aligned}$$

$$\begin{aligned}
&\le \mathbb{P}(d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \ge \theta^2) \\
&\quad + \theta^{-2}\mathbb{E}\big[\mathbb{E}_V[|d^{w,i}_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T})|^2]\,\mathbb{I}(d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \le \theta^2)\big] \\
&\le \mathbb{P}(d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \ge \theta^2) + \theta^2.
\end{aligned}$$

The bound on $d_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T})$ gives

$$\inf_{\text{coupling of } (\mathcal{W},\hat{\mathcal{W}})} \mathbb{P}\big(d^{w,i}_{J,T}(\mathcal{W}_{J,T}, \hat{\mathcal{W}}_{J,T}) \ge \tilde{O}_{\epsilon,J;\delta,T,L}(1)\big) \le 2\delta + 2o_{J;L}.$$

Then, by taking $\theta = \sqrt{o_{\epsilon,J;\delta,T,L}}$, we have

$$\inf_{\text{coupling of } (\mathcal{W},\hat{\mathcal{W}})} \mathbb{P}\big(d^{w,i}_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T}) \ge \sqrt{o_{\epsilon,J;\delta,T,L}}\big) \le e_{T,L}(\delta,\epsilon,J),$$

$$e_{T,L}(\delta,\epsilon,J) = 2\delta + 2o_{J;L} + o_{\epsilon,J;\delta,T,L}.$$

A similar fact holds for $d^{b,i}_{I,T}$, and therefore by the union bound,

$$\inf_{\text{coupling of } (\mathcal{W},\hat{\mathcal{W}})} \mathbb{P}\big(d_{I,T}(\mathcal{W}_{I,T}, \hat{\mathcal{W}}_{I,T}) \ge \sqrt{o_{\epsilon,J;\delta,T,L}(\epsilon,J)}\big) \le 2Le_{T,L}(\delta,\epsilon,J).$$

In particular, this implies, for any $f \in \mathcal{F}_{I,T}$,

$$|\mathbb{E}[f(\mathcal{W}_{I,T})] - \mathbb{E}[f(\hat{\mathcal{W}}_{I,T})]| \le 4Le_{T,L}(\delta,\epsilon,J) + \sqrt{o_{\epsilon,J;\delta,T,L}}.$$

This describes closeness between $\mathrm{Law}(\mathcal{W}_{I,T})$ and $\mathrm{Law}(\hat{\mathcal{W}}_{I,T})$. Note that this is not sufficient to conclude the proof (via taking $\epsilon \to 0$, $J \to \infty$, $\delta \to 0$): the left-hand side involves the random variables $C_i(V_i(r_i))$, which firstly does not remove $\bar{\eta}$-independence and secondly is not independent of J since $V_i(r_i) \in [N_i]$.

Step 3: *Removing $\bar{\eta}$-independence.* Let $\{U_i(j_i)\}_{j_i\in[N_i]}$ be drawn i.i.d. from P_i, independently for each $i \in [L]$, as in the statement of the corollary. First we recall that $\{C_i(j_i)\}_{j_i\in[N_i]}$ are η_i-independent for $i \le L-1$ with $\eta_i \to 0$ as $N_i \to \infty$. We also note $N_L = n_L = 1$ and hence $U_L(1) = C_L(1)$. As such, for any 1-bounded function g,

$$\big|\mathbb{E}_{\sim V}[g(S_i(r_i) : r_i \in [n_i], i \in [L])] - \mathbb{E}_{\sim V}[g(U_i(V_i(r_i)) : r_i \in [n_i], i \in [L])]\big| \le o_{J;L,I},$$

where $o_{J;L,I}$ is a deterministic quantity such that $o_{J;L,I} \to 0$ as $J \to \infty$ and $\mathbb{E}_{\sim V}$ denotes the expectation with respect to everything excluding the sub-sampling procedure. Indeed, suppose that $\{\tilde{V}_1(1),\dots,\tilde{V}_1(n_1)\}$ is a permutation of $\{V_1(1),\dots,V_1(n_1)\}$ such that $\tilde{V}_1(1) > \cdots > \tilde{V}_1(n_1)$. Using the $\bar{\eta}$-independence property, we have the following for $|\zeta_1| \le \eta_1$:

$$\begin{aligned}
&\mathbb{E}_{\sim V}[g(S_i(r_i) : r_i \in [n_i], i \in [L])] \\
&\quad = \mathbb{E}_{\sim V}\mathbb{E}_{\sim C_1(\tilde{V}_1(1))}\mathbb{E}_{C_1(\tilde{V}_1(1))|\sim C_1(\tilde{V}_1(1))}\big[g\big(C_1(\tilde{V}_1(r_1)), C_i(V_i(r_i)) : \\
&\qquad r_1 \in [n_1], r_i \in [n_i] \text{ for } i \ge 2\big)\big]
\end{aligned}$$

$$= \mathbb{E}_{\sim V}\mathbb{E}_{\sim C_1(\tilde{V}_1(1))}\mathbb{E}_{U_1(\tilde{V}_1(1))}\big[g\big(U_1(\tilde{V}_1(1)), C_1(\tilde{V}_1(r_1)), C_i(V_i(r_i)) : r_1 \in [n_1] \setminus \{1\}, r_i \in [n_i] \text{ for } i \geq 2\big)\big] + \zeta_1,$$

where conditioning on the sub-sampling, $\mathbb{E}_{\sim C_1(\tilde{V}_1(1))}$ is the expectation with respect to everything excluding $C_1(\tilde{V}_1(1))$, and $\mathbb{E}_{C_1(\tilde{V}_1(1))|\sim C_1(\tilde{V}_1(1))}$ is the expectation with respect to $C_1(\tilde{V}_1(1))$ conditioning on everything else. (Here we have assumed that $V_1(1), \ldots, V_1(n_1)$ are all distinct, since any repeated elements can be removed without affecting the argument.) By iterating this decomposition, we obtain the claim.

On the other hand, since $\{U_i(j_i)\}_{j_i \in [N_i]}$ are i.i.d., it is easy to see that

$$\mathbb{E}[g(U_i(V_i(r_i)) : r_i \in [n_i], i \in [L])] = \mathbb{E}[g(U_i(r_i) : r_i \in [n_i], i \in [L])].$$

Together with the result from the previous step, we thus have, for any $f \in \mathcal{F}_{I,T}$,

$$|\mathbb{E}[f(\mathcal{W}(I,T))] - \mathbb{E}[f(\hat{\mathcal{W}}(I,T))]| \leq 2o_{J;L,I} + 8Le_{T,L}(\delta,\epsilon,J) + 2\sqrt{o_{\epsilon,J;\delta,T,L}},$$

where we recall

$$\mathcal{W}(I,T) = \big\{w_1(t,U_1(j_1)), w_i(t,U_{i-1}(j_{i-1}),U_i(j_i)), b_i(t,U_i(j_i)) : j_i \in [n_i], i \in [L], t \in [0,T]\big\},$$

and similarly for $\hat{\mathcal{W}}(I,T)$. Note that the left-hand side is completely independent of J, ϵ and δ. So, by taking $\epsilon \to 0$, $J \to \infty$, $\delta \to 0$, we have

$$\mathbb{E}[f(\mathcal{W}(I,T))] = \mathbb{E}[f(\hat{\mathcal{W}}(I,T))],$$

which completes the proof. ∎

Proof of Corollary 4.11. For any test function $\psi\colon \mathbb{W}_1 \times \mathbb{W}_2 \to \mathbb{R}$ that is bounded with bounded gradient, we have

$$\begin{aligned}
&\frac{d}{dt}\int \psi(u_1,u_2)\,d\rho_t(u_1,u_2)\\
&= \frac{d}{dt}\,\mathbb{E}_{C_1}[\psi(w_1(t,C_1),w_2(t,C_1,1))]\\
&= -\mathbb{E}_{C_1}\big[\big\langle \nabla_1\psi(w_1(t,C_1),w_2(t,C_1,1)), \xi_1^{\mathbf{w}}(t)\,\mathbb{E}_Z[\Delta_1^w(t,Z,C_1)]\big\rangle\big]\\
&\quad - \mathbb{E}_{C_1}\big[\big\langle \nabla_2\psi(w_1(t,C_1),w_2(t,C_1,1)), \xi_2^{\mathbf{w}}(t)\,\mathbb{E}_Z[\Delta_2^w(t,Z,C_1,1)]\big\rangle\big]\\
&\overset{\text{(a)}}{=} -\mathbb{E}_{C_1}\big[\big\langle \nabla_1\psi(w_1(t,C_1),w_2(t,C_1,1)), \xi_1^{\mathbf{w}}(t)\,\mathbb{E}_Z[\underline{\Delta}_1^w(u_1,u_2;Z,\rho_t)]\big\rangle\big]\\
&\quad - \mathbb{E}_{C_1}\big[\big\langle \nabla_2\psi(w_1(t,C_1),w_2(t,C_1,1)), \xi_2^{\mathbf{w}}(t)\,\mathbb{E}_Z[\underline{\Delta}_2^w(u_1,u_2;Z,\rho_t)]\big\rangle\big]\\
&= -\int \langle \nabla\psi(u_1,u_2), G(u_1,u_2;\rho_t)\rangle\,d\rho_t(u_1,u_2),
\end{aligned}$$

where step (a) can be checked easily by inspection. This shows that ρ_t satisfies the claimed distributional partial differential equation. The rest of the claims follow in a similar vein to the proof of Corollary 4.9. ∎

D. Remaining details for Section 5

D.1. Infinite-M limit of the canonical MF limit under i.i.d. initializations

We give the full description of the infinite-M limit W^* of the canonical MF limit, described in Section 5.1.2. To that end, let us first consider depth $L \geq 5$. Let $\{w_i^*\}_{i=1}^L$ and $\{b_i^*\}_{i=2}^L$ be functions satisfying the following dynamics:

$$\begin{aligned}
\frac{\partial}{\partial t} w_1^*(t,u_1) &= -\xi_1^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_1^{w*}(t,Z,u_1)],\\
\frac{\partial}{\partial t} w_2^*(t,u_1,u_2,v_2) &= -\xi_2^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_2^{w*}(t,Z,u_1,u_2,v_2)],\\
\frac{\partial}{\partial t} w_i^*(t,u_i,v_{i-1},v_i) &= -\xi_i^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_i^{w*}(t,Z,u_i,v_{i-1},v_i)],\\
&\qquad i=3,\dots,L-2,\\
\frac{\partial}{\partial t} w_{L-1}^*(t,u_{L-1},u_L,v_{L-2},v_{L-1}) &= -\xi_{L-1}^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_{L-1}^{w*}(t,Z,u_{L-1},u_L,v_{L-2},v_{L-1})],\\
\frac{\partial}{\partial t} w_L^*(t,u_L,v_{L-1}) &= -\xi_L^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_L^{w*}(t,Z,u_L,v_{L-1})],\\
\frac{\partial}{\partial t} b_i^*(t,v_i) &= -\xi_i^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_i^{b*}(t,Z)],\\
&\qquad i=2,\dots,L-2,\\
\frac{\partial}{\partial t} b_{L-1}^*(t,u_L,v_{L-1}) &= -\xi_{L-1}^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_{L-1}^{b*}(t,Z,u_L)],\\
\frac{\partial}{\partial t} b_L^*(t) &= -\xi_L^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_L^{b*}(t,Z)],\\
&\qquad \text{for all } u_i \in \operatorname{supp}(\rho_{\mathbf{w}}^i) \text{ for } i=1,\dots,L,\\
&\qquad \text{for all } v_i \in \operatorname{supp}(\rho_{\mathbf{b}}^i) \text{ for } i=2,\dots,L-1,
\end{aligned}$$

with the initialization $w_1^*(0,u_1)=u_1$, $w_2^*(0,\cdot,u_2,\cdot)=u_2$, $w_i^*(0,u_i,\cdot,\cdot)=u_i$ for $i=3,\dots,L-2$, $w_{L-1}^*(0,u_{L-1},\cdot,\cdot,\cdot)=u_{L-1}$, $w_L^*(0,u_L,\cdot)=u_L$, $b_i^*(0,v_i)=v_i$ for $i=2,\dots,L-2$, $b_{L-1}^*(0,\cdot,v_{L-1})=v_{L-1}$ and $b_L^*(0)$ a deterministic constant that $\rho_{\mathbf{b}}^L(b_L^*(0))=1$ (i.e., $b_L^*(0)=\mathfrak{p}_L(1)$ according to equation (5.5)). Here the quantities are defined by the following forward and backward recursions.

Forward recursion:

$$\begin{aligned}
&H_1^*(t,x,u_1)=\phi_1(w_1^*(t,u_1),x),\\
&H_2^*(t,x,v_2)=\int \phi_2(w_2^*(t,u_1,u_2,v_2),b_2^*(t,v_2),H_1^*(t,x,u_1))\rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2),\\
&H_i^*(t,x,v_i)\\
&\quad=\int \phi_i(w_i^*(t,u_i,v_{i-1},v_i),b_i^*(t,v_i),H_{i-1}^*(t,x,v_{i-1}))\rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1}),\\
&\qquad i=3,\dots,L-2,
\end{aligned}$$

$$
\begin{aligned}
&H_{L-1}^{*}(t,x,u_L,v_{L-1})\\
&\quad=\int \phi_{L-1}\big(w_{L-1}^{*}(t,u_{L-1},u_L,v_{L-2},v_{L-1}),b_{L-1}^{*}(t,u_L,v_{L-1}),H_{L-2}^{*}(t,x,v_{L-2})\big)\\
&\qquad\times \rho_{\mathbf{w}}^{L-1}(du_{L-1})\rho_{\mathbf{b}}^{L-2}(dv_{L-2}),\\
&H_{L}^{*}(t,x)\\
&\quad=\int \phi_L(w_L^{*}(t,u_L,v_{L-1}),b_L^{*}(t),H_{L-1}^{*}(t,x,u_L,v_{L-1}))\rho_{\mathbf{w}}^{L}(du_L)\rho_{\mathbf{b}}^{L-1}(dv_{L-1}),\\
&\hat{y}^{*}(t,x)=\phi_{L+1}(H_L^{*}(t,x)).
\end{aligned}
$$

Backward recursion:

$$
\begin{aligned}
&\Delta_L^{H*}(t,z)=\sigma_L^{\mathbf{H}}(y,\hat{y}^{*}(t,x),H_L^{*}(t,x)),\\
&\Delta_L^{w*}(t,z,u_L,v_{L-1})\\
&=\sigma_L^{\mathbf{w}}\big(\Delta_L^{H*}(t,z),w_L^{*}(t,u_L,v_{L-1}),b_L^{*}(t),H_L^{*}(t,x),H_{L-1}^{*}(t,x,u_L,v_{L-1})\big),\\
&\Delta_L^{b*}(t,z)\\
&=\int \sigma_L^{\mathbf{b}}(\Delta_L^{H*}(t,z),w_L^{*}(t,u_L,v_{L-1}),b_L^{*}(t),H_L^{*}(t,x),H_{L-1}^{*}(t,x,u_L,v_{L-1}))\\
&\qquad\times \rho_{\mathbf{w}}^{L}(du_L)\rho_{\mathbf{b}}^{L-1}(dv_{L-1}),\\
&\Delta_{L-1}^{H*}(t,z,u_L,v_{L-1})\\
&=\sigma_{L-1}^{\mathbf{H}}(\Delta_L^{H*}(t,z),w_L^{*}(t,u_L,v_{L-1}),b_L^{*}(t),H_L^{*}(t,x),H_{L-1}^{*}(t,x,u_L,v_{L-1})),\\
&\Delta_{L-1}^{w*}(t,z,u_{L-1},u_L,v_{L-2},v_{L-1})\\
&=\sigma_{L-1}^{\mathbf{w}}\big(\Delta_{L-1}^{H*}(t,z,u_L,v_{L-1}),w_{L-1}^{*}(t,u_{L-1},u_L,v_{L-2},v_{L-1}),b_{L-1}^{*}(t,u_L,v_{L-1}),\\
&\qquad H_{L-1}^{*}(t,x,u_L,v_{L-1}),H_{L-2}^{*}(t,x,v_{L-2})\big),\\
&\Delta_{L-1}^{b*}(t,z,u_L,v_{L-1})\\
&=\int \sigma_{L-1}^{\mathbf{b}}\big(\Delta_{L-1}^{H*}(t,z,u_L,v_{L-1}),w_{L-1}^{*}(t,u_{L-1},u_L,v_{L-2},v_{L-1}),b_{L-1}^{*}(t,u_L,v_{L-1}),\\
&\qquad H_{L-1}^{*}(t,x,u_L,v_{L-1}),H_{L-2}^{*}(t,x,v_{L-2})\big)\rho_{\mathbf{w}}^{L-1}(du_{L-1})\rho_{\mathbf{b}}^{L-2}(dv_{L-2}),\\
&\Delta_{L-2}^{H*}(t,z,v_{L-2})\\
&=\int \sigma_{L-2}^{\mathbf{H}}\big(\Delta_{L-1}^{H*}(t,z,u_L,v_{L-1}),w_{L-1}^{*}(t,u_{L-1},u_L,v_{L-2},v_{L-1}),b_{L-1}^{*}(t,u_L,v_{L-1}),\\
&\qquad H_{L-1}^{*}(t,x,u_L,v_{L-1}),H_{L-2}^{*}(t,x,v_{L-2})\big)\rho_{\mathbf{w}}^{L}(du_L)\rho_{\mathbf{w}}^{L-1}(du_{L-1})\rho_{\mathbf{b}}^{L-1}(dv_{L-1}),\\
&\Delta_i^{w*}(t,z,u_i,v_{i-1},v_i)\\
&=\sigma_i^{\mathbf{w}}(\Delta_i^{H*}(t,z,v_i),w_i^{*}(t,u_i,v_{i-1},v_i),b_i^{*}(t,v_i),H_i^{*}(t,x,v_i),H_{i-1}^{*}(t,x,v_{i-1})),
\end{aligned}
$$

$$
\begin{aligned}
&\Delta_i^{b*}(t,z,v_i)\\
&= \int \sigma_i^{\mathbf{b}}(\Delta_i^{H*}(t,z,v_i), w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_i^*(t,x,v_i), H_{i-1}^*(t,x,v_{i-1}))\\
&\qquad \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1}),\\
&\Delta_{i-1}^{H*}(t,z,v_{i-1})\\
&= \int \sigma_{i-1}^{\mathbf{H}}(\Delta_i^{H*}(t,z,v_i), w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_i^*(t,x,v_i), H_{i-1}^*(t,x,v_{i-1}))\\
&\qquad \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^i(dv_i) \quad \text{for } i = L-2,\dots,3,\\
&\Delta_2^{w*}(t,z,u_1,u_2,v_2)\\
&= \sigma_2^{\mathbf{w}}(\Delta_2^{H*}(t,z,v_2), w_2^*(t,u_1,u_2,v_2), b_2^*(t,v_2), H_2^*(t,x,v_2), H_1^*(t,x,u_1)),\\
&\Delta_2^{b*}(t,z,v_2)\\
&= \int \sigma_2^{\mathbf{b}}(\Delta_2^{H*}(t,z,v_2), w_2^*(t,u_1,u_2,v_2), b_2^*(t,v_2), H_2^*(t,x,v_2), H_1^*(t,x,u_1))\\
&\qquad \times \rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^1(du_1),\\
&\Delta_1^{H*}(t,z,u_1)\\
&= \int \sigma_1^{\mathbf{H}}(\Delta_2^{H*}(t,z,v_2), w_2^*(t,u_1,u_2,v_2), b_2^*(t,v_2), H_2^*(t,x,v_2), H_1^*(t,x,u_1))\\
&\qquad \times \rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{b}}^2(dv_2),\\
&\Delta_1^{w*}(t,z,u_1) = \sigma_1^{\mathbf{w}}(\Delta_1^{H*}(t,z,u_1), w_1^*(t,u_1), x).
\end{aligned}
$$

In the case $L = 3$ and $L = 4$, we define the dynamics of w_i^* and b_i^* similarly. In particular, for $L = 4$, one can simply disregard all above equations that are with invalid indices. For $L = 3$, we define

$$
\begin{aligned}
\frac{\partial}{\partial t} w_1^*(t,u_1) &= -\xi_1^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_1^{w*}(t,Z,u_1)],\\
\frac{\partial}{\partial t} w_2^*(t,u_1,u_2,u_3,v_2) &= -\xi_2^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_2^{w*}(t,Z,u_1,u_2,u_3,v_2)],\\
\frac{\partial}{\partial t} w_3^*(t,u_3,v_2) &= -\xi_3^{\mathbf{w}}(t)\mathbb{E}_Z[\Delta_3^{w*}(t,Z,u_3,v_2)],\\
\frac{\partial}{\partial t} b_2^*(t,u_3,v_2) &= -\xi_2^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_2^{b*}(t,Z,u_3,v_2)],\\
\frac{\partial}{\partial t} b_3^*(t) &= -\xi_3^{\mathbf{b}}(t)\mathbb{E}_Z[\Delta_3^{b*}(t,Z)],
\end{aligned}
$$

for all $u_i \in \operatorname{supp}(\rho_{\mathbf{w}}^i)$ for $i = 1,2,3$, for all $v_i \in \operatorname{supp}(\rho_{\mathbf{b}}^i)$ for $i = 2,3$, in which

$$
\begin{aligned}
H_1^*(t,x,u_1) &= \phi_1(w_1^*(t,u_1),x),\\
H_2^*(t,x,u_3,v_2) &= \int \phi_2(w_2^*(t,u_1,u_2,u_3,v_2), b_2^*(t,u_3,v_2), H_1^*(t,x,u_1))\\
&\qquad \times \rho_{\mathbf{w}}^1(du_1)\,\rho_{\mathbf{w}}^2(du_2),
\end{aligned}
$$

$$
\begin{aligned}
H_3^*(t,x) &= \int \phi_3(w_3^*(t,u_3,v_2), b_3^*(t), H_2^*(t,x,u_3,v_2))\,\rho_{\mathbf{w}}^3(du_3)\,\rho_{\mathbf{b}}^2(dv_2),\\
\hat{y}^*(t,x) &= \phi_4(H_3^*(t,x)),\\
\Delta_3^{H*}(t,z) &= \sigma_3^{\mathbf{H}}(y, \hat{y}^*(t,x), H_3^*(t,x)),\\
\Delta_3^{w*}(t,z,u_3,v_2) &= \sigma_3^{\mathbf{w}}(\Delta_3^{H*}(t,z), w_3^*(t,u_3,v_2), b_3^*(t), H_3^*(t,x), H_2^*(t,x,u_3,v_2)),\\
\Delta_3^{b*}(t,z) &= \int \sigma_3^{\mathbf{b}}(\Delta_3^{H*}(t,z), w_3^*(t,u_3,v_2), b_3^*(t), H_3^*(t,x), H_2^*(t,x,u_3,v_2))\\
&\qquad\qquad \times \rho_{\mathbf{w}}^3(du_3)\,\rho_{\mathbf{b}}^2(dv_2),\\
\Delta_2^{H*}(t,z,u_3,v_2) &= \sigma_3^{\mathbf{H}}(\Delta_3^{H*}(t,z), w_3^*(t,u_3,v_2), b_3^*(t), H_3^*(t,x), H_2^*(t,x,u_3,v_2)),\\
\Delta_2^{w*}(t,z,u_1,u_2,u_3,v_2) &= \sigma_2^{\mathbf{w}}\big(\Delta_2^{H*}(t,z,u_3,v_2), w_2^*(t,u_1,u_2,u_3,v_2),\\
&\qquad\qquad b_2^*(t,u_3,v_2), H_2^*(t,x,u_3,v_2), H_1^*(t,x,u_1)\big),\\
\Delta_2^{b*}(t,z,u_3,v_2) &= \int \sigma_2^{\mathbf{b}}\big(\Delta_2^{H*}(t,z,u_3,v_2), w_2^*(t,u_1,u_2,u_3,v_2), b_2^*(t,u_3,v_2),\\
&\qquad\qquad H_2^*(t,x,u_3,v_2), H_1^*(t,x,u_1)\big)\,\rho_{\mathbf{w}}^1(du_1)\,\rho_{\mathbf{w}}^2(du_2),\\
\Delta_1^{H*}(t,z,u_1) &= \int \sigma_1^{\mathbf{H}}\big(\Delta_2^{H*}(t,z,u_3,v_2), w_2^*(t,u_1,u_2,u_3,v_2), b_2^*(t,u_3,v_2),\\
&\qquad\qquad H_2^*(t,x,u_3,v_2), H_1^*(t,x,u_1)\big)\,\rho_{\mathbf{w}}^2(du_2)\,\rho_{\mathbf{w}}^3(du_3), \rho_{\mathbf{b}}^2(dv_2),\\
\Delta_1^{w*}(t,z,u_1) &= \sigma_1^{\mathbf{w}}(\Delta_1^{H*}(t,z,u_1), w_1^*(t,u_1), x).
\end{aligned}
$$

Finally, let $W^*(t) = \{w_1^*(t,\cdot), w_i^*(t,\cdot), b_i^*(t,\cdot),\ i = 2,\ldots,L\}$. The existence and uniqueness of such dynamics follow similarly to the proof of Theorem 3.1.

Theorem D.1 (Complete statement of Theorem 5.3). *Given* $(\rho_{\mathbf{w}}^1,\ldots,\rho_{\mathbf{w}}^L,\rho_{\mathbf{b}}^2,\ldots,\rho_{\mathbf{b}}^L)$ *and an integer* M*, construct the canonical neuronal ensemble* (Ω^M, P^M)*, the random variables* $(C_1,\ldots,C_L) \sim P^M = \prod_{i=1}^L P_i^M$ *and the canonical MF limit* W^M *as described in Section 5.1.1. Also construct the dynamics* W^* *described in Section 5.1.2. For* $L \geq 5$*, define the following*:

$$
\begin{aligned}
w_1^\infty(t,c_1) &= w_1^*(t, w_1^0(c_1)),\\
w_2^\infty(t,c_1,c_2) &= w_2^*(t, w_1^0(c_1), w_2^0(c_1,c_2), b_2^0(c_2)),\\
w_i^\infty(t,c_{i-1},c_i) &= w_i^*(t, w_i^0(c_{i-1},c_i), b_{i-1}^0(c_{i-1}), b_i^0(c_i)), \quad i = 3,\ldots,L-2,\\
w_{L-1}^\infty(t,c_{L-2},c_{L-1}) &= w_{L-1}^*(t, w_{L-1}^0(c_{L-2},c_{L-1}), w_L^0(c_{L-1},1),\\
&\qquad b_{L-2}^0(c_{L-2}), b_{L-1}^0(c_{L-1})),\\
w_L^\infty(t,c_{L-1},1) &= w_L^*(t, w_L^0(c_{L-1},1), b_{L-1}^0(c_{L-1})),\\
b_i^\infty(t,c_i) &= b_i^*(t, b_i^0(c_i)), \qquad i = 2,\ldots,L-2,\\
b_{L-1}^\infty(t,c_{L-1}) &= b_{L-1}^*(t, w_L^0(c_{L-1},1), b_{L-1}^0(c_{L-1})),\\
b_L^\infty(t,1) &= b_L^*(t),\\
c_i \in \Omega_i &= \Lambda \times \mathbb{N}_{>0}, \qquad i = 1,\ldots,L-1.
\end{aligned}
$$

For $L = 4$, we define similarly by disregarding the equations with invalid indices. For $L = 3$, we define

$$\begin{aligned} w_1^\infty(t, c_1) &= w_1^*(t, w_1^0(c_1)), \\ w_2^\infty(t, c_1, c_2) &= w_2^*(t, w_2^0(c_1, c_2), w_3^0(c_2, 1), w_1^0(c_1), b_2^0(c_2)), \\ w_3^\infty(t, c_2, 1) &= w_3^*(t, w_3^0(c_2, 1), b_2^0(c_2)), \\ b_2^\infty(t, c_2) &= b_2^*(t, w_3^0(c_2, 1), b_2^0(c_2)), \\ b_3^\infty(t) &= b_3^*(t). \end{aligned}$$

We also let $W^\infty(t) = \{w_1^\infty(t, \cdot), w_i^\infty(t, \cdot, \cdot), b_i^\infty(t, \cdot),\ i = 2, \dots, L\}$. Let us consider

$$\begin{aligned} \langle W^M - W^\infty \rangle_t &= \max\Big(\max_{1 \le i \le L} \langle w_i^M - w_i^\infty \rangle_t, \max_{2 \le i \le L} \langle b_i^M - b_i^\infty \rangle_t \Big), \\ \langle w_i^M - w_i^\infty \rangle_t &= \mathbb{E}[|w_i^M(t, C_{i-1}, C_i) - w_i^\infty(t, C_{i-1}, C_i)|^2]^{1/2}, \\ \langle b_i^M - b_i^\infty \rangle_t &= \mathbb{E}[|b_i^M(t, C_i) - b_i^\infty(t, C_i)|^2]^{1/2}, \quad i = 2, \dots, L, \\ \langle w_1^M - w_1^\infty \rangle_t &= \mathbb{E}[|w_1^M(t, C_1) - w_1^\infty(t, C_1)|^2]^{1/2}. \end{aligned}$$

Then, under Assumptions 2.4–2.6 and 4.6, for any $T \ge 0$ and $L \ge 2$,

$$\sup_{t \le T} \langle W^M - W^\infty \rangle_t \le \frac{K_{T,L}}{M^{0.499}},$$

for sufficiently large $M = M(T, L)$, where $K_{T,L}$ is a constant that depends on T and L. Furthermore, for $L \ge 4$ and $2 \le i \le L - 2$,

$$\sup_{t \le T} \mathbb{E}[|H_i(X, C_i; W^M(t)) - H_i^*(t, X, b_i^0(C_i))|^2]^{1/2} \le \frac{K_{T,L}}{M^{0.499}}.$$

D.2. Proof of Theorem D.1

Proof of Theorem D.1. Let us consider the case $L \ge 5$; the case where $L \le 4$ is similarly proven. We use $K_{T,L}$ to denote a generic constant that depends on T and L and may change from line to line.

Step 1. By following the argument of Lemma 3.2, one can show that $\|W^\infty\|_T$ as well as the following quantities are all bounded by $K_{T,L}$:

$$\begin{aligned} &\mathbb{E}\Big[\sup_{t \le T} \operatorname*{ess\text{-}sup}_{Z \sim \mathcal{P}} |H_1^*(t, X, w_1^0(C_1))|^{50} \Big], \\ &\max_{2 \le i \le L-2} \mathbb{E}\Big[\sup_{t \le T} \operatorname*{ess\text{-}sup}_{Z \sim \mathcal{P}} |H_i^*(t, X, b_i^0(C_i))|^{50} \Big], \\ &\mathbb{E}\Big[\sup_{t \le T} \operatorname*{ess\text{-}sup}_{Z \sim \mathcal{P}} |H_{L-1}^*(t, X, w_L^0(C_{L-1}, 1), b_{L-1}^0(C_{L-1}))|^{50} \Big], \end{aligned}$$

$$\sup_{t\le T}\ \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|H_L^*(t,X)|,$$
$$\mathbb{E}\Big[\sup_{t\le T}\ \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_1^{H*}(t,Z,w_1^0(C_1))|^{50}\Big],$$
$$\max_{2\le i\le L-2}\ \mathbb{E}\Big[\sup_{t\le T}\ \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_i^{H*}(t,Z,b_i^0(C_i))|^{50}\Big],$$
$$\mathbb{E}\Big[\sup_{t\le T}\ \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_{L-1}^{H*}(t,Z,w_L^0(C_{L-1},1),b_{L-1}^0(C_{L-1}))|^{50}\Big],$$
$$\sup_{t\le T}\ \operatorname*{ess\text{-}sup}_{Z\sim\mathcal{P}}|\Delta_L^{H*}(t,Z)|.$$

Likewise one can also show that for any $B \ge 0$,

$$\mathbb{P}\big(\mathsf{max}_T^w(W^\infty) \ge K_{T,L}B\big) \le 2K_{T,L}e^{-KB^2},$$

in which

$$\mathsf{max}_T^w(W^\infty) = \max_{2\le i\le L}\ \sup_{t\le T}|w_i^\infty(t,C_{i-1},C_i)|.$$

Step 2. Let us define

$$\begin{aligned}
D_1(t) &= \mathbb{E}[|H_1(X,C_1;W^\infty(t)) - H_1^*(t,X,w_1^0(C_1))|^2],\\
D_i(t) &= \mathbb{E}[|H_i(X,C_i;W^\infty(t)) - H_i^*(t,X,b_i^0(C_i))|^2], \quad i=2,\dots,L-2,\\
D_{L-1}(t) &= \mathbb{E}\Big[\Big|H_{L-1}(X,C_{L-1};W^\infty(t))\\
&\qquad - H_{L-1}^*(t,X,w_L^0(C_{L-1},1),b_{L-1}^0(C_{L-1}))\Big|^2\Big],\\
D_L(t) &= \mathbb{E}[|H_L(X,1;W^\infty(t)) - H_L^*(t,X)|^2].
\end{aligned}$$

We claim that for $t \le T$,

$$D_i(t) \le \frac{K_{T,L}}{M}, \quad i \in [L].$$

Firstly, it is immediate that

$$H_1(X,C_1;W^\infty(t)) = \phi_1(w_1^*(t,w_1^0(C_1)),X) = H_1^*(t,X,w_1^0(C_1)),$$

and hence $D_1(t) = 0$. For $i = 2$, we have

$$\begin{aligned}
D_2(t) = \mathbb{E}\Big[\Big|&\mathbb{E}_{C_1}\big[\phi_2\big(w_2^*(t,w_1^0(C_1),w_2^0(C_1,C_2),b_2^0(C_2)),b_2^*(t,b_2^0(C_2)),\\
&H_1(X,C_1;W^\infty(t))\big)\big]\\
&-\int \phi_2(w_2^*(t,u_1,u_2,b_2^0(C_2)),b_2^*(t,b_2^0(C_2)),H_1^*(t,x,u_1))\\
&\times \rho_{\mathbf{w}}^1(du_1)\,\rho_{\mathbf{w}}^2(du_2)\Big|^2\Big].
\end{aligned}$$

Recalling that $w_2^0(C_1, C_2) = \mathfrak{q}_2(\theta_1, \theta_2)(\lambda_2)$, $b_2^0(C_2) = \mathfrak{p}_2(\theta_2)(\lambda_2)$ and $w_1^0(C_1) = \mathfrak{p}_1(\theta_1)(\lambda_1)$ from the construction of Section 5.1.1, we have

$$
\begin{aligned}
&\mathbb{E}_{C_2}\Big[\big|\mathbb{E}_{C_1}[\phi_2(w_2^*(t, w_1^0(C_1), w_2^0(C_1, C_2), b_2^0(C_2)), b_2^*(t, b_2^0(C_2)),\\
&\qquad H_1(x, C_1; W^\infty(t)))]\big|^2\Big]\\
&\overset{\text{(a)}}{=} \mathbb{E}_{\theta_1,\lambda_1,\theta_1',\lambda_1',\theta_2,\lambda_2}\Big[\big\langle\phi_2\big(w_2^*(t, \mathfrak{p}_1(\theta_1)(\lambda_1), \mathfrak{q}_2(\theta_1, \theta_2)(\lambda_2), \mathfrak{p}_2(\theta_2)(\lambda_2)),\\
&\qquad b_2^*(t, \mathfrak{p}_2(\theta_2)(\lambda_2)), H_1^*(t, x, \mathfrak{p}_1(\theta_1)(\lambda_1))\big),\\
&\qquad \phi_2\big(w_2^*(t, \mathfrak{p}_1(\theta_1')(\lambda_1'), \mathfrak{q}_2(\theta_1', \theta_2)(\lambda_2), \mathfrak{p}_2(\theta_2)(\lambda_2)),\\
&\qquad b_2^*(t, \mathfrak{p}_2(\theta_2)(\lambda_2)), H_1^*(t, x, \mathfrak{p}_1(\theta_1')(\lambda_1'))\big)\big\rangle\Big]\\
&\overset{\text{(b)}}{=} \mathbb{E}_{\theta_1,\theta_1'}\Big[\mathbb{I}(\theta_1 = \theta_1')\int\big\langle\phi_2(w_2^*(t, u_1, u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1)),\\
&\qquad \phi_2(w_2^*(t, u_1', u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1'))\big\rangle\\
&\quad \times \rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^1(du_1')\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{b}}^2(dv_2)\Big]\\
&\quad + \mathbb{E}_{\theta_1,\theta_1'}\Big[\mathbb{I}(\theta_1 \neq \theta_1')\int\big\langle\phi_2(w_2^*(t, u_1, u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1)),\\
&\qquad \phi_2(w_2^*(t, u_1', u_2', v_2), b_2^*(t, v_2), H_1^*(t, x, u_1'))\big\rangle\\
&\quad \times \rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{w}}^1(du_1')\rho_{\mathbf{w}}^2(du_2')\rho_{\mathbf{b}}^2(dv_2)\Big]\\
&= \frac{1}{M}\int\Big\langle\phi_2(w_2^*(t, u_1, u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1)),\\
&\qquad \phi_2(w_2^*(t, u_1', u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1'))\Big\rangle\\
&\quad \times \rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^1(du_1')\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{b}}^2(dv_2)\\
&\quad + \frac{M-1}{M}\int\Big|\int\phi_2(w_2^*(t, u_1, u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1))\rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2)\Big|^2\\
&\quad \times \rho_{\mathbf{b}}^2(dv_2),
\end{aligned}
$$

where in step (a), $(\theta_1', \lambda_1') \sim \mathrm{Unif}([M]) \times P_0$ is an independent copy of (θ_1, λ_1) and is independent of (θ_2, λ_2), and step (b) is by the construction of $\mathfrak{p}_1$, $\mathfrak{p}_2$ and $\mathfrak{q}_2$. It is also easy to see that

$$
\begin{aligned}
&\mathbb{E}_{C_2}\Big[\Big\langle\mathbb{E}_{C_1}\big[\phi_2\big(w_2^*(t, w_1^0(C_1), w_2^0(C_1, C_2), b_2^0(C_2)), b_2^*(t, b_2^0(C_2)),\\
&\qquad H_1(x, C_1; W^\infty(t))\big)\big],\\
&\qquad \int\phi_2(w_2^*(t, u_1, u_2, b_2^0(C_2)), b_2^*(t, b_2^0(C_2)), H_1^*(t, x, u_1))\rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2)\Big\rangle\Big]\\
&= \int\Big|\int\phi_2(w_2^*(t, u_1, u_2, v_2), b_2^*(t, v_2), H_1^*(t, x, u_1))\rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2)\Big|^2\rho_{\mathbf{b}}^2(dv_2).
\end{aligned}
$$

Therefore, for $t \le T$,

$$\begin{aligned}
D_2(t) &\le \frac{1}{M}\int \mathbb{E}\big[\big|\big\langle \phi_2(w_2^*(t,u_1,u_2,v_2), b_2^*(t,v_2), H_1^*(t,X,u_1)),\\
&\qquad \phi_2(w_2^*(t,u_1',u_2,v_2), b_2^*(t,v_2), H_1^*(t,X,u_1'))\big\rangle\big|\big]\\
&\quad \times \rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^1(du_1')\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{b}}^2(dv_2)\\
&\quad + \frac{1}{M}\int \mathbb{E}\Big[\Big|\int \phi_2(w_2^*(t,u_1,u_2,v_2), b_2^*(t,v_2), H_1^*(t,X,u_1))\\
&\quad \times \rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2)\Big|^2\Big]\rho_{\mathbf{b}}^2(dv_2)\\
&\le \frac{K}{M}\int \mathbb{E}[|\phi_2(w_2^*(t,u_1,u_2,v_2), b_2^*(t,v_2), H_1^*(t,X,u_1))|^2]\\
&\quad \times \rho_{\mathbf{w}}^1(du_1)\rho_{\mathbf{w}}^2(du_2)\rho_{\mathbf{b}}^2(dv_2)\\
&\le \frac{K_{T,L}}{M},
\end{aligned}$$

where we use Step 1 and Assumption 2.5 in the last step. For $i \in \{3,\dots,L-2\}$, recall that $w_i^0(C_{i-1},C_i) = \mathfrak{q}_i(\theta_{i-1},\theta_i)(\lambda_i)$, $b_i^0(C_i) = \mathfrak{p}_i(\theta_i)(\lambda_i)$ and $b_{i-1}^0(C_{i-1}) = \mathfrak{p}_{i-1}(\theta_{i-1})(\lambda_{i-1})$ from the construction of Section 5.1.1. Then, similar to the argument for $i = 2$,

$$\begin{aligned}
&\mathbb{E}_{C_i}\big[\big|\mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i)),\\
&\quad b_i^*(t,b_i^0(C_i)), H_{i-1}^*(t,x,b_{i-1}^0(C_{i-1}))\big)\big]\big|^2\big]\\
&= \frac{1}{M}\int \big\langle \phi_i(w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_{i-1}^*(t,x,v_{i-1})),\\
&\quad \phi_i(w_i^*(t,u_i,v_{i-1}',v_i), b_i^*(t,v_i), H_{i-1}^*(t,x,v_{i-1}'))\big\rangle\\
&\quad \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\rho_{\mathbf{b}}^{i-1}(dv_{i-1}')\rho_{\mathbf{b}}^i(dv_i)\\
&\quad + \frac{M-1}{M}\int\Big|\int \phi_2(w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_{i-1}^*(t,x,v_{i-1}))\\
&\quad \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})|^2\rho_{\mathbf{b}}^i(dv_i),\\
&\mathbb{E}_{C_i}\Big[\Big\langle \mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i)),\\
&\quad b_i^*(t,b_i^0(C_i)), H_{i-1}^*(t,x,b_{i-1}^0(C_{i-1}))\big)\big],\\
&\quad \int \phi_i(w_i^*(t,u_i,v_{i-1},b_i^0(C_i)), b_i^*(t,b_i^0(C_i)), H_{i-1}^*(t,x,v_{i-1}))\\
&\quad \times \rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big\rangle\Big]\\
&= \int\Big|\int \phi_i(w_i^*(t,u_i,v_{i-1},v_i), b_i^*(t,v_i), H_{i-1}^*(t,x,v_{i-1}))\rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big|^2\\
&\quad \times \rho_{\mathbf{b}}^i(dv_i),
\end{aligned}$$

which then gives, by Step 1 and Assumption 2.5,

$$
\begin{aligned}
&\mathbb{E}_{C_i}\Big[\Big|\mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i)),\\
&\qquad b_i^*(t,b_i^0(C_i)),H_{i-1}^*(t,x,b_{i-1}^0(C_{i-1}))\big)\big]-H_i^*(t,X,b_i^0(C_i))\Big|^2\Big]\\
&=\mathbb{E}_{C_i}\Big[\Big|\mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i)),\\
&\qquad b_i^*(t,b_i^0(C_i)),H_{i-1}^*(t,x,b_{i-1}^0(C_{i-1}))\big)\big]\\
&\qquad-\int\phi_i(w_i^*(t,u_i,v_{i-1},b_i^0(C_i)),b_i^*(t,b_i^0(C_i)),H_{i-1}^*(t,x,v_{i-1}))\\
&\qquad\times\rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\Big|^2\Big]\\
&\le\frac{K}{M}\int|\phi_i(w_i^*(t,u_i,v_{i-1},v_i),b_i^*(t,v_i),H_{i-1}^*(t,x,v_{i-1}))|^2\\
&\qquad\times\rho_{\mathbf{w}}^i(du_i)\rho_{\mathbf{b}}^{i-1}(dv_{i-1})\rho_{\mathbf{b}}^i(dv_i)\\
&\le\frac{K_{T,L}}{M}.
\end{aligned}
$$

Next, notice that again by Step 1 and Assumption 2.5,

$$
\begin{aligned}
&\mathbb{E}\big[\big|H_i(X,C_i;W^\infty(t))-\mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i)),\\
&\qquad b_i^*(t,b_i^0(C_i)),H_{i-1}^*(t,X,b_{i-1}^0(C_{i-1}))\big)\big]\big|^2\big]\\
&=\mathbb{E}\big[\big|\mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i)),\\
&\qquad b_i^*(t,b_i^0(C_i)),H_{i-1}(X,C_{i-1};W^\infty(t))\big)\big]\\
&\qquad-\mathbb{E}_{C_{i-1}}\big[\phi_i\big(w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i)),\\
&\qquad b_i^*(t,b_i^0(C_i)),H_{i-1}^*(t,X,b_{i-1}^0(C_{i-1}))\big)\big]\big|^2\big]\\
&\le K\mathbb{E}\big[\big|\mathbb{E}_{C_{i-1}}\big[(1+|w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i))|+|b_i^*(t,b_i^0(C_i))|)\\
&\qquad\times|H_{i-1}(X,C_{i-1};W^\infty(t))-H_{i-1}^*(t,X,b_{i-1}^0(C_{i-1}))|\big]\big|^2\big]\\
&\le K\mathbb{E}[(1+|w_i^*(t,w_i^0(C_{i-1},C_i),b_{i-1}^0(C_{i-1}),b_i^0(C_i))|+|b_i^*(t,b_i^0(C_i))|)^2]\\
&\qquad\times\mathbb{E}[|H_{i-1}(X,C_{i-1};W^\infty(t))-H_{i-1}^*(t,X,b_{i-1}^0(C_{i-1}))|^2]\\
&\le K_{T,L}D_{i-1}(t).
\end{aligned}
$$

Hence,

$$
D_i(t)\le\frac{K_{T,L}}{M}+K_{T,L}\,D_{i-1}(t).
$$

This proves the claim for $i\le L-2$. The other claims are similar.

Step 3. Let us define

$$
D_1^H(t)=\mathbb{E}[|\Delta_1^H(Z,C_1;W^\infty(t))-\Delta_1^{H*}(t,Z,w_1^0(C_1))|^2],
$$

$$D_i^H(t) = \mathbb{E}[|\Delta_i^H(Z, C_i; W^\infty(t)) - \Delta_i^{H*}(t, Z, b_i^0(C_i))|^2], \quad i = 2, \dots, L-2,$$
$$D_{L-1}^H(t) = \mathbb{E}[|\Delta_{L-1}^H(Z, C_{L-1}; W^\infty(t)) - \Delta_{L-1}^{H*}(t, Z, w_L^0(C_{L-1}, 1), b_{L-1}^0(C_{L-1}))|^2],$$
$$D_L^H(t) = \mathbb{E}[|\Delta_L^H(Z, 1; W^\infty(t)) - \Delta_L^{H*}(t, Z)|^2].$$

We claim that for $t \le T$,

$$D_i^H(t) \le K_{T,L} \frac{\log^{1/2} M}{M}, \quad i \in [L].$$

The derivation is similar to Step 2; let us give a sketch and highlight the difference. The last claim for $i = L$ is immediate from Assumption 2.6 and Step 2. Let us consider the claim for $2 \le i \le L-3$; the rest of the claims are similar. We have

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\mathbb{E}_{C_{i+1}}\big[\sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})), w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i),\\
&\quad b_{i+1}^0(C_{i+1})), b_{i+1}^*(t, b_{i+1}^0(C_{i+1})), H_{i+1}^*(t, X, b_{i+1}^0(C_{i+1})), H_i^*(t, X, b_i^0(C_i))\big)\big]\\
&\quad - \Delta_i^{H*}(t, Z, b_i^0(C_i))\Big|^2\Big]\\
&= \mathbb{E}\Big[\Big|\mathbb{E}_{C_{i+1}}\big[\sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})), w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i),\\
&\quad b_{i+1}^0(C_{i+1})), b_{i+1}^*(t, b_{i+1}^0(C_{i+1})), H_{i+1}^*(t, X, b_{i+1}^0(C_{i+1})), H_i^*(t, X, b_i^0(C_i))\big)\big]\\
&\quad - \int \sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H*}(t, Z, v_{i+1}), w_{i+1}^*(t, u_{i+1}, b_i^0(C_i), v_{i+1}), b_{i+1}^*(t, v_{i+1}),\\
&\quad H_{i+1}^*(t, X, v_{i+1}), H_i^*(t, X, b_i^0(C_i))\big)\rho_{\mathbf{w}}^{i+1}(du_{i+1})\rho_{\mathbf{b}}^{i+1}(dv_{i+1})\Big|^2\Big]\\
&\overset{(a)}{\le} \frac{K}{M}\int \mathbb{E}\big[\big|\sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H*}(t, Z, v_{i+1}), w_{i+1}^*(t, u_{i+1}, v_i, v_{i+1}), b_{i+1}^*(t, v_{i+1}),\\
&\quad H_{i+1}^*(t, X, v_{i+1}), H_i^*(t, X, v_i)\big)\big|^2\big]\rho_{\mathbf{w}}^{i+1}(du_{i+1})\rho_{\mathbf{b}}^i(dv_i)\rho_{\mathbf{b}}^{i+1}(dv_{i+1})\\
&\overset{(b)}{\le} \frac{K}{M}\int \mathbb{E}[(1 + |\Delta_{i+1}^{H*}(t, Z, v_{i+1})|^2)(1 + |w_{i+1}^*(t, u_{i+1}, v_i, v_{i+1})|^2\\
&\quad + |b_{i+1}^*(t, v_{i+1})|^2)]\rho_{\mathbf{w}}^{i+1}(du_{i+1})\rho_{\mathbf{b}}^i(dv_i)\rho_{\mathbf{b}}^{i+1}(dv_{i+1})\\
&\overset{(c)}{\le} \frac{K_{T,L}}{M},
\end{aligned}$$

where (a) is similar to Step 2, in which we use the fact that $w_{i+1}^0(C_i, C_{i+1}) = \mathfrak{q}_{i+1}(\theta_i, \theta_{i+1})(\lambda_{i+1})$, $b_i^0(C_i) = \mathfrak{p}_i(\theta_i)(\lambda_i)$ and $b_{i+1}^0(C_{i+1}) = \mathfrak{p}_{i+1}(\theta_{i+1})(\lambda_{i+1})$, from the construction of Section 5.1.1, (b) is by Assumption 2.6, and (c) follows from Step 1. We also note

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\mathbb{E}_{C_{i+1}}\big[\sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})), w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i),\\
&\quad b_{i+1}^0(C_{i+1})), b_{i+1}^*(t, b_{i+1}^0(C_{i+1})), H_{i+1}^*(t, X, b_{i+1}^0(C_{i+1})), H_i^*(t, X, b_i^0(C_i))\big)\big]\\
&\quad - \Delta_i^H(Z, b_i^0(C_i); W^\infty(t))\Big|^2\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\big|\mathbb{E}_{C_{i+1}}\big[\sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})), w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i), \\
&\quad b_{i+1}^0(C_{i+1})), b_{i+1}^*(t, b_{i+1}^0(C_{i+1})), H_{i+1}^*(t, X, b_{i+1}^0(C_{i+1})), H_i^*(t, X, b_i^0(C_i))\big)\big] \\
&\quad - \mathbb{E}_{C_{i+1}}\big[\sigma_i^{\mathbf{H}}\big(\Delta_{i+1}^{H}(Z, C_{i+1}; W^\infty(t)), w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i), \\
&b_{i+1}^0(C_{i+1})), b_{i+1}^*(t, b_{i+1}^0(C_{i+1})), H_{i+1}(X, C_{i+1}; W^\infty(t)), H_i(X, C_i; W^\infty(t))\big)\big]\big|^2\Big] \\
&\overset{(a)}{\leq} K\mathbb{E}\big[\mathbb{E}_{C_{i+1}}\big[\big(1 + |w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i), b_{i+1}^0(C_{i+1}))| \\
&\quad + |b_{i+1}^*(t, b_{i+1}^0(C_{i+1}))|\big)|\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})) - \Delta_{i+1}^{H}(Z, C_{i+1}; W^\infty(t))|\big]^2\big] \\
&\quad + K\mathbb{E}\big[\mathbb{E}_{C_{i+1}}\big[\big(1 + |w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i), b_{i+1}^0(C_{i+1}))| \\
&\quad + |b_{i+1}^*(t, b_{i+1}^0(C_{i+1}))|\big)\big(1 + |\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1}))| \\
&\quad + |\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})) - \Delta_{i+1}^{H}(Z, C_{i+1}; W^\infty(t))|\big) \\
&\quad \times \big(|H_{i+1}^*(t, X, b_{i+1}^0(C_{i+1})) - H_{i+1}(X, C_{i+1}; W^\infty(t))| \\
&\quad + |H_i^*(t, X, b_i^0(C_i)) - H_i(X, C_i; W^\infty(t))|\big)\big]^2\big] \\
&\overset{(b)}{\leq} K_{T,L}(D_{i+1}^H(t) + D_{i+1}(t) + D_i(t)) + KQ_i(t),
\end{aligned}
$$

where (a) follows from Assumption 2.6, (b) follows from Step 1, and we define

$$
\begin{aligned}
Q_i(t) = \mathbb{E}\big[\mathbb{E}_{C_{i+1}}\big[&|w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i), b_{i+1}^0(C_{i+1}))| \\
&\times \big(|\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1}))| \\
&+ |\Delta_{i+1}^{H*}(t, Z, b_{i+1}^0(C_{i+1})) - \Delta_{i+1}^{H}(Z, C_{i+1}; W^\infty(t))|\big) \\
&\times |H_i^*(t, X, b_i^0(C_i)) - H_i(X, C_i; W^\infty(t))|\big]^2\big].
\end{aligned}
$$

The bounding of $Q_i(t)$ requires some more care. In particular, for $B > 0$, define

$$
E = \big\{|w_{i+1}^*(t, w_{i+1}^0(C_i, C_{i+1}), b_i^0(C_i), b_{i+1}^0(C_{i+1}))| \geq B\big\}.
$$

Upon decomposing the inner expectation of $Q_i(t)$ into the sum of $\mathbb{I}(E)$ and $\mathbb{I}(\neg E)$, together with Step 1, via an appropriate use of Cauchy–Schwarz's inequality, it is easy to see that

$$
\begin{aligned}
Q_i(t) &\leq K_{T,L}B(D_i(t) + D_{i+1}^H(t)) + K_{T,L}(1 + D_i(t) + D_{i+1}^H(t))\mathbb{P}(E)^{1/8} \\
&\leq K_{T,L}(1 + B)(D_i(t) + D_{i+1}^H(t)) + K_{T,L}e^{-KB^2},
\end{aligned}
$$

which holds for any $B > 0$. Combining these bounds together and Step 2, we obtain

$$
\begin{aligned}
D_i^H(t) &\leq \frac{K_{T,L}}{M} + K_{T,L}(1 + B)(D_{i+1}^H(t) + D_{i+1}(t) + D_i(t)) + K_{T,L}e^{-KB^2} \\
&\leq \frac{K_{T,L}}{M} + K_{T,L}(1 + B)\Big(D_{i+1}^H(t) + \frac{1}{M}\Big) + K_{T,L}e^{-KB^2}.
\end{aligned}
$$

Then choosing $B = c\sqrt{\log M}$ for an appropriate constant c leads to the desired conclusion.

Step 4. Let us define

$$
\begin{aligned}
D_1^w(t) &= \mathbb{E}[|\Delta_1^w(Z, C_1; W^\infty(t)) - \Delta_1^{w*}(t, Z, w_1^0(C_1))|^2],\\
D_2^w(t) &= \mathbb{E}[|\Delta_2^w(Z, C_1, C_2; W^\infty(t)) - \Delta_2^{w*}(t, Z, w_1^0(C_1), w_2^0(C_1, C_2), b_2^0(C_2))|^2],\\
D_i^w(t) &= \mathbb{E}[|\Delta_i^w(Z, C_{i-1}, C_i; W^\infty(t))\\
&\qquad - \Delta_i^{w*}(t, Z, w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i))|^2], \quad i = 3, \dots, L-2,\\
D_{L-1}^w(t) &= \mathbb{E}\big[\big|\Delta_{L-1}^w(Z, C_{L-2}, C_{L-1}; W^\infty(t)) - \Delta_{L-1}^{w*}(t, Z, w_{L-1}^0(C_{L-2}, C_{L-1}),\\
&\qquad w_L^0(C_{L-1}, 1), b_{L-2}^0(C_{L-2}), b_{L-1}^0(C_{L-1}))\big|\big],\\
D_L^w(t) &= \mathbb{E}[|\Delta_L^w(Z, C_{L-1}, 1; W^\infty(t)) - \Delta_L^{w*}(t, Z, w_L^0(C_{L-1}, 1), b_{L-1}^0(C_{L-1}))|^2],\\
D_i^b(t) &= \mathbb{E}[|\Delta_i^b(Z, C_i; W^\infty(t)) - \Delta_i^{b*}(t, Z, b_i^0(C_i))|^2], \qquad i = 2, \dots, L-2,\\
D_{L-1}^b(t) &= \mathbb{E}[|\Delta_{L-1}^b(Z, C_{L-1}; W^\infty(t)) - \Delta_{L-1}^{b*}(t, Z, w_L^0(C_{L-1}, 1), b_{L-1}^0(C_{L-1}))|^2],\\
D_L^b(t) &= \mathbb{E}[|\Delta_L^b(Z, 1; W^\infty(t)) - \Delta_L^{b*}(t, Z)|^2].
\end{aligned}
$$

We claim that for any $t \le T$,

$$
\max_{1\le i\le L} D_i^w(t) \le K_{T,L}\frac{\log^{1/2} M}{M}, \quad \max_{2\le i\le L} D_i^b(t) \le K_{T,L}\frac{\log^{1/2} M}{M}.
$$

Indeed, by Assumption 2.6, for $3 \le i \le L-2$,

$$
\begin{aligned}
D_i^w(t) \le K\mathbb{E}\big[1 &+ |\Delta_i^{H*}(t, Z, b_i^0(C_i))|^2\\
&+ |\Delta_i^{H*}(t, Z, b_i^0(C_i)) - \Delta_i^H(Z, C_i; W^\infty(t))|^2\big] D_{i-1}(t)\\
&+ K(D_i^H(t) + D_i(t)).
\end{aligned}
$$

The claim for D_i^w then follows from Steps 1, 2 and 3. The rest are similar.

Step 5. With the same argument as Lemma B.4, given Step 1, one gets that for $2 \le i \le L$, any $t \le T$ and any $B \ge 0$,

$$
\begin{aligned}
&\mathbb{E}\big[\big|\mathbb{E}_Z[\Delta_i^w(Z, C_{i-1}, C_i; W^M(t)) - \Delta_i^w(Z, C_{i-1}, C_i; W^\infty(t))]\big|^2\big]^{1/2}\\
&\qquad \le K_{T,L}((1+B)\langle W^M - W^\infty\rangle_t + e^{-KB^2}).
\end{aligned}
$$

As such, by Step 4,

$$
\begin{aligned}
&\mathbb{E}[|\Delta_i^w(Z, C_{i-1}, C_i; W^M(t)) - \Delta_i^{w*}(t, Z, w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}), b_i^0(C_i))|^2]^{1/2}\\
&\le |D_i^w(t)|^{1/2} + \mathbb{E}[|\Delta_i^w(Z, C_{i-1}, C_i; W^M(t)) - \Delta_i^w(Z, C_{i-1}, C_i; W^\infty(t))|^2]^{1/2}\\
&\le K_{T,L}\Big(\frac{\log^{1/4} M}{M^{1/2}} + (1+B)\langle W^M - W^\infty\rangle_t + e^{-KB^2}\Big).
\end{aligned}
$$

One can obtain similar results for Δ_1^w and Δ_i^b. Hence, we obtain that for all $t \le T$,

$$\langle W^M - W^\infty\rangle_t \le K_{T,L}\int_0^t \Big(\frac{\log^{1/4} M}{M^{1/2}} + (1+B)\langle W^M - W^\infty\rangle_t + e^{-KB^2}\Big)\,ds.$$

Since $\langle W^M - W^\infty\rangle_0 = 0$, Gronwall's inequality implies that

$$\sup_{t\le T}\langle W^M - W^\infty\rangle_t \le K_{T,L}\inf_{B>0}\Big[\Big(\frac{\log^{1/4} M}{M^{1/2}} + e^{-KB^2}\Big)e^{K_{T,L}(1+B)}\Big] \le K_{T,L}\frac{1}{M^{0.499}},$$

for sufficiently large M.

Furthermore, with the same argument as Lemma B.3, given Step 1, one gets that for $2 \le i \le L-2$ and any $t \le T$,

$$\mathbb{E}[|H_i(X, C_i; W^M(t)) - H_i(X, C_i; W^\infty(t))|^2]^{1/2} \le K_{T,L}\langle W^M - W^\infty\rangle_t.$$

As such, together with Step 2, we get

$$\sup_{t\le T}\mathbb{E}[|H_i(X, C_i; W^M(t)) - H_i^*(t, X, b_i^0(C_i))|^2]^{1/2} \le K_{T,L}\frac{1}{M^{0.499}},$$

for sufficiently large M. ∎

D.3. Proof of Proposition 5.2

Proof of Proposition 5.2. It is easy to see that under the canonical neuronal ensemble (Ω^M, P^M), the functions $\{w_i^0\}_{i=1}^L$ and $\{b_i^0\}_{i=2}^L$ satisfy the i.i.d. initialization law, according to equation (5.1)–(5.5). To derive the $\bar\eta$-independence property, recall from the construction that for $i \le L-1$, $C_i(j_i) = (\lambda_i(j_i), \theta_i(j_i))$ and $\{C_i(j_i)\}_{j_i\in[n_i]}$ are sampled from $(P_0 \times \mathrm{Unif}([M]))^{n_i}$ conditional on that $\{\theta_i(j_i)\}_{j_i\in[n_i]}$ are all distinct. Notice then, for $i \le L-1$ and any $j \in [n_i]$,

$$\mathbb{E}[f(C_i(j)) \mid \{C_i(h), h \ne j\}] = \frac{1}{M - n_i + 1}\sum_{\theta\notin\{\theta_i(h):h\ne j\}}\mathbb{E}[f(C_i(j)) \mid \theta_i(h) = \theta].$$

Thus, for 1-bounded function f, we have

$$\begin{aligned}
&\big|\mathbb{E}[f(C_i(j)) \mid \{C_i(h), h \ne j\}] - \mathbb{E}[f(C_i(j))]\big| \\
&\quad\le \frac{1}{M}\sum_{\theta\in\{\theta_i(h):h\ne j\}}|\mathbb{E}[f(C_i(j)) \mid \theta_i(h) = \theta]| \\
&\qquad + \frac{n_i - 1}{M(M - n_i + 1)}\sum_{\theta\notin\{\theta_i(h):h\ne j\}}|\mathbb{E}[f(C_i(j)) \mid \theta_i(h) = \theta]| \\
&\quad\le 2\frac{n_i - 1}{M}.
\end{aligned}$$

The claim is trivial for $i = L$. ∎

D.4. Proofs of Corollaries 5.5 and 5.6

Proof of Corollary 5.5. By Proposition 5.2 and Corollary 4.9 (in particular, one of the intermediate steps in its proof), we have that for sufficiently large M, with probability at least $1 - 3\delta - KLn_{\max}\exp(-Kn_{\min}^{c_2})$,

$$\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|\mathbf{H}_i(\lfloor t/\epsilon\rfloor, X, j_i) - H_i(X, C_i(j_i); W^M(\lfloor t/\epsilon\rfloor\epsilon))|^2]\Big)^{1/2} = \tilde{O}(n_{\min}^{-c_1} + \epsilon^{c_1}),$$

where we recall that $\{C_i(j_i)\}_{j_i\in[n_i]}$ are sampled according to the sampling rule $\bar{P}_{\mathbf{n}}^M$ as described in Section 5.1.1. On the other hand, since $\mathrm{Law}(C_i(j_i)) = P_i^M$, by Theorem 5.3,

$$\mathbb{E}\Big[\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|H_i(X, C_i(j_i); W^M(\lfloor t/\epsilon\rfloor\epsilon)) - H_i^*(\lfloor t/\epsilon\rfloor\epsilon, X, B_i)|^2]\Big]^{1/2} \le \frac{K_{T,L}}{M^{0.499}},$$

which yields, for any $\gamma > 0$,

$$\mathbb{P}\Big(\frac{1}{n_i}\sum_{j_i=1}^{n_i}\mathbb{E}_Z[|H_i(X, C_i(j_i); W^M(\lfloor t/\epsilon\rfloor\epsilon)) - H_i^*(\lfloor t/\epsilon\rfloor\epsilon, X, B_i)|^2] \ge \gamma\Big) \le \frac{K_{T,L}}{\gamma M^{0.9}}.$$

Finally, by following the argument in the proof of Corollary 4.9, we have

$$\mathbb{E}[|H_i^*(\lfloor t/\epsilon\rfloor\epsilon, X, B_i) - H_i^*(t, X, B_i)|^2]^{1/2} \le K_{T,L}\epsilon.$$

The proof concludes by combining this with the previous two probability bounds and taking $M \to \infty$ and then $\gamma \to 0$. ∎

Proof of Corollary 5.6. For $3 \le i \le L-2$, since $b_i^0(C_i) = B_i$ is a constant,

$$\begin{aligned}
&w_i^\infty(t, C_{i-1}, C_i) - w_i^\infty(0, C_{i-1}, C_i)\\
&\quad= -\int_0^t \xi_i^{\mathbf{w}}(s)\mathbb{E}_Z\big[\sigma_i^{\mathbf{w}}\big(\Delta_i^{H*}(s, Z, b_i^0(C_i)), w_i^*(s, w_i^0(C_{i-1}, C_i), b_{i-1}^0(C_{i-1}),\\
&\qquad b_i^0(C_i)), b_i^*(s, b_i^0(C_i)), H_i^*(s, X, b_i^0(C_i)), H_{i-1}^*(s, X, b_{i-1}^0(C_{i-1}))\big)\big]\,ds\\
&\quad= -\int_0^t \xi_i^{\mathbf{w}}(s)\mathbb{E}_Z\big[\bar{\sigma}_i^{\mathbf{w}}\big(\Delta_i^{H*}(s, Z, b_i^0(C_i)), b_i^*(s, b_i^0(C_i)),\\
&\qquad H_i^*(s, X, b_i^0(C_i)), H_{i-1}^*(s, X, b_{i-1}^0(C_{i-1}))\big)\big]\,ds\\
&\quad= -\int_0^t \xi_i^{\mathbf{w}}(s)\mathbb{E}_Z\big[\bar{\sigma}_i^{\mathbf{w}}\big(\Delta_i^{H*}(s, Z, B_i), b_i^*(s, B_i), H_i^*(s, X, B_i),\\
&\qquad H_{i-1}^*(s, X, B_{i-1})\big)\big]\,ds,
\end{aligned}$$

which is independent of C_{i-1} and C_i. The desired claim readily follows. ∎

E. Remaining proofs for Section 6

E.1. Proof of Theorem 6.2

First we show that if $w_1(0, C_1)$ has full support, then so does $w_1(t, C_1)$ at any time t. Note that the following result holds beyond the setting of Theorem 6.2.

Lemma E.1. *Consider the MF ODEs (as described in Section 2.2) with $L = 2$ and $\mathbb{W}_1 = \mathbb{R}^d$ (for some positive integer d), under Assumptions 2.4–2.6 and 4.6. Let us disregard the bias of the second layer by considering $\xi_2^{\mathfrak{b}}(\cdot) = 0$ and $b_2(0, \cdot) = 0$. Suppose that the support of $\mathrm{Law}(w_1(0, C_1), w_2(0, C_1, 1))$ contains the graph of a continuous function $F\colon \mathbb{W}_1 \to \mathbb{W}_2$ such that $|F(u)| \le K$ for all $u \in \mathbb{W}_1$. Then for all finite time t, the support of $\mathrm{Law}(w_1(t, C_1))$ is $\mathbb{W}_1$.*

Proof. Since the support of $\mathrm{Law}(w_1(0, C_1), w_2(0, C_1, 1))$ contains the graph of $F\colon \mathbb{W}_1 \to \mathbb{W}_2$, we can choose the neuronal embedding so that there is a choice $C_1(u)$ for each $u \in \mathbb{W}_1$ such that $w_1(0, C_1(u)) = u$ and $w_2(0, C_1(u), 1) = F(u)$, and furthermore, for any neighborhood U of $(u, F(u))$, $(w_1(0, C_1), w_2(0, C_1, 1))$ lies in U with positive probability. For an arbitrary $T \ge 0$, let us define $M\colon [0, T] \times \mathbb{W}_1 \to \mathbb{W}_1$ by $M(t, u) = w_1(t, C_1(u))$.

We show that M is continuous. In the following, we define K_t to be a generic constant that changes with t and is finite with finite t. We first have from Assumption 2.6 that

$$|\Delta_2^H(t, z, 1)| \le K, \quad |\Delta_2^w(t, z, c_1, 1)| \le K(1 + |\Delta_2^H(t, z, 1)|) \le K,$$

which implies, by Assumption 2.4,

$$|w_2(t, c_1, 1)| \le |w_2(0, c_1, 1)| + K_t.$$

In particular, for any $u \in \mathbb{W}_1$,

$$|w_2(t, C_1(u), 1)| \le F(u) + K_t \le K_t.$$

We then have from Assumptions 2.5–2.6 that

$$\begin{aligned}
&|H_1(t, x, c_1) - H_1(t, x, c_1')| \le K|w_1(t, c_1) - w_1(t, c_1')|,\\
&|\Delta_2^w(t, z, c_1, 1) - \Delta_2^w(t, z, c_1', 1)|\\
&\quad \le K(1 + |\Delta_2^H(t, z, 1)|)|H_1(t, x, c_1) - H_1(t, x, c_1')| + K|w_2(t, c_1, 1) - w_2(t, c_1', 1)|\\
&\quad \le K(|w_2(t, c_1, 1) - w_2(t, c_1', 1)| + |w_1(t, c_1) - w_1(t, c_1')|),\\
&|\Delta_1^H(t, z, c_1) - \Delta_1^H(t, z, c_1')|\\
&\quad \le K(1 + |\Delta_2^H(t, z, 1)|)\big(|w_2(t, c_1, 1) - w_2(t, c_1', 1)|\\
&\qquad + (1 + |w_2(t, c_1, 1)| + |w_2(t, c_1', 1)|)|H_1(t, x, c_1) - H_1(t, x, c_1')|\big)
\end{aligned}$$

$$
\begin{aligned}
&\le K|w_2(t,c_1,1)-w_2(t,c_1',1)| \\
&\quad + K_t(1+|w_2(0,c_1,1)|+|w_2(0,c_1',1)|)|w_1(t,c_1)-w_1(t,c_1')|, \\
&|\Delta_1^w(t,z,c_1)-\Delta_1^w(t,z,c_1')| \\
&\le K(|\Delta_1^H(t,z,c_1)-\Delta_1^H(t,z,c_1')|+|w_1(t,c_1)-w_1(t,c_1')|) \\
&\le K|w_2(t,c_1,1)-w_2(t,c_1',1)| \\
&\quad + K_t(1+|w_2(0,c_1,1)|+|w_2(0,c_1',1)|)|w_1(t,c_1)-w_1(t,c_1')|.
\end{aligned}
$$

Defining

$$
R(t)=|w_2(t,C_1(u),1)-w_2(t,C_1(u'),1)|^2+|w_1(t,C_1(u))-w_1(t,C_1(u'))|^2
$$

for some $u,u'\in\mathbb{W}_1$, we then have for any $t\le T$,

$$
\begin{aligned}
\frac{d}{dt}R(t) &\le K_T(1+|w_2(0,C_1(u),1)|+|w_2(0,C_1(u'),1)|)^2R(t) \\
&= K_T(1+|F(u)|+|F(u')|)^2R(t) \\
&\le K_T R(t),
\end{aligned}
$$

which implies that $R(t)\le R(0)\exp(K_T t)$. In addition, by Assumption 2.6,

$$
\begin{aligned}
|\Delta_1^H(t,z,c_1)| &\le K(1+|\Delta_2^H(t,z,1)|)(1+|w_2(t,c_1,1)|) \\
&\le K|w_2(0,c_1,1)|+K_t, \\
|\Delta_1^w(t,z,c_1)| &\le K(1+|\Delta_1^H(t,z,c_1)|) \\
&\le K|w_2(0,c_1,1)|+K_t,
\end{aligned}
$$

which leads to

$$
|w_1(t,c_1)-w_1(t',c_1)|\le K_{t\vee t'}(1+|w_2(0,c_1,1)|)|t-t'|.
$$

Since $R(0)=|F(u)-F(u')|^2+|u-u'|^2\to 0$ as $u\to u'$, we deduce, for $t,t'\le T$, that

$$
\begin{aligned}
&|w_1(t,C_1(u))-w_1(t',C_1(u'))| \\
&\quad\le |w_1(t,C_1(u))-w_1(t',C_1(u))|+|w_1(t',C_1(u))-w_1(t',C_1(u'))| \\
&\quad\le K_T(1+|F(u)|)|t-t'|+\sqrt{R(0)}\exp(K_T T)\to 0
\end{aligned}
$$

as $(u,t)\to(u',t')$. This shows that $M(t,u)=w_1(t,C_1(u))$ is continuous.

Recall that $\mathbb{W}_1=\mathbb{R}^d$, and consider the sphere $\mathbb{S}^d$ which is a compactification of $\mathbb{W}_1$. We extend $M\colon[0,T]\times\mathbb{S}^d\to\mathbb{S}^d$ by fixing the point at infinity, which remains a continuous map since

$$
\begin{aligned}
|M(t,u)-u| = |M(t,u)-M(0,u)| &= |w_1(t,C_1(u))-w_1(0,C_1(u))| \\
&\le K_T(1+|F(u)|)t\le K_T t.
\end{aligned}
$$

Let $M_t\colon \mathbb{W}_1 \to \mathbb{W}_1$ be defined by $M_t(u) = M(t,u)$. Observe that if M_t is surjective for all t, then the support of $\mathrm{Law}(w_1(t,C_1))$ is $\mathbb{W}_1$, since for a neighborhood B of $M(t,u) = w_1(t,C_1(u))$, $\mathbb{P}(w_1(t,C_1) \in B) = \mathbb{P}(w_1(0,C_1) \in M_t^{-1}(B)) > 0$. It is indeed true that M_t is surjective for all t for the following reason. If M_t fails to be surjective for some t, then for some $p \in \mathbb{S}^d$, $M_t\colon \mathbb{S}^d \to \mathbb{S}^d \setminus \{p\} \to \mathbb{S}^d$ is homotopic to the constant map, but M then gives a homotopy from the identity map M_0 on the sphere to a constant map, which is a contradiction as the sphere $\mathbb{S}^d$ is not contractible. This finishes the proof of the claim. ∎

We are ready to prove Theorem 6.2. We recall the setting of Theorem 6.2, and in particular, the neural network (6.1).

Proof of Theorem 6.2. It is easy to check that Assumptions 2.4–2.6 hold. Therefore, by Theorem 3.1, the solution to the MF ODEs exists uniquely, and by Lemma E.1, the support of $\mathrm{Law}(w_1(t,C_1))$ is $\mathbb{R}^d$ at all t. We recall from the convergence assumption the limits $\bar{w}_1$ and $\bar{w}_2$, and we shall first prove $(\bar{w}_1, \bar{w}_2)$ is a global minimizer of $\mathcal{L}$ in Case 1 and $\mathcal{L}(\bar{w}_1, \bar{w}_2) = 0$ in Case 2.

By the convergence assumption, we have that for any $\epsilon > 0$, there exists $T(\epsilon)$ such that for all $t \geq T(\epsilon)$ and P-almost every c_1,

$$\begin{aligned}\epsilon &\geq |\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(t,X))\varphi_2'(H_2(t,X,1))\varphi_1(\langle w_1(t,c_1), X\rangle)]| \\ &= |\langle \mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(t,X)) \mid X = x]\varphi_2'(H_2(t,x,1)), \varphi_1(\langle w_1(t,c_1), x\rangle)\rangle_{L^2(\mathcal{P}_X)}|.\end{aligned}$$

Let $\mathcal{H}(f_1, f_2, x) = \mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \hat{y}(X; f_1, f_2)) \mid X = x]\varphi_2'(H_2(x; f_1, f_2))$. Since $\mathrm{Law}(w_1(t,C_1))$ has full support, we obtain that, for u in a dense subset of $\mathbb{R}^d$,

$$|\langle \mathcal{H}(w_1(t,\cdot), w_2(t,\cdot,1), x), \varphi_1(\langle u, x\rangle)\rangle_{L^2(\mathcal{P}_X)}| \leq \epsilon.$$

Since φ_1' is bounded and $|X| \leq K$, $\mathbb{E}_X[|\varphi_1(\langle u', X\rangle) - \varphi_1(\langle u, X\rangle)|^2] \to 0$ as $u' \to u$. Hence,

$$|\langle \mathcal{H}(w_1(t,\cdot), w_2(t,\cdot,1), x), \varphi_1(\langle u, x\rangle)\rangle_{L^2(\mathcal{P}_X)}| \leq \epsilon,$$

for all $u \in \mathbb{R}^d$. We claim that $\mathcal{H}(w_1(t,\cdot), w_2(t,\cdot,1), X) \to \mathcal{H}(\bar{w}_1, \bar{w}_2, X)$ in $L^1(\mathcal{P}_X)$ as $t \to \infty$. Assuming this claim, since φ_1 is bounded, we have for every $u \in \mathbb{R}^d$,

$$\langle \mathcal{H}(\bar{w}_1, \bar{w}_2, x), \varphi_1(u, x)\rangle_{L^2(\mathcal{P}_X)} = 0.$$

Since $\{\varphi_1(\langle u, \cdot\rangle) : u \in \mathbb{R}^d\}$ has dense span in $L^2(\mathcal{P}_X)$,

$$\mathcal{H}(\bar{w}_1, \bar{w}_2, x) = \mathbb{E}[\partial_2 \mathcal{L}(Y, \hat{y}(X; \bar{w}_1, \bar{w}_2)) \mid X = x]\varphi_2'(H_2(x; \bar{w}_1, \bar{w}_2)) = 0,$$

for $\mathcal{P}_X$-almost every x.

In Case 1, φ_2' is non-zero, and we get $\mathbb{E}[\partial_2 \mathcal{L}(Y, \hat{y}(X; \bar{w}_1, \bar{w}_2)) \mid X = x] = 0$ for $\mathcal{P}_X$-almost every x. For $\mathcal{L}$ convex in the second variable, for any measurable

function $\tilde{y}(x)$, we have

$$\mathcal{L}(y, \tilde{y}(x)) - \mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2)) \geq \partial_2 \mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2))(\tilde{y}(x) - \hat{y}(x; \bar{w}_1, \bar{w}_2)).$$

Taking expectation, we get $\mathbb{E}_Z[\mathcal{L}(Y, \tilde{y}(X))] \geq \mathcal{L}(\bar{w}_1, \bar{w}_2)$, i.e., $(\bar{w}_1, \bar{w}_2)$ is a global minimizer of $\mathcal{L}$.

Now, in Case 2, since y is a function of x and φ_2' is non-zero, we obtain that $\partial_2 \mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2)) = 0$ and hence $\mathcal{L}(y, \hat{y}(x; \bar{w}_1, \bar{w}_2)) = 0$ for $\mathcal{P}_X$-almost every x. That is, $\mathcal{L}(\bar{w}_1, \bar{w}_2) = 0$.

We now prove the claim. Using the assumptions and recalling the coupling π_t in Assumption 6.1 (4), we have

$$\begin{aligned}
&\mathbb{E}[|\mathcal{H}(w_1(t,\cdot), w_2(t,\cdot,1), X) - \mathcal{H}(\bar{w}_1(t,\cdot), \bar{w}_2(t,\cdot,1), X)|] \\
&\quad \leq \mathbb{E}[|\partial_2 \mathcal{L}(Y, \hat{y}(X; w_1, w_2))\varphi_2'(H_2(X; w_1, w_2)) \\
&\qquad - \partial_2 \mathcal{L}(Y, \hat{y}(X; \bar{w}_1, \bar{w}_2))\varphi_2'(H_2(X; \bar{w}_1, \bar{w}_2))|] \\
&\quad \leq K\mathbb{E}[|\varphi_2'(H_2(X; w_1, w_2)) - \varphi_2'(H_2(X; \bar{w}_1, \bar{w}_2))|] \\
&\qquad + K\mathbb{E}[|\varphi_2(H_2(X; w_1, w_2)) - \varphi_2(H_2(X; \bar{w}_1, \bar{w}_2))|] \\
&\quad \leq K\mathbb{E}[|H_2(X; w_1, w_2) - H_2(X; \bar{w}_1, \bar{w}_2)|] \\
&\quad \leq K\mathbb{E}_{\pi_t}[|\bar{w}_2(C_1) - w_2(t, C_1', 1)| \\
&\qquad + |\bar{w}_2(C_1)||\varphi_1(\langle w_1(t, C_1'), x\rangle) - \varphi_1(\langle \bar{w}_1(C_1), x\rangle)|] \\
&\quad \leq K\mathbb{E}_{\pi_t}[|\bar{w}_2(C_1) - w_2(t, C_1', 1)| + |\bar{w}_2(C_1)||w_1(t, C_1') - \bar{w}_1(C_1)|],
\end{aligned}$$

which converges to 0 by assumption. This proves the claim.

Finally, to connect $\mathcal{L}(\bar{w}_1, \bar{w}_2)$ with $\mathcal{L}(W(t))$ in the limit $t \to \infty$, we have

$$\begin{aligned}
&|\mathcal{L}(W(t)) - \mathcal{L}(\bar{w}_1, \bar{w}_2)| \\
&\quad = |\mathbb{E}_Z[\mathcal{L}(Y, \hat{y}(X; W(t))) - \mathcal{L}(Y, \hat{y}(X; \bar{w}_1, \bar{w}_2))]| \\
&\quad \leq K|\mathbb{E}_Z[\hat{y}(X; W(t)) - \hat{y}(X; \bar{w}_1, \bar{w}_2)]| \\
&\quad \leq K\mathbb{E}_{\pi_t}[|\bar{w}_2(C_1)||w_1(t, C_1') - \bar{w}_1(C_1)| + |w_2(t, C_1', 1) - \bar{w}_2(C_1)|],
\end{aligned}$$

which again converges to 0 by assumption. This completes the proof. ∎

E.2. Proof of Proposition 6.8

Proof of Proposition 6.8. We recall

$$\begin{aligned}
\frac{\partial}{\partial t} w_2^*(t, u_1, u_2, u_3) = -\mathbb{E}_Z\big[&\partial_2 \mathcal{L}(Y, \hat{y}^*(t, X))\, w_3^*(t, u_3) \\
&\times \varphi_3'(H_3^*(t, X))\varphi_2'(H_2^*(t, X, u_3))\varphi_1(\langle w_1^*(t, u_1), X\rangle)\big],
\end{aligned}$$

for $u_1 \in \mathbb{R}^d$, $u_2 \in \text{supp}(\rho^2)$, $u_3 \in \text{supp}(\rho^3)$. By the regularity assumption,

$$\Big|\frac{\partial}{\partial t} w_2^*(t,u_1,u_2,u_3)\Big| \le K\mathbb{E}_Z[|\partial_2 \mathcal{L}(Y,\hat{y}^*(t,X))|]|w_3^*(t,u_3)|.$$

Note that the right-hand side is independent of u_1 and u_2. Since

$$\int |w_3^*(t,u_3') - \bar{w}_3(u_3)|\, d\pi_t^3(u_3,u_3') \to 0$$

as $t \to \infty$ for a coupling π_t^3 of ρ_3 and itself, we have for some finite $t_0 \le K$,

$$\mathbb{E}[|\bar{w}_3(U_3)|] \le \mathbb{E}[|w_3^*(t_0,U_3)|] + K \le K,$$

where the last step is by an argument similar to the proof of Lemma 3.2 and the initialization assumption. As such, for all t sufficiently large, we have

$$\begin{aligned}
&\sup_{u_1\in\mathbb{R}^d,\, u_2\in\text{supp}(\rho^2)} \mathbb{E}_{U_3\sim\rho^3}\Big[\Big|\frac{\partial}{\partial t} w_2^*(t,u_1,u_2,U_3)\Big|\Big] \\
&\quad \le K\mathbb{E}_Z[|\partial_2\mathcal{L}(Y,\hat{y}^*(t,X))|]\,\mathbb{E}[|w_3^*(t,U_3)|] \\
&\quad \le K\mathbb{E}_Z[|\partial_2\mathcal{L}(Y,\hat{y}^*(t,X))|](K + \mathbb{E}[|\bar{w}_3(U_3)|]) \\
&\quad \le K\mathbb{E}_Z[|\partial_2\mathcal{L}(Y,\hat{y}^*(t,X))|].
\end{aligned}$$

The proof concludes once we show that $\mathbb{E}_Z[|\partial_2\mathcal{L}(Y,\hat{y}^*(t,X))|] \to 0$ as $t \to \infty$.

For a fixed x, let us write $\mathcal{L}(t,x) = \mathbb{E}[\mathcal{L}(Y,\hat{y}^*(t,X)) \mid X = x]$ and $\partial_2\mathcal{L}(t,x) = \mathbb{E}[\partial_2\mathcal{L}(Y,\hat{y}^*(t,X)) \mid X = x]$ for brevity. Consider Case 1. We claim that if there is an increasing sequence of time t_i so that $\lim_{i\to\infty}[\mathcal{L}(t_i,x) - \inf_{\hat{y}} \mathbb{E}[\mathcal{L}(Y,\hat{y}) \mid X = x]] = 0$, then $\lim_{i\to\infty} |\partial_2\mathcal{L}(t_i,x)| = 0$. Indeed, it suffices to show that for any subsequence t_{i_j} of t_i, there exists a further subsequence $t_{i_{j_k}}$ such that $\lim_{k\to\infty} |\partial_2\mathcal{L}(t_{i_{j_k}},x)| = 0$. In any subsequence t_{i_j} of t_i, using that $\mathcal{L}(t_{i_j},x)$ is convergent and the fact $\mathcal{L}(y,\hat{y}) \to \infty$ as $|\hat{y}| \to \infty$, we have that $\hat{y}^*(t_{i_j},x)$ is bounded. Hence, we obtain a subsequence $t_{i_{j_k}}$ for which $\hat{y}^*(t_{i_{j_k}},x)$ converges to some limit $\hat{y}^*$. By continuity, we have

$$\mathbb{E}[\mathcal{L}(Y,\hat{y}^*) \mid X = x] = \lim_{k\to\infty} \mathcal{L}(t_{i_{j_k}},x) = \inf_{\hat{y}} \mathbb{E}[\mathcal{L}(Y,\hat{y}) \mid X = x].$$

Thus, since $\mathcal{L}$ is convex in the second variable, we have $\mathbb{E}[\partial_2\mathcal{L}(Y,\hat{y}^*)|X = x] = 0$. Hence,

$$\lim_{k\to\infty} |\partial_2\mathcal{L}(t_{i_{j_k}},x)| = |\mathbb{E}[\partial_2\mathcal{L}(Y,\hat{y}^*)|X = x]| = 0,$$

as claimed. Similarly, we obtain in Case 2 that if there is an increasing sequence of time t_i so that $\lim_{i\to\infty} \mathcal{L}(t_i,x) = 0$, then $\lim_{i\to\infty} |\partial_2\mathcal{L}(t_i,x)| = 0$.

To show that $\mathbb{E}_Z[|\partial_2\mathcal{L}(t,X)|] \to 0$ as $t \to \infty$, it suffices to show that for any increasing sequence of times t_i tending to infinity, there exists a subsequence t_{i_j}

of t_i such that $\mathbb{E}_Z[|\partial_2 \mathscr{L}(t_{i_j}, X)|] \to 0$. In Case 1, we have $\lim_{i\to\infty} \mathcal{L}(W^*(t_i)) = \inf_{\tilde{y}} \mathbb{E}_Z[\mathscr{L}(Y, \tilde{y}(X))]$, so $\lim_{i\to\infty} \mathbb{E}_Z[\mathscr{L}(t_i, X) - \inf_{\tilde{y}(X)} \mathbb{E}_Z[\mathscr{L}(Y, \tilde{y}(X)) \mid X]] = 0$. Since $\mathscr{L}(t_i, X) - \inf_{\tilde{y}(X)} \mathbb{E}_Z[\mathscr{L}(Y, \tilde{y}(X)) \mid X]$ is non-negative, it converges to 0 in probability. Therefore, there is a further subsequence t_{i_j} for which $\mathscr{L}(t_{i_j}, X) - \inf_{\tilde{y}(X)} \mathbb{E}_Z[\mathscr{L}(Y, \tilde{y}(X)) \mid X]$ converges to 0 $\mathcal{P}$-almost surely. By the previous claim, $|\partial_2 \mathscr{L}(t_{i_j}, X)|$ converges to 0 $\mathcal{P}$-almost surely. Since $|\partial_2 \mathscr{L}(t_{i_j}, X)|$ is bounded $\mathcal{P}$-almost surely, we obtain that $\mathbb{E}_Z[|\partial_2 \mathscr{L}(t_{i_j}, X)|] \to 0$ from the bounded convergence theorem. The result in Case 2 can be established similarly. ∎

F. Remaining proofs for Section 7

F.1. Proof of Proposition 7.3

Proof of Proposition 7.3. We recall

$$\begin{aligned}&\frac{\partial}{\partial t} w_L(t, c_{L-1}, 1)\\ &\quad = -\mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \hat{y}(X; W(t)))\varphi_L'(H_L(X, 1; W(t)))\varphi_{L-1}(H_{L-1}(X, c_{L-1}; W(t)))],\end{aligned}$$

for $c_{L-1} \in \Omega_{L-1}$. By the regularity assumption,

$$\left|\frac{\partial}{\partial t} w_L(t, c_{L-1}, 1)\right| \le K\mathbb{E}_Z[|\partial_2 \mathscr{L}(Y, \hat{y}(X; W(t)))|].$$

Note that the right-hand side is independent of c_{L-1}. Then as argued in the proof of Proposition 6.8 (Section E.2), $\mathbb{E}_Z[|\partial_2 \mathscr{L}(Y, \hat{y}(X; W(t)))|] \to 0$ as $t \to \infty$. This completes the proof. ∎

G. Remaining proofs for Section 8

G.1. Proof of Theorem 8.2

Proof of Theorem 8.2. For brevity, let us write

$$\bar{H}_2(x) = H_2(x, 1; \bar{W}), \quad \bar{y}(x) = \hat{y}(x; \bar{W}).$$

We also define

$$\begin{aligned}G_2(t, u_1) &= \mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \hat{y}(t, X))\varphi_2'(H_2(t, X, 1))\varphi_1(\langle u_1, X\rangle)],\\ G_1(t, u_1) &= \mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \hat{y}(t, X))\varphi_2'(H_2(t, X, 1))\varphi_1'(\langle u_1, X\rangle)X],\\ \bar{G}_2(u_1) &= \mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \bar{y}(X))\varphi_2'(\bar{H}_2(X))\varphi_1(\langle u_1, X\rangle)],\\ \bar{G}_1(u_1) &= \mathbb{E}_Z[\partial_2 \mathscr{L}(Y, \bar{y}(X))\varphi_2'(\bar{H}_2(X))\varphi_1'(\langle u_1, X\rangle)X].\end{aligned}$$

We claim that as $t \to \infty$,

$$\mathbb{E}[|H_2(t, X, 1) - \bar{H}_2(X)|] \to 0, \quad \mathbb{E}[|\hat{y}(t, X) - \bar{y}(X)|] \to 0,$$
$$\mathbb{E}[|\partial_2 \mathscr{L}(Y, \hat{y}(t, X)) - \partial_2 \mathscr{L}(Y, \bar{y}(X))|] \to 0,$$

and uniformly in u_1,

$$|G_1(t, u_1) - \bar{G}_1(u_1)| \to 0, \quad |G_2(t, u_1) - \bar{G}_2(u_1)| \to 0.$$

Indeed, recall the coupling π_t in Assumption 8.1, we have from Assumption 6.1 (3):

$$\begin{aligned}
&\mathbb{E}_X[|H_2(t, X, 1) - \bar{H}_2(X)|] \\
&\quad = \mathbb{E}_X\big[\big|\mathbb{E}_{(C_1, C_1') \sim \pi_t}[w_2(t, C_1', 1)\varphi_1(\langle w_1(t, C_1'), X\rangle) - \bar{w}_2(C_1)\varphi_1(\langle \bar{w}_1(C_1), X\rangle)]\big|\big] \\
&\quad \le K\mathbb{E}_{\pi_t}[|\bar{w}_2(C_1)||w_1(t, C_1') - \bar{w}_1(C_1)| + |w_2(t, C_1', 1) - \bar{w}_2(C_1)|],
\end{aligned}$$

which tends to 0 as $t \to \infty$ by Assumption 8.1. The other claims can be derived similarly.

Consider the limit potential $\bar{\mathcal{F}}$ given by

$$\bar{\mathcal{F}}(u_1) = \frac{1}{2}|\bar{G}_2(u_1)|^2.$$

By Assumption 6.1 (3), $u_1 \mapsto \bar{\mathcal{F}}(u_1)$ is continuous. Notice that

$$\nabla \bar{\mathcal{F}}(u_1) = \frac{1}{2} \cdot 2\bar{G}_2(u_1)\nabla(\bar{G}_2(u_1)) = \bar{G}_2(u_1)\bar{G}_1(u_1).$$

Let $\bar{\mathcal{F}}^\infty \colon \mathbb{S}^{d-1} \to \mathbb{R}$ be defined by $\bar{\mathcal{F}}^\infty(\tilde{u}_1) = \lim_{r\to\infty} \bar{\mathcal{F}}(r\tilde{u}_1)$, which exists by Assumption 8.1. We shall argue that $\bar{\mathcal{F}}(u_1) = 0$ for all $u_1 \in \mathbb{R}^d$, by contradiction. To that end, let us assume that $\bar{\mathcal{F}}(u_1) \neq 0$ for some u_1. Note that $\bar{\mathcal{F}}$ is bounded by a constant by Assumption 6.1 (3). Thus, either there is a local maximizer u_1^* of $\bar{\mathcal{F}}$ with $\bar{\mathcal{F}}(u_1^*) > 0$ or there is a local maximizer $\tilde{u}_1^*$ of $\bar{\mathcal{F}}^\infty$ with $\bar{\mathcal{F}}^\infty(\tilde{u}_1^*) > 0$.

First consider the case that $\bar{\mathcal{F}}$ has a local maximizer u_1^* with $\bar{\mathcal{F}}(u_1^*) > 0$. Under Assumption 8.1, there exists $\delta \in (0, \bar{\mathcal{F}}(u_1^*))$ arbitrarily small so that for S_δ the connected component of the set $\{u \colon \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}(u_1^*) - \delta\}$ that contains u_1^*, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u_1)| > \xi$ for all $u_1 \in \partial \mathrm{cl}(S_\delta)$. Let T_0 be sufficiently large so that for $t \ge T_0$, we have, for $u_1 \in \partial \mathrm{cl}(S_\delta)$, $|\bar{G}_1(u_1) - G_1(t, u_1)| \le \xi / \sqrt{8\bar{\mathcal{F}}(u_1^*)}$, which implies

$$\begin{aligned}
\langle \bar{G}_1(u_1), G_1(t, u_1)\rangle &\ge |\bar{G}_1(u_1)|^2 - |\bar{G}_1(u_1)||\bar{G}_1(u_1) - G_1(t, u_1)| \\
&\ge |\bar{G}_1(u_1)|^2 - \frac{\xi}{\sqrt{8\bar{\mathcal{F}}(u_1^*)}}|\bar{G}_1(u_1)|
\end{aligned}$$

$$\overset{(a)}{\geq} |\bar{G}_1(u_1)|^2 - \frac{|\nabla \bar{\mathcal{F}}(u_1)|}{2|\bar{G}_2(u_1)|}|\bar{G}_1(u_1)|$$
$$= \frac{|\nabla \bar{\mathcal{F}}(u_1)|^2}{4\bar{\mathcal{F}}(u_1)} > \frac{\xi^2}{4\bar{\mathcal{F}}(u_1^*)}, \tag{G.1}$$

where (a) is because $2\bar{\mathcal{F}}(u_1^*) > 2\bar{\mathcal{F}}(u_1) = |\bar{G}_2(u_1)|^2$ for any $u_1 \in \partial \mathrm{cl}(S_\delta)$ by local maximality of u_1^* and continuity of $\bar{\mathcal{F}}$. Also, we further enlarge T_0 so that for $t \geq T_0$ and any $u_1 \in \mathrm{cl}(S_\delta)$, $|\bar{G}_2(u_1) - G_2(t, u_1)| \leq \frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_1^*) - \delta}$, and hence

$$G_2(t, u_1) \geq \bar{G}_2(u_1) - \frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_1^*) - \delta}$$
$$> \bar{G}_2(u_1) - \frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_1)} = \bar{G}_2(u_1) - \frac{1}{2}|\bar{G}_2(u_1)|, \tag{G.2}$$
$$G_2(t, u_1) \leq \bar{G}_2(u_1) + \frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_1^*) - \delta}$$
$$< \bar{G}_2(u_1) + \frac{1}{2}\sqrt{\bar{\mathcal{F}}(u_1)} = \bar{G}_2(u_1) + \frac{1}{2}|\bar{G}_2(u_1)|. \tag{G.3}$$

Furthermore, notice that

$$\frac{\partial}{\partial t}\bar{G}_2(w_1(t, C_1)) = \left\langle \bar{G}_1(w_1(t, C_1)), \frac{\partial}{\partial t} w_1(t, C_1) \right\rangle$$
$$= -w_2(t, C_1, 1)\langle \bar{G}_1(w_1(t, C_1)), G_1(t, w_1(t, C_1))\rangle. \tag{G.4}$$

Let $\tilde{\Omega}_1$ be the subset of Ω_1 consisting of c_1 where $|w_2(0, c_1, 1)| < |F(w_1(0, c_1))| + 1$ for F given in Assumption 6.1 (2). The proof of Lemma E.1 in fact shows that for any $t \geq 0$ and any open subset B of $\mathbb{R}^d$, there exists a positive mass of $C_1 \in \tilde{\Omega}_1$ such that $w_1(t, C_1) \in B$. In the following, we consider $C_1 \in \tilde{\Omega}_1$. We further divide the argument into two cases: $\bar{G}_2(u_1^*) > 0$ and $\bar{G}_2(u_1^*) < 0$.

Let us consider the case that $\bar{G}_2(u_1^*) > 0$. Then we can choose sufficiently small δ such that $\bar{G}_2(u_1) > 0$ for all $u_1 \in \mathrm{cl}(S_\delta)$. Furthermore, consider the scenario that there exists $T \geq T_0$ such that a positive mass of $(w_1(T, C_1), w_2(T, C_1, 1))$ with $C_1 \in \tilde{\Omega}_1$ has $w_1(T, C_1) \in S_\delta$ and $w_2(T, C_1, 1) < 0$. Note that if $w_1(t, C_1) \in \mathrm{cl}(S_\delta)$,

$$\frac{\partial}{\partial t} w_2(t, C_1, 1) = -G_2(t, w_1(t, C_1)) \leq -\left(\bar{G}_2(w_1(t, C_1)) - \frac{1}{2}|\bar{G}_2(w_1(t, C_1))|\right) < 0$$

by equation (G.2). Define $T_1 = \inf\{t \geq T : w_1(t, C_1) \notin S_\delta\}$. Then $t \mapsto w_2(t, C_1, 1)$ is decreasing on $t \in [T, T_1)$. Let us argue that $T_1 = \infty$. Indeed, suppose T_1 is finite. We then have, by continuity, $w_1(T_1, C_1) \in \partial \mathrm{cl}(S_\delta)$ and $w_2(T_1, C_1, 1) \leq w_2(T, C_1, 1) < 0$. As such, $\frac{\partial}{\partial t}\bar{G}_2(w_1(T_1, C_1)) > 0$ by equation (G.1) and (G.4). By continuity, for some $\gamma > 0$, we have $\frac{\partial}{\partial t}\bar{G}_2(w_1(T_1 + t, C_1)) > 0$ for all $t \in [0, \gamma]$. But then

$$\bar{G}_2(w_1(T_1 + t, C_1)) \geq \bar{G}_2(w_1(T_1, C_1)) \geq \sqrt{2(\bar{\mathcal{F}}(u_1^*) - \delta)},$$

and hence $w_1(T_1 + t, C_1) \in S_\delta$ for all $t \le \gamma$, contradicting the definition of T_1. Therefore, $T_1 = \infty$, i.e., for $t \ge T$ and $C_1 \in \tilde{\Omega}_1$ with $w_1(T, C_1) \in S_\delta$ and $w_2(T, C_1, 1) < 0$, we have $w_1(t, C_1) \in S_\delta$, which implies

$$G_2(t, w_1(t, C_1)) \overset{(a)}{\ge} \frac{1}{2} \bar{G}_2(w_1(t, C_1)) = \sqrt{\frac{1}{2} \bar{\mathcal{F}}(w_1(t, C_1))} \ge \sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u_1^*) - \delta)},$$

where (a) is by equation (G.2) and the fact $\bar{G}_2(u_1) > 0$ for all $u_1 \in \mathrm{cl}(S_\delta)$. In particular, there is a positive mass of $(w_1(t, C_1), w_2(t, C_1, 1))$ with $G_2(t, w_1(t, C_1)) \ge \sqrt{(\bar{\mathcal{F}}(u_1^*) - \delta)/2}$ for all $t \ge T$. Noting that

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\mathcal{L}(Y, \hat{y}(t, X))] &= -\mathbb{E}[|G_2(t, w_1(t, C_1))|^2] - \mathbb{E}[|G_1(t, w_1(t, C_1))|^2] \\ &\le -\mathbb{E}[|G_2(t, w_1(t, C_1))|^2], \end{aligned}$$

we obtain that $\frac{d}{dt}\mathbb{E}[\mathcal{L}(Y, \hat{y}(t, X))]$ is bounded above by a strictly negative constant for all $t \ge T$, which is a contradiction since $\mathcal{L}$ is bounded below.

Next consider the scenario that for all $t \ge T_0$, the probability that $w_1(t, C_1) \in S_\delta$ and $w_2(t, C_1, 1) < 0$ on $C_1 \in \tilde{\Omega}_1$ is zero. Let us argue that for any $t \ge T_0$ and for a.e. $C_1 \in \tilde{\Omega}_1$ with $w_1(t, C_1) \in S_\delta$, we have $w_1(s, C_1) \in S_\delta$ for all $s \in [T_0, t]$. Indeed, consider t and $C_1 \in \tilde{\Omega}_1$ such that $w_1(t, C_1) \in S_\delta$ and $w_1(T', C_1) \notin S_\delta$ for some $T' \in [T_0, t)$. Let $t' = \sup\{s \in [T', t] : w_1(s, C_1) \notin S_\delta\} < t$. By continuity, $w_1(t', C_1) \in \partial \mathrm{cl}(S_\delta)$, and so, by equation (G.1),

$$\langle \bar{G}_1(w_1(t', C_1)), G_1(t', w_1(t', C_1)) \rangle > \frac{\xi^2}{4\bar{\mathcal{F}}(u_1^*)}.$$

By continuity, there exists $t'' \in (t', t)$ such that for all $s \in [t', t'']$,

$$\langle \bar{G}_1(w_1(s, C_1)), G_1(s, w_1(s, C_1)) \rangle \ge \frac{\xi^2}{100\bar{\mathcal{F}}(u_1^*)}.$$

By definition of t', we also have $w_1(s, C_1) \in S_\delta$ and therefore $w_2(s, C_1, 1) \ge 0$ for any $s \in (t', t]$. Then, by equation (G.4), $\frac{\partial}{\partial t} \bar{G}_2(w_1(s, C_1)) \le 0$ for all $s \in (t', t'']$, and therefore

$$\bar{G}_2(w_1(t'', C_1)) \le \bar{G}_2(w_1(t', C_1)) = \sqrt{2(\bar{\mathcal{F}}(u_1^*) - \delta)},$$

where the equality follows from $w_1(t', C_1) \in \partial \mathrm{cl}(S_\delta)$. However, this contradicts with $w_1(t'', C_1) \in S_\delta$. Therefore, it holds that for any $t \ge T_0$, for a.e. $C_1 \in \tilde{\Omega}_1$ with $w_1(t, C_1) \in S_\delta$, $w_1(s, C_1) \in S_\delta$ and therefore $w_2(s, C_1, 1) \ge 0$ for all $s \in [T_0, t]$. Since $w_1(t, C_1)$ on $C_1 \in \tilde{\Omega}_1$ has full support at any $t \ge 0$, we have, for any $t_0 \ge T_0$, that there is a positive mass on $C_1 \in \tilde{\Omega}_1$ such that $w_1(t_0, C_1) \in S_\delta$, and hence, as shown, $w_1(s, C_1) \in S_\delta$ and $w_2(s, C_1, 1) \ge 0$ for all $s \in [T_0, t_0]$. Note that we have

$w_2(T_0, C_1, 1) \le M(T_0)$ for some finite $M(T_0) > 0$ for $C_1 \in \tilde{\Omega}_1$ (which follows from the fact $|\frac{\partial}{\partial t} w_2(t, \cdot, 1)| \le K$ by Assumption 6.1 (3) and that $|w_2(0, C_1, 1)| < |F(w_1(0, C_1))| + 1 \le K$). Also note that for $w_1(s, C_1) \in S_\delta$ and $s \ge T_0$,

$$\frac{\partial}{\partial t} w_2(s, C_1, 1) = -G_2(s, w_1(s, C_1)) \overset{(a)}{\le} -\frac{1}{2} \bar{G}_2(w_1(s, C_1)) = -\sqrt{\frac{1}{2} \bar{\mathcal{F}}(w_1(s, C_1))} \le -\sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u_1^*) - \delta)},$$

a strictly negative constant, where (a) is by equation (G.2) and the fact $\bar{G}_2(u_1) > 0$ for all $u_1 \in \mathrm{cl}(S_\delta)$. As such, for any $t_0 \ge T_0$ such that

$$M(T_0) - (t_0 - T_0)\sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u_1^*) - \delta)} < 0,$$

there is a positive mass on $C_1 \in \tilde{\Omega}_1$ such that firstly $w_2(s, C_1, 1) \ge 0$ for all $s \in [T_0, t_0]$ and secondly there exists $t \in [T_0, t_0]$ in which

$$w_2(t, C_1, 1) \le M(T_0) - (t - T_0)\sqrt{\frac{1}{2}(\bar{\mathcal{F}}(u_1^*) - \delta)} < 0.$$

We again obtain a contradiction.

The case $\bar{G}_2(u_1^*) < 0$ can be treated similarly, with the use of equation (G.2) replaced by equation (G.3). Both cases lead to a contradiction, ruling out the possibility that there is a local maximizer u_1^* of $\bar{\mathcal{F}}$ with $\bar{\mathcal{F}}(u_1^*) > 0$.

Next consider the case where $\bar{\mathcal{F}}$ does not have any local maximizer in $\mathbb{R}^d$ but $\bar{\mathcal{F}}^\infty$ has a local maximizer $\tilde{u}_1^*$ with $\bar{\mathcal{F}}^\infty(\tilde{u}_1^*) > 0$. Under Assumption 8.1 (and with the same argument in the discussion that follows), there exists $\delta \in (0, \bar{\mathcal{F}}^\infty(\tilde{u}_1^*))$ arbitrarily small so that for S_δ the connected component of the set $\{u \in \mathbb{R}^d : \bar{\mathcal{F}}(u) > \bar{\mathcal{F}}^\infty(\tilde{u}_1^*) - \delta\}$ which contains $r\tilde{u}_1^*$ for all r sufficiently large, there is $\xi > 0$ such that $|\nabla \bar{\mathcal{F}}(u)| > \xi$ for all $u \in \partial \mathrm{cl}(S_\delta)$. The rest of the argument can be repeated as before to yield a contradiction.

In short, we have shown that $\bar{\mathcal{F}}(u_1) = \frac{1}{2}|\bar{G}_2(u_1)|^2 = 0$ and, equivalently,

$$\mathbb{E}_Z[\partial_2 \mathcal{L}(Y, \bar{y}(X)) \varphi_2'(\bar{H}_2(X)) \varphi_1(\langle u_1, X \rangle)] = 0$$

for all $u_1 \in \mathbb{R}^d$. The remaining proof follows identically as in the proof of Theorem 6.2. ∎

Acknowledgments. Author ordering is randomized. The work was done in parts when P.-M. Nguyen was at Department of Electrical Engineering, Stanford University and H. T. Pham was at the University of Cambridge. A conference version [28] of the work appears in the International Conference on Learning Representations (ICLR) 2021.

H. T. Pham would like to thank Jan Vondrak for many helpful discussions and in particular for the shorter proof of Lemma C.1. We would like to thank Andrea Montanari for the succinct description of the difficulty in extending the mean field formulation to the multilayer case, in that there are multiple symmetry group actions in a multilayer network.

Funding. The work of P.-M. Nguyen was partially supported by grants NSF IIS-1741162 and ONR N00014-18-1-2729.

References

[1] A. Agazzi and J. Lu, Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. 2020, arXiv:2010.11858

[2] Z. Allen-Zhu, Y. Li, and Z. Song, A convergence theory for deep learning via over-parameterization. 2020, arXiv:1811.03962

[3] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*. Second edn., Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2008 Zbl 1145.35001 MR 2401600

[4] D. Araújo, R. I. Oliveira, and D. Yukimura, A mean-field limit for certain deep neural networks. 2019, arXiv:1906.00193

[5] F. Bach and L. Chizat, Gradient descent on infinitely wide neural networks: Global convergence and generalization. 2021, arXiv:2110.08084

[6] T. Chen and H. Chen, Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Network Sci. Eng.* **6** (1995), no. 4, 911–917

[7] L. Chizat, Sparse optimization on measures with over-parameterized gradient descent. *Math. Program.* **194** (2022), no. 1-2, Ser. A, 487–532 Zbl 1494.90082 MR 4445462

[8] L. Chizat and F. Bach, A note on lazy training in supervised differentiable programming. 2018, arXiv:1812.07956

[9] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS 2018 – Advances in Neural Information Processing Systems 31*, pp. 3040–3050, Curran Associates, 2018

[10] G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** (1989), no. 4, 303–314 Zbl 0679.94019 MR 1015670

[11] S. S. Du, X. Zhai, B. Poczos, and A. Singh, Gradient descent probably optimizes over-parameterized neural networks. In *ICLR 2019 – International conference on learning representations*, 2019

[12] W. E, C. Ma, and L. Wu, Machine learning from a continuous viewpoint, I. *Sci. China Math.* **63** (2020), no. 11, 2233–2266 Zbl 1472.68136 MR 4170870

[13] W. E and S. Wojtowytsch, On the Banach spaces associated with multi-layer ReLU networks: Function representation, approximation theory and gradient descent dynamics. 2020, arXiv:2007.15623

[14] C. Fang, J. D. Lee, P. Yang, and T. Zhang, Modeling from features: a mean-field framework for over-parameterized deep neural networks. 2020, arXiv:2007.01452

[15] V. Feldman and J. Vondrak, Generalization bounds for uniformly stable algorithms. In *NeurIPS 2018 – Advances in Neural Information Processing Systems 31*, pp. 9747–9757, Curran Associates, 2018

[16] E. A. Golikov, Dynamically stable infinite-width limits of neural classifiers. 2020, arXiv:2006.06574

[17] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS 2018 – Advances in Neural Information Processing Systems 31*, pp. 8580–8589, Curran Associates, 2018

[18] A. Javanmard, M. Mondelli, and A. Montanari, Analysis of a two-layer neural network via displacement convexity. *Ann. Statist.* **48** (2020), no. 6, 3619–3642 Zbl 1464.62401 MR 4185822

[19] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent. *J. Stat. Mech. Theory Exp.* (2020), no. 12, article no. 124002 Zbl 07330523 MR 4241355

[20] Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying, A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, pp. 6426–6436, 2020

[21] S. Mei, T. Misiakiewicz, and A. Montanari, Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. 2019, arXiv:1902.06015

[22] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** (2018), no. 33, E7665–E7671 MR 3845070

[23] P.-M. Nguyen, Mean field limit of the learning dynamics of multilayer neural networks. 2019, arXiv:1902.02880

[24] P.-M. Nguyen, *Mean field limit in neural network learning: Autoencoders and multilayer networks*. Ph.D. thesis, Stanford University, 2020

[25] P.-M. Nguyen, Analysis of feature learning in weight-tied autoencoders via the mean field lens. 2021, arXiv:2102.08373

[26] A. Nitanda and T. Suzuki, Stochastic particle gradient descent for infinite ensembles. 2017, arXiv:1712.05438

[27] A. Nitanda, D. Wu, and T. Suzuki, Particle dual averaging: optimization of mean field neural network with global convergence rate analysis. *J. Stat. Mech. Theory Exp.* (2022), no. 11, Paper No. 114010 Zbl 07632727 MR 4535581

[28] H. T. Pham and P.-M. Nguyen, Global convergence of three-layer neural networks in the mean field regime. In *ICLR 2021 – International conference on learning representations*, 2021

[29] H. T. Pham and P.-M. Nguyen, Limiting fluctuation and trajectorial stability of multilayer neural networks with mean field training. In *NeurIPS 2021 – Advances in Neural Information Processing Systems 34*, pp. 4843–4855, Curran Associates, 2021

[30] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.* **22** (1994), no. 4, 1679–1706 Zbl 0836.60015 MR 1331198

[31] G. Rotskoff, S. Jelassi, J. Bruna, and E. Vanden-Eijnden, Global convergence of neuron birth-death dynamics. 2019, arXiv:1902.01843

[32] G. M. Rotskoff and E. Vanden-Eijnden, Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. 2018, arXiv:1805.00915

[33] A. Shevchenko and M. Mondelli, Landscape connectivity and dropout stability of SGD solutions for over-parameterized neural networks. 2019, arXiv:1912.10095

[34] J. Sirignano and K. Spiliopoulos, Mean field analysis of neural networks: a law of large numbers. *SIAM J. Appl. Math.* **80** (2020), no. 2, 725–752 Zbl 1493.68333 MR 4074020

[35] J. Sirignano and K. Spiliopoulos, Mean field analysis of deep neural networks. *Math. Oper. Res.* **47** (2022), no. 1, 120–152

[36] A.-S. Sznitman, Topics in propagation of chaos. In *École d'été de Probabilités de Saint-Flour XIX—1989*, pp. 165–251, Lecture Notes in Math. 1464, Springer, Berlin, 1991 MR 1108185

[37] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pp. 210–268, Cambridge University Press, Cambridge, 2012 MR 2963170

[38] C. Wei, J. D. Lee, Q. Liu, and T. Ma, Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *NeurIPS 2019 – Advances in Neural Information Processing Systems 32*, pp. 9712–9724, Curran Associates, 2019

[39] S. Wojtowytsch, On the convergence of gradient descent training for two-layer ReLU-networks in the mean field regime. 2020, arXiv:2005.13530

[40] G. Yang and E. J. Hu, Feature learning in infinite-width neural networks. 2021, arXiv:2011.14522

[41] D. Zou, Y. Cao, D. Zhou, and Q. Gu, Gradient descent optimizes over-parameterized deep ReLU networks. *Mach. Learn.* **109** (2020), no. 3, 467–492 Zbl 1494.68245 MR 4075425

Received 23 May 2022; revised 30 November 2022.

Phan-Minh Nguyen
Department of Electrical Engineering, Stanford University, 350 Jane Stanford Way, David Packard Building, Stanford, CA 94305, USA; Current affiliation: The Voleon Group, Berkeley, CA 94704, USA; npminh@alumni.stanford.edu

Huy Tuan Pham
Department of Mathematics, Stanford University, 450 Jane Stanford Way, Building 380, Stanford, CA 94305, USA; huypham@stanford.edu